An Tausenden von Kochvideos lernt der Computer, welche Bewegung »schneiden« ist.

Foto: Uwe Dettmar

Videoerkennung: Ist es Kochen oder Winken?

von Markus Bernards

Gegenstände und Gesichter können Computer schon recht gut erkennen, auch dass sich etwas bewegt und in welche Richtung. Schwierigkeiten bereiten der Künstlichen Intelligenz aber noch zu erfassen, um welche Art von Bewegungen es sich handelt. Das lernen Computer jetzt im Labor von Prof. Hilde Kühne an der Goethe-Universität.

Nach ihrem 80. Geburtstag meldete sich Lydia S., eine alleinstehende Dame, beim Hausnotrufdienst der Caritas an. Sie erhielt einen Druckschalter an einer Schnur. Ein Druck auf den Knopf löste in der Zentrale des Hausnotrufdienstes Alarm aus, und eine Mitarbeiterin meldete sich per Telefon bei Frau S. und fragte, ob alles in Ordnung sei. Wenn sie in ihrer Wohnung war, hängte sie sich den Schalter um den Hals. Er gab ihr das gute Gefühl, Hilfe herbeirufen zu können, sollte sie zum Beispiel einmal stürzen. Als Lydia S. dann einen leichten Schlaganfall erlitt, stürzte und den Notruf dringend gebraucht hätte, lag der Schalter auf dem Sessel, wo sie ihn beim Fernsehen hingelegt hatte – unerreichbar für die am Boden Liegende. Es dauerte eine Weile, bis sie sich aus eigener Kraft aufrappeln und zum Telefonapparat schleppen konnte, um Hilfe zu rufen.

Notrufsysteme für ältere Menschen in Form von Handsendern zum Umhängen oder an Armbändern haben den Nachteil, dass sie abgelegt und im entscheidenden Moment nicht erreichbar sein können. Videokameras etwa in der Wohnung sind allerdings kaum eine akzeptable Alternative. »Niemand möchte sein Zuhause per Video von einem Notdienst überwachen las-

sen«, meint Hilde Kühne, Juniorprofessorin für Bilderkennungssysteme und maschinelles Lernen an der Goethe-Universität. »Schon gar nicht das Schlafzimmer oder das Bad, wo man aber schnell in eine kritische Situation kommen kann.«

ZUR PERSON

Prof.in Dr. Hilde Kühne, Jahrgang 1979, studierte Informatik (Computervisualistik) an der Universität Koblenz-Landau und promovierte am Karlsruher Institut für Technologie. Stationen ihres wissenschaftlichen Arbeitens waren das Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie und die Universität Bonn, bevor sie als Forscherin an das MIT-IBM-Watson AI Lab des Massachusetts Institute of Technology in den USA wechselte. Seit 2021 ist sie zusätzlich Juniorprofessorin für Bilderkennungssysteme und maschinelles Lernen an der Goethe-Universität. Sie ist Mitgründerin des Unternehmens KS-Research und wurde jüngst mit dem ICCV Helmholtz Prize ausgezeichnet zur Würdigung einer Arbeit, die zehn Jahre nach ihrer Veröffentlichung ihre Relevanz für die Forschung bewiesen hat.

kuehne@em.uni-frankfurt.de

Schutz der Privatsphäre

Vielleicht wäre es anders, wenn zwar in der Wohnung Videokameras aufgehängt würden, die Bilder aber nur ein Computer und kein Mensch zu sehen bekäme? Wenn der Computer bei einem Sturz den Notdienst alarmieren würde, ohne die Videodaten selbst zu übermitteln? Dann bliebe die Privatsphäre gewahrt, denn »der Computer interessiert sich nicht für die Person, die durch die Wohnung geht«, meint Kühne, die am automatischen Erkennen von Bewegungen forscht. »Für den Computer sind Videos schlicht Zahlenkolonnen.«

Um Stürze zu melden, müsste der Computer sie allerdings erst einmal von anderen Bewegungen unterscheiden können. Das ist allerdings schwerer als die Identifizierung von Gesichtern und Objekten auf Fotos. Denn das Computertraining mit Videos ist aufwändiger, alleine schon wegen der riesigen Datenmengen, die dafür verarbeitet werden müssen. 50 bis 100 Video-Einzelbilder (Frames) sind nötig, damit eine Bewegung sichtbar wird – also die 50- bis 100-fache Datenmenge eines Fotos.

Viele Begriffe für dieselbe Bewegung

Zudem trainieren Computer klassischerweise mit Texten, die beschreiben, was auf Fotos oder Videosequenzen zu sehen ist. Solche Verschlagwortungen, Annotationen genannt, werden von Menschen gemacht, die die Bilder ansehen und beschreiben. Auf diese Weise lernt der Computer, was zum Beispiel eine Tasse ist, wenn er sehr viele Bilder sieht, die mit der Annotation »Tasse« versehen sind. Bei Videos ist es viel zeitaufwändiger, Annotationen zu erstellen und so genug Trainingsmaterial zu erhalten, alleine wegen der großen Datenmengen und der längeren Zeitspannen, die für Videos veranschlagt werden müssen.

Dazu kommen zwei weitere Probleme: Zum einen gibt es für dieselbe Bewegung oft unterschiedliche Begriffe, die auch davon abhängen, wie lange eine Bewegung beobachtet werden kann. Kühne: »Wenn ich jemandem nur drei Sekunden zusehe, kann ich zum Beispiel sagen ›er rennt‹ oder ›er läuft‹. Wenn ich ihn 20 Sekunden beobachte, weiß ich ›er sprintet‹ oder ›er joggt‹. Sehe ich noch mehr von dem Video, und es taucht ein Hund auf oder eine Bushaltestelle, erkenne ich: ›Er flüchtet vor dem Hund‹ oder ›er hastet zur Bushaltestelle‹. Die Aufgabe, Bewegungen zu erkennen, ist daher schlecht definierbar, für Mensch und Computer.«

Lösung: Autonomes Lernen

Das zweite Problem liegt darin, wie Menschen den Datenstrom verarbeiten, den sie über Augen und Ohren empfangen. Wir nehmen Bewegungen nicht als etwas Kontinuierliches wahr, sondern unterteilen sie in kleinere Abschnitte, um sie uns zu merken. Im Gehirn werden diese Abschnitte dann wieder zu einem kontinuierlichen Bewegungsablauf zusammengefügt. Wie viele einzelne Abschnitte wahrgenommen werden, hängt dabei von den individuellen Erfahrungen und Fähigkeiten jedes Betrachters ab. Hilde Kühne nennt als Beispiel das Kunstturnen bei Olympia: »Die geschulten Wertungsrichter können den Bewegungsablauf einer Kür genau analysieren. Ich als Laie sehe dieselbe Abfolge, kann aber die einzelnen Elemente kaum unterscheiden.«

Wie also soll der Computer lernen? Autonom, findet Hilde Kühne, nicht mehr anhand von Annotationen, sondern selbstständig. Kühnes Team schöpft dazu aus einem Pool von 100 Millionen YouTube-Videos. Zum Lernen ist der Computer mit einem künstlichen neuronalen Netz bestückt. Dabei handelt es sich um Algorithmen, die im Prinzip so funktionieren wie Nervenzellen in einem Gehirn. »Eigentlich aber sind es mathematische Funktionen, die Zahlenkolonnen in andere Zahlenkolonnen umwandeln«, sagt Kühne.

Computer-Training

Aus jedem Videoclip erhält der Computer drei Informationen: Die eigentliche Videosequenz,

AUF DEN PUNKT GEBRACHT

- An 100 Millionen YouTube-Videos lernen Computer im Labor, Bewegungen zu erkennen.
- Da Textbeschreibungen für Bewegungsvideos verhältnismäßig aufwändig sind, trainieren die Computer autonom. Das Ziel: Videobild, Ton und Untertitel einer Bewegung werden durch Algorithmen miteinander verknüpft.
- Anwendungen sind beim Assistant Living denkbar oder im Erkennen gefährlicher Situationen in der Videoüberwachung.

die eine Bewegung zeigt, den Ton des Videos und eventuell noch Untertitel, die im Video eingeblendet werden. Ein Beispiel wäre eine Sequenz aus einem Kochvideo, in der die YouTube-Köchin eine Paprika schneidet und dabei spricht: »Jetzt schneiden wir die Paprika in Würfel.« Gleichzeitig erscheint im Untertitel »Paprika in Würfel schneiden«.

Für den Computer sind die Informationen Video, Ton und Untertitel drei Zahlenkolonnen, aus denen er mithilfe einer mathematischen Funktion drei Punkte in einem sogenannten Embedding Space errechnet, den man sich als großen, durchsichtigen Würfel vorstellen kann. Kühne erklärt: »Wir wollen eine mathematische Funktion finden, die die drei Zahlenkolonnen zur Bewegung ›schneiden‹ so übersetzt, dass sie drei nahe beieinanderliegende Punkte im Embedding Space bilden. Video-, Ton- und Untertiteldaten einer anderen Bewegung wie zum Beispiel ›winken‹ sollte entsprechend drei Punkte an einer anderen Stelle des Embedding Space generieren.«

Das Training des Computers besteht nun darin, viele Videos zu analysieren und für verschiedene Bewegungen jeweils Punktegruppen im Embedding Space zu generieren. Im nächsten Schritt zeigen die Informatikerinnen und Informatiker dem Computer annotierte Videos, sodass er die Punktegruppen mit den dazugehörigen Begriffen wie »schneiden« oder »winken« verknüpfen kann und nun »weiß«, wie die jeweiligen Bewegungen genannt werden.

Viele Anwendungen

Irgendwann soll der Computer dann die verschiedensten Bewegungen erkennen können, auch wenn sie Teile eines längeren Videos mit vielen Szenen sind, und er wird dieselben Bewegungen der Klassifikation »schneiden« zuordnen, auch wenn die Sprecher in den Videos statt »schneiden« von »schnibbeln«, »zerteilen«, »auslösen« oder »klein-«, »ab-« oder »zuschneiden« sprechen. Und er wird auch unterscheiden können, ob Gemüse, die Gartenhecke oder ein Video geschnitten wird.

Und wenn er zwischen »stürzen« und »hinknien«, »sich bücken« oder »sich setzen« differenzieren kann, ist er vielleicht reif als diskreter Helfer im Hausnotdienst. Weitere Anwendungen könnten das autonome Fahren sein, wo er zur Vermeidung von Unfällen beiträgt, oder auch die Wissenschaft, die er bei der Auswertung von Verhaltensstudien unterstützt.

Soweit sind das gute Aussichten für die Zukunft. Aber wird diese Technik nicht auch dazu führen, dass man uns noch besser überwachen kann, als das jetzt schon der Fall ist? »Überwachung ist erst einmal nichts Negatives«, findet Hilde Kühne. »Überwachung ist meiner Meinung nach erst einmal neutral. Man kann sie natürlich missbrauchen wie nahezu jede Technologie, das sollten wir gut im Auge behalten. Doch gerade Bewegungserkennung könnte wie zum Beispiel beim Assistant Living helfen, die Privatsphäre zu schützen. Wenn man auf U-Bahnhöfen eine Gefahrensituation wie eine Schlägerei erkennen möchte, kann der Computer helfen, gewisse Aktivitäten höher zu priorisieren als zum Beispiel das Bild von einem Bahnsteig mit Kindern bei einem Abklatschspiel. Denn niemand kann sich sämtliche Überwachungsvideos dauernd ansehen. Die Idee ist also nicht, dass der Computer die Weltherrschaft übernimmt, sondern dass wir dadurch, dass wir in der Lage sind, große Datenmengen automatisiert zu verarbeiten und zu filtern, für den Menschen Entscheidungen leichter machen und ihnen so ermöglichen, bessere Entscheidungen zu treffen.« ●

Der Autor

Markus Bernards, Jahrgang 1968, ist promovierter Molekularbiologe, Wissenschaftsjournalist und Redakteur von Forschung Frankfurt.

bernards@em.uni-frankfurt.de