Foto: MPI für Informatik

Der digitale **Bildreporter**

Die Hörfassung eines Films lässt Blinde die Handlung verstehen. Könnte nicht ein Computer diese Übersetzungsarbeit leisten? **Anna Rohrbach**, Wissenschaftlerin am **Max-Planck-Institut für Informatik** in Saarbrücken, und ihr Mann **Marcus Rohrbach**, bis vor Kurzem ebenfalls am Institut beschäftigt, arbeiten genau daran. In Zukunft soll ein Rechner automatisch Filmbeschreibungen generieren und vorlesen.

TEXT **TIM SCHRÖDER**

Der Pianist“, „Gandhi“, „Men in Black“, „X-Men“. Anna Rohrbach besitzt ziemlich viele Videos und jede Menge Blockbuster. Gut 200 DVDs stehen säuberlich aufgereiht in ihrem Büroregal. Während die meisten anderen Menschen Videos sammeln, um sich gemütliche Fernsehabende auf dem Sofa zu machen, bedeuten die Filme für Anna Rohrbach vor allem einen Haufen Arbeit.

Anna Rohrbach ist Informatikerin. Gemeinsam mit ihrem Mann Marcus versucht sie, dem Computer etwas beizubringen, das zunächst unmöglich klingt: Videos anzuschauen und zu beschreiben, was auf dem Bildschirm passiert. Für den Menschen ist das trivial. „Schatz, komm mal schnell, jetzt wird es spannend“, hat wohl jeder schon einmal durch die Wohnung gerufen. Wenn der Gangster im Film die Waffe hebt oder die Polizei den Killer durch dunkle Gassen jagt, dann weiß der Mensch, was abgeht.

Aber ein Computer? Der muss zunächst einmal erkennen können, dass eine Pistole, die jemand in der Hand hält, eine Waffe und keine Fernbedienung ist, dass eine Umarmung nichts mit Nahkampf zu tun hat oder dass es beim Sportfechten nicht um Leben und Tod geht. Schon das ist eine Herausforderung. Dazu muss die bewegte Szene in eine verständliche und grammatikalisch saubere Sprache übersetzt werden.

Anna und Marcus Rohrbach sind Spezialisten für „Computer Vision“, für automatische Bilderkennung. Auf diesem Fachgebiet gab es in den vergangenen zehn Jahren große Fortschritte. >

Der Videokoch: Marcus Rohrbach hat am Max-Planck-Institut für Informatik eine Küche eingerichtet und mit Videokameras ausstaffiert. Die Kochszenen, die er hier dreht, kann ein von ihm entwickeltes Computerprogramm beschreiben.

Lernende Software: Marcus Rohrbach hat dem Computerprogramm beigebracht, verschiedene Tätigkeiten in der Küche zu erkennen, indem er Helfer die Szenen zunächst beschreiben ließ. Hier assistiert ihm die Doktorandin Siyu Tang.

Computer können heute auf Fotos Gesichter erkennen und verschiedenen Personen zuordnen. Auch Landschaftsaufnahmen können sie richtig interpretieren. Rotes Licht, Segel, horizontale Linien? Na sicher: ein Sonnenuntergang am Meer. „Eine bewegte Filmszene korrekt in klaren Worten zu beschreiben ist aber etwas ganz anderes“, sagt Anna Rohrbach.

BILDBESCHREIBUNGEN FÜR BLINDE SIND EINE ANWENDUNG

Die Wissenschaftlerin forscht am Saarbrücker Max-Planck-Institut für Informatik. Auch Marcus Rohrbach hat dort bis vor Kurzem gearbeitet, ist jetzt aber für eine Postdoc-Stelle an die University of California in Berkeley gewechselt. Doch hält er einen Draht zu den Kollegen nach Saarbrücken. Das Projekt beschäftigt aber nicht nur die Rohrbachs. Die Idee entstand nämlich aus einer Zusammenarbeit zwischen der Max-Planck-Arbeitsgruppe von Bernt Schiele, in der Anna und Marcus Rohrbach arbeiten beziehungsweise gearbeitet haben, und dem Fachbereich Computational Linguistics der Universität des Saarlandes, der von Manfred Pinkal geleitet wird.

Den Forschern schweben gleich mehrere Anwendungen vor. Zukünftig könnte der Computer Filmbeschreibungen für Blinde automatisch generieren – und vorlesen. Heute ist das noch recht aufwendig, weil die Offstimme für einen Film von einem Profi eingesprochen werden muss. Anwendungsfall Nummer 2 besteht darin, Videos auf Internetportalen automatisiert zu beschreiben. Anhand solcher Kurztexte könnten Internetnutzer schneller relevante Videos finden, ohne sich wie bisher durch etliche Filmchen klicken zu müssen, bis sie endlich das passende gefunden haben.

Etwas futuristisch mutet Anwendung Nummer 3 an. Wenn ein Computer Filmszenen interpretieren und in Worte fassen kann, versteht er auch Geschehnisse in der realen Welt und kann sie entsprechend in Worten wiedergeben. Daher halten es die Rohrbachs für möglich, dass Serviceroboter oder Handy-Apps schon in einigen Jahren menschliche Handlungen begreifen und sich in natürlicher Sprache mit dem Menschen unterhalten. Sie könnten beispielsweise beantworten, wo der Besitzer seine Brille hat liegen lassen, oder mit ihm diskutieren, was er zum Abendessen kochen sollte – weil sie ja beobachtet haben, was in den Tagen zuvor aufgetischt wurde.

Marcus Rohrbach hat vor etwa fünf Jahren damit begonnen, dem Computer das Beschreiben von Filmen beizubringen – und sich dem großen Ziel in kleinen Schritten genähert. „Man kann ja nicht erwarten, dass eine Software sofort die ganze Welt mit sämtlichen vorstellbaren Szenen erkennt“, erklärt der Wissenschaftler. „Wir haben uns deshalb zunächst auf eine überschaubare Szene beschränkt – auf eine Küche, in der wir Personen beim Kochen gefilmt haben.“ Dafür ließ Marcus Rohrbach im Max-Planck-Institut eigens eine moderne Küche mit Ceranfeld und schicken Einbauschränken einrichten.

Der Unterschied zur Küche zu Hause ist, dass einige Kameras aufnehmen,

Foto: MPI für Informatik

» Der wichtigste Schritt: Marcus Rohrbach musste das Wissen über Bewegungen und Objekte mit Beschreibungen der Aktivitäten verknüpfen – ein komplexer Vorgang, der in mehreren Stufen abläuft.

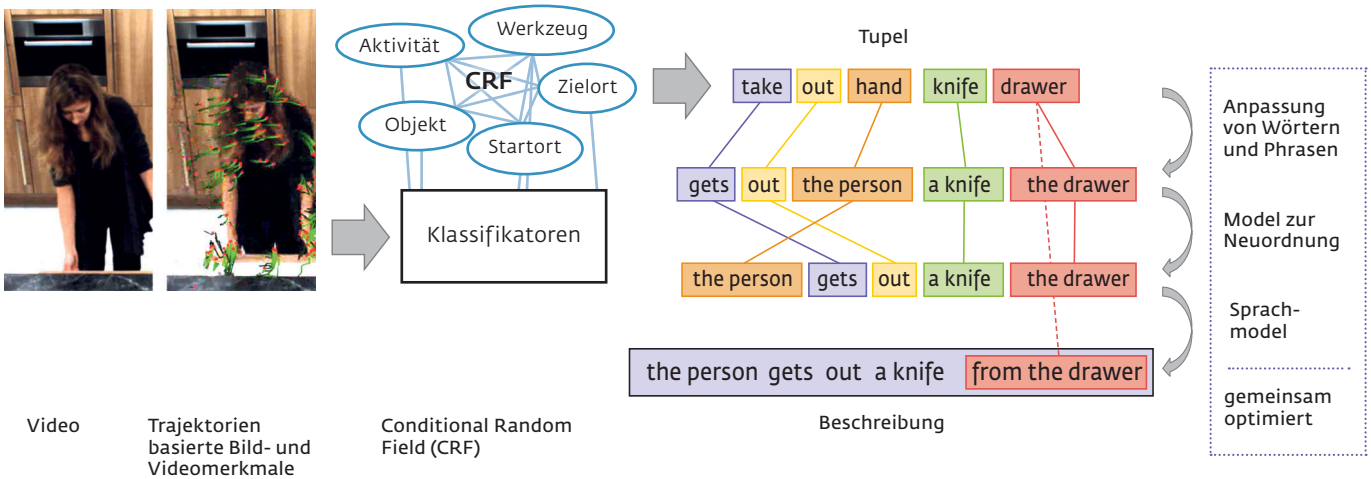
was passiert. Im ersten Schritt filmte er Probanden bei verschiedenen Tätigkeiten – dem Schälen einer Orange, beim Kochen von Spaghetti oder beim Gurkeschneiden. Diese Filmsequenzen ließ er anschließend von Helfern mit natürlichen Worten beschreiben – zum Beispiel: „Ein Mann steht in der Küche und schneidet eine Gurke mit dem Messer."

Da diese Beschreibungen frei und ohne eine feste Struktur sind, wurden die Daten zusätzlich nach einem festen Muster mit Bemerkungen versehen – annotiert. So mussten die Helfer Informationen zu den folgenden Kategorien notieren: Objekt (etwa Gurke), Aktivität (zum Beispiel Schälen oder Schneiden), Werkzeug (Messer), Ort (Tischplatte) und Ziel (Salatschüssel). „Diese Kategorien sind essenziell, wenn man eine Aktivität umfassend beschreiben will", erklärt Marcus Rohrbach, „denn sie enthalten die wesentlichen Satzbestandteile wie etwa Verb oder Objekt; zum Beispiel: Mann – Messer – schneiden – Gurke."

EINE SOFTWARE VERFOLGT DIE BEWEGUNGEN

Doch bevor ein Computer Objekte beschreiben kann, muss er zunächst lernen, wie diese aussehen. Zu diesem Zweck setzte Marcus Rohrbach Software ein, die verschiedene Parameter automatisch lernt. Dabei handelt es sich um Algorithmen, die mit einem Trainingsdatensatz gefüttert werden – in diesem Falle den Videofilmen aus der Küche. Der Algorithmus lernt nach und nach, wie ein Objekt aussieht, und kann es später wiedererkennen. Am Ende ermittelt er Wahrscheinlichkeitswerte, etwa: „Das Objekt entspricht zu 94 Prozent einer Banane."

Zum Erkennen einer Videosequenz gehört es natürlich auch, Bewegungen korrekt zu erfassen und zu interpretieren. Eine Hand, die sich rhythmisch bewegt, könnte schneiden oder eine Möhre schälen oder Eischnee schlagen. Der Computer muss all das unterscheiden können. Um ihm Bewegungen beizubringen, nutzte Marcus Rohrbach eine sogenannte Trackingsoftware, eine Verfolgungssoftware. Diese kann in einem Videobild die Bewegung einzelner Pixel nachverfolgen und damit den gesamten Bewegungsablauf gewissermaßen einfrieren. Auch diese Trackingdaten speiste der Forscher in die Algorithmen ein, sodass der Computer Schneiden oder Schälen zu unterscheiden lernte. >



In einem Video ermittelt die Software von Marcus Rohrbach zunächst Bild- und Videomerkmale entlang der Trajektorien (Bewegungsbahnen). Klassifikatoren identifizieren dann Objekte, Aktivitäten, Werkzeuge sowie Start- und Zielort. Diese Parameter setzt ein Wahrscheinlichkeitsfeld, englisch Conditional Random Field (CRF), miteinander in Beziehung. So entsteht ein Tupel, dessen Wörter und Phrasen zunächst an übliche Formulierungen angepasst und dabei etwa mit Artikeln versehen werden. Die Begriffe werden dann neu geordnet und schließlich von einem Sprachmodell um die fehlenden Präpositionen zur endgültigen Beschreibung ergänzt.

Grafik: MPI für Informatik

Links Anna Rohrbach hat gut 200 DVDs gesammelt, um einer Software beizubringen, Videos beliebigen Inhalts zu beschreiben.

Rechts Um eine Szene korrekt zu beschreiben, muss eine Software auch den Subtext der Bilder erfassen. Sonst hält sie eine Umarmung für einen Ringkampf oder umgekehrt und kann auch nicht zwischen Sportfechten und einem Duell unterscheiden.

„Wir sprechen bei solchen Algorithmen von Klassifikatoren", sagt Marcus Rohrbach. Je nach Wahrscheinlichkeitswert gewichten diese verschiedene Möglichkeiten und wägen ab, um welche Handlung – Schneiden oder Rühren etwa – oder welches Objekt es sich handelt – eine Gurke oder Banane. Dabei muss der Klassifikator schon bei der Identifikation von Objekten eine Fülle von Merkmalen berücksichtigen wie Farbe, Form oder Größe.

EIN WAHRSCHEINLICHKEITSFELD BEWERTET DIE TÄTIGKEIT

Außerdem ist es wichtig, das Zusammenspiel verschiedener Objekte und Aktivitäten zu modellieren. Zum Beispiel ist es unwahrscheinlich, dass jemand eine Gurke im Topf mit einem Löffel schält; man erwartet eher, dass jemand Zucchini im Topf mit einem Kochlöffel umrührt, auch wenn beide Szenarien auf den ersten Blick visuelle Ähnlichkeiten haben.

Um welche Bewegung oder Tätigkeit es sich am wahrscheinlichsten handelt, bewertet Marcus Rohrbach in einem sogenannten Conditional Random Field, einem Wahrscheinlichkeitsfeld. Das Conditional Random Field stellt Beziehungen zwischen Objekt, Aktivität, Werkzeug und Ort her – es bildet eine Gruppe von Parametern, die Fachleute als Tupel bezeichnen; in diesem Falle ein Objekt-Aktivität-Werkzeug-Ort-Tupel. Auch das Conditional-Random-Field-Modell wird mithilfe von Trainingsdaten angelernt.

Dann folgte der wichtigste Schritt. Marcus Rohrbach musste das Wissen über Bewegungen und Objekte mit Beschreibungen der Aktivitäten verknüpfen – ein komplexer Vorgang, der in mehreren Stufen abläuft. Zunächst erkennt der Klassifikator die Wahrscheinlichkeit einzelner Elemente. Wenn eine Person eine Zwiebel auf das Schneidebrett legt, handelt es sich für den Klassifikator mit einer hohen Wahrscheinlichkeit um die folgenden Elemente: „Hand", „legen", "stellen", „Zwiebel", „Brett", „Tischplatte". Begriffe mit geringen Wahrscheinlichkeiten wie „Löffel" oder „Topf" schließt der Klassifikator aus. Dann berechnet das Conditional Random Field, welches Tupel die Szene am besten beschreibt, in diesem Fall etwa: Hand, legen, Zwiebel, Brett.

„Um dann aus den Tupeln natürliche Sprache zu erzeugen, haben wir Software verwendet, wie man sie ähnlich von Übersetzungsprogrammen kennt, die beispielsweise vom Englischen ins Deutsche übersetzen", sagt Marcus Rohrbach. Diese bringen die zu einem Tupel verknüpften Begriffe zunächst in eine sinnvolle Reihenfolge wie etwa: „Hand legt Zwiebel auf Brett."

Anschließend werden die Begriffe nach einem sogenannten Sprachmodell durch Artikel und eventuell fehlende Präpositionen ergänzt, sodass sich ein semantisch sinnvolles Konstrukt ergibt, ein Satz mit vernünftigem Aufbau wie: „Die Hand legt die Zwiebel auf das Brett." Zudem werden bestimmte Begriffe durch üblichere Formulierungen ersetzt, die dem Sprachmodell eher vertraut sind – etwa „Hand" durch „Person". So entsteht Rechenschritt für Rechenschritt eine grammatikalisch korrekte Formulierung wie: „Eine Person legt eine Zwiebel auf das Brett."

DETAILLIERT – ODER SCHLICHT ZUSAMMENGEFASST

„Über das Küchenprojekt habe ich vor einiger Zeit meine Doktorarbeit geschrieben", erzählt Marcus Rohrbach. „Dieses Verfahren zur Videobeschreibung hat ziemlich gut funktioniert und die Szenen korrekt in Sprache übersetzt." Anna Rohrbach hat es dann so

Foto: Oliver Dietze



Fotos: gruizza/istockphoto, PaStarr/shutterstock, katatonia82/shutterstock (von links nach rechts)

erweitert, dass es Szenen unterschiedlich detailliert oder abstrahiert beschreiben kann, was vorher noch keiner anderen Forschergruppe gelungen ist. So ist die Methode in der Lage, einzelne Arbeitsschritte wie: „Eine Frau holt Spaghetti aus dem Schrank, nimmt einen Topf aus der Schublade und füllt ihn mit Wasser“ detailliert aufzuzählen oder die Tätigkeit schlicht in einem einzigen Satz zusammenzufassen: „Eine Frau kocht Spaghetti.“

Doch hatte dieses erste Projekt seine Grenzen, sagt Marcus Rohrbach. Immerhin war das Videoanalysesystem auf das Umfeld Küche beschränkt. Zudem empfand er das ganze System als zu komplex. Der Weg von der Szenenanalyse über die Tupel und die semantische Verknüpfung der Begriffe bis hin zum fertigen Satz erschien ihm zu weit. „Deshalb haben wir uns zwei neue Ziele gesetzt: Wir wollen Szenen aus jedem beliebigen Umfeld analysieren können und außerdem den Weg von der Szenenanalyse zur Sprachausgabe auf einen Schritt reduzieren.“

An dieser Stelle kommt Anna Rohrbachs inzwischen stattliche Filmsammlung ins Spiel. Bis heute hat sie 202 Videofilme und 118 000 Videoclips analysiert. Jeder dieser Clips hat etwa eine Beschreibung mit meist einem natürlichen Satz. Mit diesen Datensätzen trainiert sie ein ganz besonderes Software-Werkzeug – ein Long Short-Term Memory (LSTM).

DREI KLASSIFIKATOREN ERKENNEN EINE SZENE

Dabei handelt es sich um ein künstliches neuronales Netz, das wie alle Varianten dieser Software die Funktionsweise des Gehirns nachahmt. Ein LSTM erinnert sich jedoch über einen längeren Zeitraum an bereits verarbeitete Daten als andere künstliche neuronale Netzwerke und verarbeitet die Eingabedaten daher auch zuverlässig, wenn die entscheidenden Signale wie bei der Erkennung von Szenen oder Sprache in unregelmäßigen Abständen eintreffen.

Sofern man ein solches LSTM ordentlich mit Trainingsdaten gefüttert hat, kann es aufgrund seiner Erfahrung selbst darüber entscheiden, welche Information relevant ist und im System gespeichert werden muss oder welche gelöscht werden kann. Das LSTM kann damit die Relevanz von Informationen einschätzen. LSTM werden heute oft für die Übersetzung von Sprache oder die Erkennung von Handschrift eingesetzt.

Ein LSTM ist das Herz von Anna Rohrbachs Arbeit. Es verknüpft unmittelbar die visuelle Information, den Input, mit der Sprachanalyse – und reduziert die Videobeschreibung damit tatsächlich auf einen Schritt. Auch das LSTM arbeitet mit Wahrscheinlichkeiten. Als Input dienen ihm visuelle Daten, die wiederum von Klassifikatoren geliefert werden. Um eine Szene vollständig zu erkennen, setzt die Wissenschaftlerin drei Klassifikatoren ein. Diese geben Auskunft über folgende drei Aspekte: die Tätigkeit, die Objekte im Bild und den Ort, an dem sich die Szene abspielt.

Anna Rohrbach greift dabei durchaus auf Entwicklungen anderer Arbeitsgruppen zurück – etwa einen Klassifikator, den Forscher des Massachusetts Institute of Technology in den USA entwickelt haben. Der Klassifikator wurde mit vielen Daten darauf trainiert, Umgebungen zu erkennen – eine Küche, ein Schlafzimmer oder ein Restaurant etwa. Wie gehabt, liefern die Klassifikatoren Wahrscheinlichkeitswerte, die zu einem Wahrscheinlichkeitsvektor – zu einer Art Wolke von Wahrscheinlichkeitswerten – verknüpft werden, ehe sie in das LSTM eingespeist werden.

Das LSTM generiert aus dieser visuellen Information unmittelbar die

Das neuronale Netzwerk (LSTM) von Anna Rohrbach beschreibt Videosequenzen wie etwa eine Tanzszene genauer als andere Computerprogramme, aber noch nicht so gut wie die Worte eines Menschen.

Sprachdaten. „Eine Stärke des LSTM ist, dass es aus einer Sequenz von Worten auf die folgenden Worte schließen kann“, sagt Anna Rohrbach. Es kann sehr gut entscheiden, welches Wort auf ein Wort folgen muss, und irrelevante Daten aussortieren. Das LSTM fügt Artikel und Präpositionen hinzu und liefert so sinnvolle, natürliche Sprache.

„Im Grunde ist das wie beim Menschen. Wir merken uns ja auch, welche Worte wir gerade gesagt haben, und formulieren auf dieser Basis den nächsten Satzteil.“ Anna Rohrbachs LSTM hat also so etwas wie ein Sprachgefühl. Es benötigt keine Tupel mehr, mit deren Hilfe Worte zunächst aneinandergereiht und dann Stück für Stück zu einem vollständigen Satz zurechtgeschoben werden müssen.

Letztlich entscheidet das LSTM anhand von Wahrscheinlichkeiten, welches Wort als nächstes folgt. Und das funktioniert offenbar sehr gut. In einem Vergleich mit anderen Videobeschreibungsmethoden schnitt Anna Rohrbachs Verfahren am besten ab. Unter anderem konnte das LSTM eine Szene treffender und differenzierter beschreiben als die übrigen Methoden.

DAS LSTM ANALYSIERT GENAUER ALS ANDERE VERFAHREN

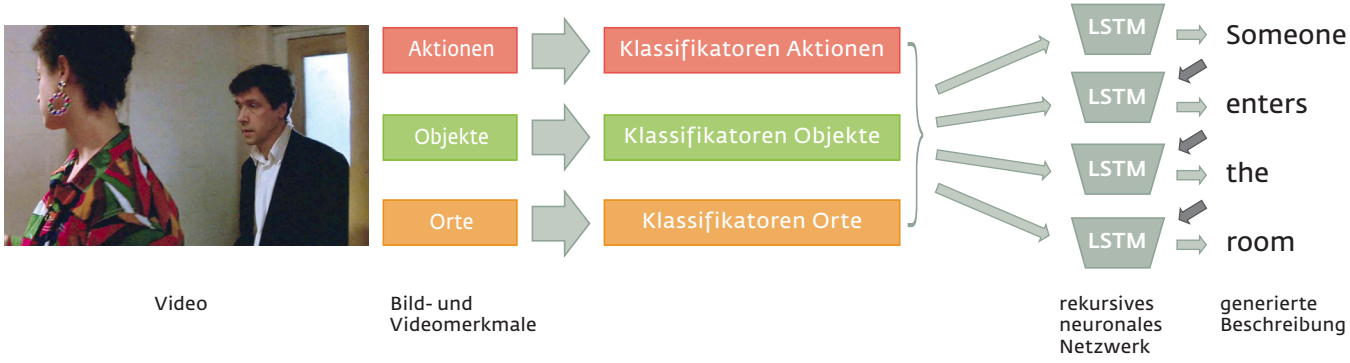
Ein Beispiel: In einer Filmszene führt jemand eine blonde Frau in weißem Kleid auf die Tanzfläche und beginnt, sie zu drehen. Anna Rohrbachs LSTM beschrieb die Szene so: „Jemand trägt ein weißes Kleid, lächelt und hat weißes Haar.“ Eine andere Software interpretiert die Szene deutlich dürftiger: „Jemand schaut jemanden an.“ Die Software des dritten Teams, mit der Anna Rohrbach ihr LSTM verglichen hat, liefert gar ein unfreiwillig komisches Ergebnis für die gegenseitigen Blicke: „Jemand schaut jemanden an. Jemand schaut jemanden an.“

Der Vergleich macht deutlich, dass das LSTM die Szene genauer analysiert als andere Verfahren. Zugleich offenbart das Beispiel, dass Anna Rohrbachs System noch Schwächen hat. Denn dass es sich um eine Szene im Tanzsaal handelt, verrät das LSTM nicht. „Tatsächlich arbeitet das Verfahren noch nicht hundertprozentig. Immer wieder gibt es grammatikalische Fehler. Und in manchen Fällen werden vor allem komplexe Szenen nicht richtig erkannt“, sagt die Forscherin.

Ein Beispiel ist eine Sequenz, in der ein Jugendlicher in Sportkleidung davonläuft. Für Blinde wurde die Szene manuell wie folgt beschrieben: „Jemand läuft in Sportkleidung eine Tribüne hoch und dann davon.“ Das LSTM interpretiert: „Jemand läuft auf der Straße.“

Vor allem abstrakte Inhalte kann das LSTM derzeit also noch nicht erkennen. Information, die sozusagen zwischen den Zeilen steckt. Die Tatsache, dass der Jugendliche davonläuft, bleibt dem LSTM verborgen, zudem ignoriert es, dass der Junge eine Tribüne hinaufrennt. „In an-

Foto: Oliver Dietze



Die Software von Anna Rohrbach lernt Aktionen, Objekte und Orte in einem Video mithilfe von Klassifikatoren, die jeweils für eine dieser Kategorien spezialisiert sind, zu erkennen. Aus diesen Bildmerkmalen erzeugt ein rekursives neuronales Netzwerk (LSTM) in mehreren Zyklen Wort für Wort eine Beschreibung des Videos.

deren Fällen konnte das System nicht erkennen, dass eine Person vor der Polizei davonläuft", sagt Anna Rohrbach.

„Es ist schwierig, dem Computer beizubringen, solche inhaltlichen Bezüge herzustellen." Genau das aber will Anna Rohrbach in nächster Zeit erreichen. Interessant wäre es für sie auch, den Computer zu lehren, die Emotionen der Schauspieler zu deuten. Denn damit ließe sich eine ganz neue Ebene erreichen, welche die Videobeschreibung noch deutlich verbessern könnte.

Wann ihre Videobeschreibung marktreif ist, kann sie noch nicht genau abschätzen. „Doch die Fortschritte in der Bilderkennung waren in den vergangenen Jahren sehr beachtlich. Manchmal geht es also sehr schnell", sagt Rohrbach. Festlegen will sie sich aber nicht. Der Gewinn für die Nutzer wäre riesig. Videos ließen sich im Handumdrehen für Blinde betexten. Und Internetnutzer könnten die Inhalte von Videos in Windeseile überfliegen – kurz und knapp – „Eine Frau kocht Spaghetti" – oder im Langtext mit allen Details. ◂

AUF DEN PUNKT GEBRACHT

- **Im vergangenen Jahrzehnt gab es im Bereich „Computer Vision", der automatischen Bilderkennung, große Fortschritte. So etwa können Rechner heute auf Fotos Gesichter erkennen und verschiedenen Personen zuordnen.**
- **Die Beschreibung von Filmszenen dagegen ist viel komplexer.**
- **Dennoch sollen Computer zukünftig automatisch Filmbeschreibungen erzeugen und diese vorlesen.**
- **Um dieses Ziel zu erreichen, arbeiten Forscher am Max-Planck-Institut für Informatik mit einem besonderen Softwarewerkzeug, dem Long Short-Term Memory (LSTM).**

GLOSSAR

Algorithmus: Eindeutige Handlungsvorschrift zur Lösung eines Problems oder einer Klasse von Problemen. Algorithmen bestehen aus endlich vielen Einzelschritten und lassen sich zur Ausführung etwa in einem Computerprogramm implementieren.

Computer Vision: Der Begriff bedeutet „maschinelles Sehen" und beschreibt die computergestützte Lösung von Aufgabenstellungen, die sich an den Fähigkeiten des menschlichen visuellen Systems orientieren. Anwendungen liegen etwa in industriellen Produktionsprozessen oder in der Verkehrstechnik.

Long Short-Term Memory (LSTM): Ein künstliches neuronales Netz, das die Funktionsweise des Gehirns nachahmt und sich über einen vergleichsweise langen Zeitraum an bereits verarbeitete Daten erinnert. Mit Trainingsdaten gefüttert, kann ein LSTM schließlich selbst entscheiden, welche Information relevant ist und im System gespeichert werden muss.

Grafik: MPI für Informatik