



Im Auge der künstlichen Intelligenz – wie Maschinen Bilder verstehen und erzeugen

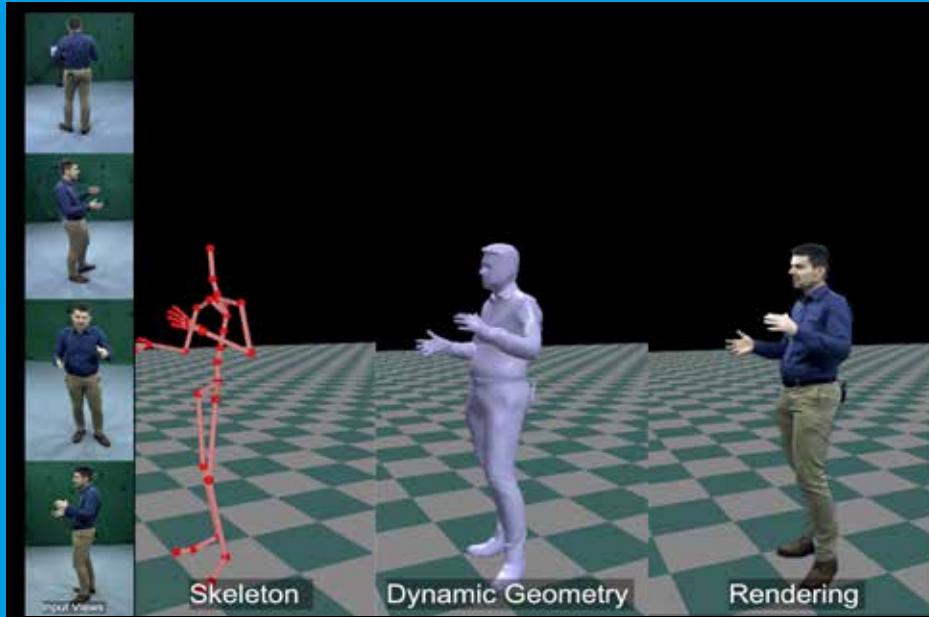
Bilder, die nicht der Wirklichkeit entsprechen, sind allgegenwärtig, zum Beispiel in Kinofilmen, auf Werbeplakaten oder im Internet. Zunehmend bearbeiten oder erzeugen künstliche Intelligenzen Bilder – mit Chancen und Risiken. Der Arbeitsgruppe von Christian Theobalt am Max-Planck-Institut für Informatik gelingt es so beispielsweise, den Gesichtsausdruck einer Person in einem Quellvideo auf eine Person in einem Zielvideo zu übertragen. Der Nutzen solcher Techniken für die Filmindustrie oder für virtuelle Treffen liegt auf der Hand. Die Risiken aber auch: gefälschte Medieninhalte (Deepfakes) können für einzelne Personen oder für die ganze Gesellschaft gefährlich werden.

Der Avatar folgt den Bewegungen Theobalts exakt, und das in Echtzeit. Während der Wissenschaftler seinen Vortrag hält, spricht, gestikuliert und bewegt sich auch sein virtueller Doppelgänger. Neben dem realistischen Abbild des Wissenschaftlers zeigt der Bildschirm parallel auch zwei einfache Modelle (Abb. A). Diese sind üblicherweise nicht zu sehen, verdeutlichen aber, auf welcher Grundlage die Bewegungen des aus vier Blickwinkeln durch Kameras aufgezeichneten Wissenschaftlers auf den Avatar übertragen werden. Christian Theobalt spricht von **holo-**

portierten Charakteren, die in virtuellen Räumen zum Einsatz kommen können. Er sagt: „So könnte in Zukunft beispielsweise eine virtuelle Telepräsenz möglich sein, die es erlaubt, über große Distanzen mit Personen realistisch zu kommunizieren, ohne reisen zu müssen.“

Anhand des holoportierten Wissenschaftlers erklärt Christian Theobalt viele Facetten seiner Arbeit. Sein Ziel ist es, neue Wege zu entwickeln, die bewegte, reale Welt technisch zu erfassen und so detailgetreue virtuelle Modelle zu erstellen. Diese Modelle sollen es Computern und zukünftigen intelligenten Maschinen ermöglichen, die reale Welt zu verstehen, sicher mit ihr zu interagieren oder sie auch zu simulieren. Bislang ist es sehr aufwendig, Bewegungen technisch aufzuzeichnen und in allen Einzelheiten mittels Computergrafik wiederzugeben. Für die technische Erfassung von Bewegungen, **Motion Capture** genannt, werden meist viele Kameras und Marker kombiniert oder eine Tiefenkamera verwendet. Bei der digitalen Erzeugung von Bildern wird außerdem viel Aufwand betrieben, damit Bewegungen natürlich erscheinen oder Details wie Lichtreflexionen, Falten in der Kleidung oder die Mimik von Menschen möglichst realistisch wiedergegeben werden. Für Filme erstellen und bearbeiten Spezialisten die

ABB. A: EIN AVATAR ENTSTEHT



Das von Theobalts Team erstellte und trainierte KI-Programm kann anhand von Kamerabildern, die aus vier Blickwinkeln aufgenommen werden (links), ein virtuelles 3D-Abbild einer Person erschaffen (rechts). Dieses lässt sich dann aus jedem beliebigen Blickwinkel betrachten bzw. darstellen und in virtuellen Treffen oder Computerspielen einsetzen. Damit der Avatar realistisch und detailgetreu ist, extrahiert das Programm zunächst die 3D-Skelettpose aus den Kamerabildern. Anschließend wird eine dynamische Textur bzw. Oberfläche erstellt und schließlich der hochauflöste Avatar erzeugt.

computergenerierten Bilder, kurz CGI (engl. Computer Generated Imagery) in aufwendiger Handarbeit. Christian Theobalt will das alles wesentlich vereinfachen: „Ziel ist es, dass eine einzige Kamera ausreicht, um Bewegungen exakt zu erfassen.“ Und auch Bilder zu erzeugen oder zu verändern soll wesentlich einfacher werden. Dazu forscht Theobalts Abteilung „Visual Computing and Artificial Intelligence“ an der Schnittstelle von Computergrafik, Computer Vision und künstlicher Intelligenz. Der erwünschte Fortschritt soll durch die Kombination künstlicher Intelligenz und etablierter Ansätze der Computergrafik, wie beispielsweise der Nutzung geometrischer Modelle, erreicht werden.

MASCHINEN WERDEN INTELLIGENT

Der Begriff **künstliche Intelligenz** beschreibt Algorithmen, die dazu dienen, Maschinen intelligent zu machen. In vielen Fällen ahmen diese Algorithmen die kognitiven Fähigkeiten von Menschen nach. Ziel der Forschung und Entwicklung im Bereich der künstlichen Intelligenz ist es, Maschinen zu schaffen, die in bestimmten Bereichen der Intelligenz an den Menschen heranreichen oder diesen sogar übertreffen. Eine gebräuchliche Abkürzung für künstliche Intelligenz ist KI. Im englischsprachigen Raum wird von „artificial intelligence“ gesprochen und manchmal wird die daraus folgende Abkürzung AI auch im Deutschen verwendet.

Die menschliche Intelligenz zeichnet sich dadurch aus, dass das Gehirn unseren Körper steuert, Sinneseindrücke verarbeitet und neue Informationen mit bekannten verbindet. Dadurch können wir Geschehnisse in unserer Umwelt einordnen und vorausschauend denken und handeln. Bekannte Bereiche der künstlichen Intelligenz sind die Robotik, also die Steuerung komplexer Bewegungen, und Computerprogramme, die komplexe Spiele wie Schach oder

Go meistern und dafür Informationen verarbeiten und vorausschauend agieren müssen. Eine weitere Komponente der Intelligenz ist das Sprachverständnis. Das Ziel des Forschungsbereichs der Computerlinguistik ist es, Maschinen zu entwickeln, die Sprache möglichst umfassend verstehen. Zuletzt machten auf diesem Feld sogenannte Chatbots wie ChatGPT oder Bard Schlagzeilen, aber auch Übersetzungsprogramme wie DeepL gehören zu den vielfältigen Anwendungen von KI im Bereich Sprache. Der Aspekt an künstlicher Intelligenz, der Christian Theobalt am meisten interessiert, ist das **Visual Computing**. Darunter fallen alle digitalen Methoden, die Bilder verarbeiten, analysieren, modifizieren und erzeugen. Seine Arbeit geht also über die Computer Vision hinaus, die aus visuellen Daten wie Bildern und Videos Informationen gewinnt und beispielsweise in selbstfahrenden Fahrzeugen zum Einsatz kommt.

In seiner Forschung setzt Theobalt auf **maschinelles Lernen**. Diese Art des

Lernens produziert künstliche Intelligenz, die nicht auf vorab formulierten Regeln basiert, sondern aus Beispielen lernt, wie eine Entscheidung zu treffen ist. Stehen der selbstlernenden Maschine hunderte oder besser tausende Beispiele zum Training zur Verfügung, entwickelt sie selbstständig einen Entscheidungsprozess, der verallgemeinert werden kann. Somit ist dieser anschließend auch auf unbekannte Datensätze anwendbar. Dazu nutzt Theobalts Forschungsteam das Deep-Learning-Verfahren. Dieses imitiert das menschliche Lernverhalten und basiert auf einem **neuronalen Netz**. Das Netz besteht aus künstlichen Neuronen, die in mehreren Schichten den Entscheidungsprozess gestalten (**Abb. B**). Jedes Neuron verarbeitet die eingehenden Daten, indem es die einzelnen Eingabegrößen gewichtet und gemäß bestimmter Regeln an die Neuronen der nächsten Schicht weitergibt. Nachdem moderne neuronale Netze oftmals aus vielen Schichten bestehen und damit tief sind, spricht man von **Deep Learning**.

SELBSTLERNENDE PROGRAMME

Der rasante Fortschritt der letzten Jahre auf dem Gebiet der KI basiert auf solchen selbstlernenden Programmen. Ausgelöst wurde diese Entwicklung durch Forschungserfolge im Deep Learning ab 2009 sowie immer größere verfügbare Rechenleistung und Datenmengen (Big Data), die es möglich machen, eine KI umfassend zu trainieren. So konnten Programmierende die Fähigkeiten von KI-Programmen rasch verbessern und erweitern. Beispielsweise erzielte 2015 das Deep-Learning-basierte Programm AlphaGo die ersten Erfolge einer KI gegen Weltklassenspieler beim Brettspiel Go. Im weniger komplexen Schach schaffte es dagegen schon 1997 der Schachcomputer Deep Blue, den amtierenden Weltmeister zu schlagen. Deep Blue war eine regelbasierte, sogenannte **symbolische KI**. Diese Art der KI ist nicht selbstlernend, sondern

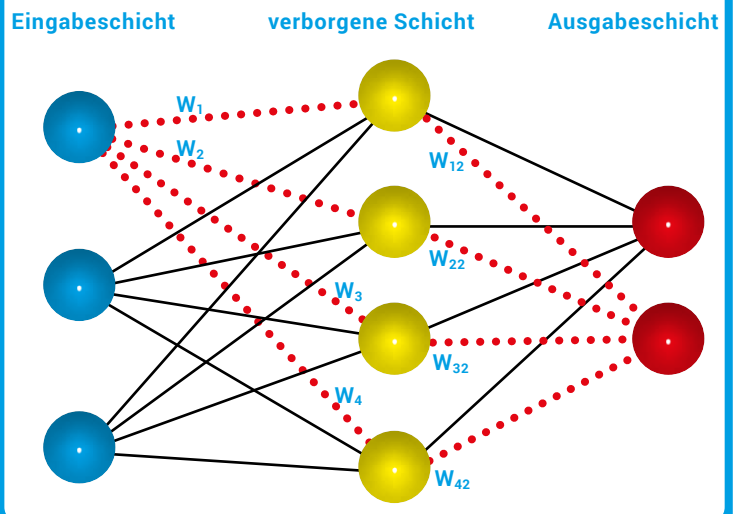
kommt zu Entscheidungen, indem sie anhand klarer, vorab im Programmcode festgelegter Regeln Symbole wie z. B. Wörter oder Ziffern kombiniert. Die rein regelbasierte KI ist allerdings stark limitiert. Denn abgesehen von Spielen wie Schach, in denen die Umgebung eindeutig definiert ist, versagt sie, da es kaum möglich ist, alle möglichen Fälle vorab durch Regeln abzudecken. Der Vorteil symbolischer KI ist, dass sie durch die Regeln und Symbole in der menschlichen Realität verankert ist und ihre Entscheidungen somit nachvollziehbar und interpretierbar sind. Im Gegensatz dazu sind die Entscheidungen selbstlernender Programme nicht per se nachvollziehbar. Christian Theobalt kombiniert in seiner Forschung regelbasierte und selbstlernende KI im sogenannten **neuro-expliziten Verfahren**. Wenn die KI etwa lernen soll, menschliche Bewegungen aus Kamerabil- dern zu rekonstruieren, nutzt sein Team ein vereinfachtes Skelett mit erlaubten Bewegungsrichtungen und -winkeln, um die Entscheidungen des Programms in realistische Bahnen zu lenken.

EFFIZIENTES TRAINING

Damit die KI später gute Entscheidungen trifft, sind die Trainingsdaten entscheidend. Dabei ist es sowohl wichtig, dass eine große Datenmenge verfügbar ist, als auch, dass diese Daten von hoher Qualität sind. Damit der Avatar von Theobalt erzeugt werden kann, posierte der Wissenschaftler vorab in einem speziellen Labor vor mehr als einhundert hochauflösenden Kameras. Für das Trainingsdatenset der **neuro-expliziten KI** wird einerseits ein statischer 3D-Scan von Theobalt mit dem vereinfachten Skelett versehen und andererseits Videomaterial aufgezeichnet, das die unterschiedlichsten Bewegungen und Körperhaltungen aus allen Blickwinkeln umfasst. Ein Teil des Videomaterials dient außerdem als Testdatenset. Die trainierte KI kann anschließend auf der Grundlage von Videomaterial aus nur vier Blickwinkeln den detailgetreuen, bewegten Avatar erstellen. „Der Avatar kann Bewegungen darstellen und Haltungen annehmen, die nicht im Trainingsdatenset enthalten sind. Und er kann aus jedem Blickwinkel betrachtet werden, also nicht nur aus den vier Kameraperspektiven der Eingangsdaten“, sagt Christian Theobalt.

Dazu startet das Programm mit den vier Kamerabildern, der extrahierten 3D-Skelett-Pose und den Kameraparametern (Abb. A). Das auf Basis des Trainingsdatensets erstellte neuronale Netz für das Charaktermodell nimmt die Skelett-Bewegung als Eingabe und sagt eine positionsabhängige Verformung des Gitters voraus, das die Oberfläche des Charaktermodells bildet. Anschließend wird die Textur der Person soweit möglich aus den vier Kamerabildern gewonnen. Die Textur umfasst die Oberflächenbeschaffenheit und Farbe, etwa von Haut, Haaren und Kleidung. Im nächsten Schritt erstellt ein weiteres neuronales Netz aus diesen Texturinformationen eine blickwinkelabhängige, dynamische Textur. Zu guter Letzt erzeugt ein weiteres neuronales Netz aus den gesamten, niedrig aufgelösten Merkmalen die hochauflösten Bilder des Avatars. Das ganze Programm aus mehreren zusammenspielenden neuronalen Netzwerken arbeitet so schnell, dass der Avatar in Echtzeit entsteht und keine Verzögerung zwischen den Bewegungen der realen Person und dem holoportierten Charakter festzustellen ist.

ABB. B: NEURONALES NETZ



Ein einfaches Modell eines neuronalen Netzes, das für Deep Learning genutzt wird, besteht aus mehreren Schichten künstlicher Neuronen (Kugeln). Die Eingabeschicht (blaue Kugeln) nimmt die eingehenden Daten auf. Diese werden anschließend von den Neuronen in den verborgenen Schichten (hier nur eine Schicht, gelbe Kugeln) verarbeitet. Dazu werden die Daten von einem künstlichen Neuron gewichtet (Gewicht w_{xx}) und an weitere Neuronen in der nächsten Schicht weitergegeben. Das Ergebnis des Programms in der Ausgabeschicht hängt somit von vielen verschiedenen Neuronen und Gewichten ab (rote Linien).

Der Lernprozess der neuronalen Netze, die Theobalts Team dazu nutzt, läuft überwacht ab. Beim **überwachten Lernen** hat der Algorithmus eine klare Zielvorgabe und nutzt das Trainingsdatenset, um diesem Ziel immer näher zu kommen. Im Fall des Avatars werden die Ergebnisse der neuronalen Netze mit den zugrundeliegenden Kamerabildern verglichen, um eine möglichst fotorealistische Darstellung zu erreichen. Weitere wichtige Formen des maschinellen Lernens sind das unüberwachte Lernen und das bestärkende Lernen.

KÜNSTLICHES LÄCHELN

Das Gesicht und die Hände sind die Körperpartien, die am schwierigsten technisch nachzustellen sind. Doch gerade Mimik und Handgesten werden in Zukunft wichtig für die Interaktion von Menschen mit Computer- und Robotersystemen sein. Daher liegt hier auch ein Schwerpunkt von Theobalts Forschung: Sein Team arbeitet daran, mit nur einer Kamera die Bewegung von Händen oder die Details eines Gesichts erfassen zu können. Ihre Forschung zu Gesichtern zeigt, dass sich der Gesichtsausdruck einer Person in einem Quellvideo auf eine Person in einem Zielvideo übertragen lässt. Die Forschenden entwickelten beispielsweise ein Programm, das die detaillierten Bewegungen der Augenbrauen, des Mundes, der Nase und der Kopfposition aufzeichnet. Dadurch kann etwa der ganze Ausdruck eines Synchronsprechers auf den eigentlichen Schauspieler im Film übertragen werden, wodurch die Synchronisation eines Films in einer anderen Sprache wesentlich vereinfacht wird. Noch realistischer wirkt die Synchronisation durch eine weitere Entwicklung des Forschungsteams: Die stilbewahrende Lippensynchronisation überträgt die Mimik der Quellperson (Synchronsprecherin) auf den charakteristischen

ABB. C: REALISTISCHE MIMIK



Die KI-gestützte visuelle Synchronisation kann die Lippen stilbewahrend an eine neue Tonspur anpassen, indem sie die Mimik der Quellperson auf den charakteristischen Stil der Zielperson überträgt. Wird der Gesichtsausdruck dagegen direkt übertragen, gehen die Eigenheiten, die die Zielperson ausmachen, verloren. Dies wird hier beispielsweise an der Mundpartie deutlich.

Stil der Zielperson (Schauspielerin) (**Abb. C**): Dadurch passen die Lippenbewegungen zur neuen Tonspur, während die Eigenheiten, die die Schauspielerin ausmachen, erhalten bleiben. Dazu nutzen die Forschenden einen ähnlichen Ansatz wie für den holoportierten Charakter. Die neuroexplizite KI stützt sich in diesem Fall auf ein Gesichtsmodell und neuronale Netze.

KI VERANTWORTUNGSVOLL NUTZEN

Neben vielen zukunftssträchtigen Anwendungen, die solche Forschung erschließt, birgt diese Technik auch Gefahren. Mithilfe derartiger Programme ist es möglich, Medieninhalte zu fälschen, die für einzelne Personen, aber auch ganze Gesellschaften zur Gefahr werden können. Diese durch Deep Learning erzeugten Fälschungen werden Deepfakes genannt und sind ein echtes Problem: gerade in niedrig aufgelösten Videos, die in sozialen Medien kursieren, sind Fälschungen mit bloßem Auge kaum zu identifizieren. So können falsche Informationen schnell und durchaus glaubhaft verbreitet werden. Politikern oder Politikerinnen können zu Propagandazwecken falsche Aussagen in den Mund gelegt und Prominenten kann ein künstlicher Skandal angehängt werden. Letztlich kann prinzipiell jeder Mensch, von dem Video- oder Bildmaterial zugänglich ist, Opfer eines Deepfakes werden. Theobalt plädiert dafür, dass Forschende die Ausgabe ihrer Programme mit einem Wasserzeichen versehen, das es später ermöglicht, damit erzeugte Deepfakes leicht zu identifizieren. Außerdem sagt er: „Es wird immer Menschen geben, die Technik missbrauchen. Der beste Weg, um dagegen vorzugehen, ist mit dem Fortschritt Schritt zu halten und KI-basierte Programme auch dafür zu nutzen, gefälschte Bilder oder Videos aufzuspüren. Wir entwickeln mit unserer Forschung auch das mathematische Verständnis, das dazu nötig ist, Fälschungen zu detektieren.“ Aktuell ist es meist noch möglich, Deepfake-Videos selbst zu identifizieren. Doch dazu muss man sehr aufmerksam sein und auf Details wie Lippenbewegungen, Zähne und Mundinnenraum, Augenpartie oder Schattenwurf und Reflexionen achten. Allerdings werden die Algorithmen immer bes-

ser und gefälschte Videos immer schwerer von der Wirklichkeit zu unterscheiden. Forschende entwickeln daher Programme, die Deepfakes verlässlich aufdecken sollen. Diese können allerdings wiederum dazu genutzt werden, die erzeugenden KI-Programme noch besser zu machen. Ein Wettrennen findet statt. Entsprechend ist es nach Ansicht vieler Experten entscheidend, den Einsatz von KI umfassend gesetzlich zu regulieren, damit diese Technologie sicher und zum Wohl der Menschen eingesetzt wird.

Schlüsselbegriffe

Deep Learning, holoportierter Charakter, künstliche Intelligenz, maschinelles Lernen, Motion Capture, neuro-explizite KI, neuronales Netz, symbolische KI, überwachtes Lernen, Visual Computing

Link-Tipps

- **Themensammlung KI**
www.mpg.de/t341
- **Unterrichtsmaterial**
„Meinungsbildung in der digitalen Welt“
www.mpg.de/t342

Video-Tipps

- **Holoportierter Charakter**
www.mpg.de/t343
- **So lernt KI**
www.mpg.de/t344 > QR
- **Mit KI und Algorithmen gegen Hass**
www.mpg.de/t345



Audio-Tipp

- **Was kann künstliche Intelligenz?**
www.mpg.de/t346

www.max-wissen.de

der Link zur Forschung
für Lernende und Lehrkräfte



Hier finden Sie Hintergrundinformationen und Materialien für den Unterricht zu den Ausgaben von **BIOMAX**, **GEOMAX** und **TECHMAX**. Weitere Exemplare können Sie auf der Webseite kostenfrei bestellen.

CC-Lizenztexte finden Sie unter <https://creativecommons.org/licenses> und im Detail bei den einzelnen Heften auf der Webseite www.max-wissen.de.