Wie Computer die Welt sehen

Hochspezialisierte Rechner können sich ein Bild von ihrer Umgebung machen. Doch bis sie so gut sehen können wie der Mensch, wird es wohl noch dauern

Von Markus Mandau

Sehen will gelernt sein. Als Erwachsene meinen wir, diese Fähigkeit sei so natürlich wie das Atmen. Aber Kinder trainieren bis zum sechsten Lebensjahr, bis sie ihre Umgebung so gut sehen und deuten können wie wir. Allein die ersten zwei Jahre verbringen sie damit, den Blick scharf zu stellen, indem sie ständig versuchen, ihn auf Gegenstände zu fokussieren. Daneben üben sie sich darin, ein räumliches Bild von ihrer Umgebung zu bekommen. Das Grundprinzip nutzt auch das Kino mit 3D-Brillen: Das rechte und linke Auge sehen jeweils ein etwas anderes Bild. Das Gehirn lernt, aus den Unterschieden der beiden Bilder Tiefeninformation und einen räumlichen Eindruck der Umgebung zu gewinnen. Eine Umgebung allerdings, in der nur die Objekte scharf zu sehen sind, auf die wir unser Augenmerk lenken. Nach sechs Jahren ist das Training für Auge und Hirn abgeschlossen.

Computer lernen ebenfalls sehen – schon seit vielen Jahren. Mit dem Perceptron-Algorithmus startete im Jahr 1957 das erste neuronale Netzwerk, das getarnte Panzer in einem Wald finden konnte. Jedenfalls bei gutem Wetter. Heute nutzen wir etwa die Gesichtserkennung unserer Fotoverwaltung. Doch das ist eine Software, die stur Code abarbeitet. Künftig wird sie durch ein neuronales Netz auf unserem Rechner ersetzt, das wie ein Kind aus seinen Fehlern lernt und sich ständig verbessert, indem es seinen Code anpasst. Je mehr Fotos das neuronale Netz einliest, desto besser lernt es die Motive und Blickwinkel kennen, die wir bevorzugen. Es wird zu einem Foto-Partner, der Ratschläge gibt, Ordnung schafft und weiß, worauf er sein Augenmerk legen muss. Für ein solch komplexes Szenario müssen wir neuronale Netzwerke noch trainieren (siehe Seite 38), aber die letzten Jahre haben schon große Fortschritte gebracht.

Zusätzlich geben wir Computern auch Sensoren, mit denen sie ihre Umgebung und Bewegungen räumlich erfassen – das Äquivalent des kindlichen 3D-Blicks. Microsoft hat den Trend mit der Kinect-Kamera für seine Spielkonsole Xbox schon vor ein paar Jahren angestoßen. Die Kinect erkennt den Spieler, der mit seinen Bewegungen das Spiel steuert. Googles Projekt Tango und Intels neue RealSense-Kamera bringen solche Fähigkeiten nun auf Mobilgeräte. Das Tablet Dell Venue 8 7000 verschafft sich dank der RealSense-Kamera einen Eindruck von seiner Umgebung. Die erste Generation R100 speichert zu jedem mit der Tablet-Kamera aufgenommenen Foto zusätzliche Rauminformationen. Dazu hat sie drei unterschiedlich positionierte Kameras, die gleichzeitig fotografieren. Die Kamera in der Mitte schießt das eigentliche Bild. In den beiden Aufnahmen der Kameras rechts und links sucht ein Algorithmus identische Punkte, anhand derer die nachgelagerte Analyse eine Tiefenbestimmung durchführt, die sich an das Prinzip der Triangulation anlehnt (siehe rechts). Das funktioniert ab einer Entfernung von einem Meter schon recht genau, ab mehr als fünf Meter lässt die Präzision nach. Das Dell-Tablet misst in seiner Fotogalerie auf Wunsch die Abstände zwischen zwei Objekten (siehe rechts) oder bestimmt einen Flächeninhalt. Hobby-Architekten und Wohnungsrenovierer brauchen nicht länger ihr Maßband zu zücken, ein Foto reicht.

Raumgefühl als Netz aus Messpunkten

Das ist nur der Anfang. Die nächste RealSense-Generation 200 wird genauer sein, sie ermittelt die Rauminformationen mittels eines Lasers, der via Infrarotlicht ein Netz aus Messpunkten bildet. Zwei IR-Sensoren empfangen diese Daten, ein Algorithmus verknüpft die Punkte zu Flächen und errechnet ein Tiefenmodell der Umgebung, das sich mit einem Foto kombinieren lässt (siehe rechts). So wirft die Software zum Beispiel ein Netz von 78 Punkten über ein Gesicht, um seine Position und sogar grob den Gemütszustand zu erfassen – Ärger, Freude oder Trauer. Auch den Puls misst sie anhand von Farbänderungen der Gesichtshaut. Künftig werden Mobilgeräte erkennen, wie ihre Besitzer oder deren Gesprächspartner drauf sind. Und dank neuronaler Netze werden sie mit der Zeit immer besser darin.

So weit sind wir aber noch nicht. Ähnlich wie in der Kinect arbeiten die IR-Kameras von RealSense mit einer geringen Auflösung von beispielsweise 320 x 240 oder 360 x 480 Messpunkten. Das System zeichnet ein Video auf und analysiert damit Bewegungen sowie Gesten. Bei 60 Frames pro Sekunde kommt das System also auf 18 Millionen Tiefenberechnungen pro Sekunde. Intel empfiehlt, das Gerät mit der RealSense-Kamera nur langsam zu bewegen, das aufgenommene Objekt sollte am besten ganz ruhig bleiben. Diese Empfehlungen zeigen, dass die Hardware für Mobilgeräte von räumlicher Orientierung, wie sie beispielsweise das autonome Fahren erfordert, weit entfernt ist (siehe Seite 38). Zusätzliche Herausforderung: Die RealSense-Hardware soll so schrumpfen, dass sie in Smartphones passt. Ein Research-Team von Microsoft hat schon einen Weg gefunden, den Sensorenapparat zu verkleinern. Die Kameralinse wird dazu mit einem Kranz aus LEDs umgeben, die Infrarotlicht aussenden (siehe rechts). Der in Kameras eingebaute Infrarotfilter wird dann ausgebaut, damit das Licht dieser Wellenlänge empfangen werden kann. Das Depth4Free genannte System reicht aus, Gesten im Nahbereich zu erkennen; Räume kann man damit aber nicht vermessen.

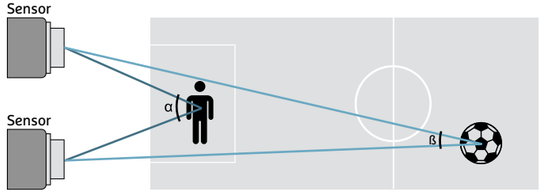
Das wiederum kann das Google-Projekt Tango, das Ende des Jahres in Consumer-Geräten Einzug halten soll. Zunächst bestückt das Raumerkennungs-System die SPHERES-Roboter der NASA. Diese fußballgroßen Kugeln sollen durch die internationale Raumstation schweben und deren Innenräume vermessen. Bisher gibt es von →

FOTOS: YOUTUBE/GOOGLE ATAP (AUFM.); INTEL CORPORATION (2, REALSENSE); MICROSOFT RESEARCH (2, GESTENERKENNUNG)

INFO

Zwei Kameras für die Entfernungsmessung

Die gebräuchlichste Methode, um den Abstand zu einem Objekt zu ermitteln, ist die Triangulation. Die Größe des Winkels aus der Sicht der beiden Sensoren bestimmt die Entfernung des Gegenstands.



Ein Tablet vermisst den Raum

Die erste Generation (R100) von Intels RealSense-Kamera steckt im Dell Tablet Venue 8 7000. Sie nutzt das obige Prinzip und kann im Nahbereich die Distanzen zwischen Gegenständen berechnen.

Tiefeninformationen mit dem Infrarotlaser

Intels neue Tablet- und Handy-Kamera RealSense R200 nutzt Infrarot-Laser und -Sensoren. Der Bildprozessor ermittelt daraus die Rauminformationen. Eine Software vervollständigt diese zu einem Tiefenbild.

Gesten erkennen mit LED-Lampen

Forscher von Microsoft haben die Raumerkennung im Nahbereich vereinfacht: LED-Lampen um eine Fotolinse senden Infrarotlicht aus, das die Kamera empfängt und eine Software auswertet.

Tango nur ein Tablet (siehe erste Seite), das mit einem ähnlichen Sensorenapparat ausgestattet ist wie RealSense oder Kinect. Zusätzlich sorgt ein Motion-Tracking-System mittels Infrarot-Messungen dafür, dass der Prototyp seine eigene Position erfasst. Wenn man mit dem Tablet durch die Wohnung läuft, zeichnet Tango den Weg auf und vermisst gleichzeitigt den Raum. Geht man den Weg wieder zurück, erkennt Tango anhand eines Datenabgleichs, dass es schon einmal hier war – und schon kennt es sich aus.

Doch warum das alles? Nun ja, in ein paar Jahren werden beispielsweise autonome Autos per Computersteuerung durch unsere Straßen fahren. Sie benötigen einen mehrstufigen Sensorenapparat, um sich in ihrer ständig wandelnden Umwelt zu orientieren. Entfernungsberechnungen mittels einer Stereokamera reichen für den Nahbereich bis etwa 30 Meter aus (siehe rechts oben) – sie können Fahrspuren und Ampeln identifizieren. Aber sie erfassen nur einen kleinen Blickwinkel zwischen 50 und 60 Grad. Ein komplettes Rundumbild von 360 Grad erhalten die autonomen Autos durch ein LiDAR-System. Das sendet Laserimpulse aus und misst, wie lange es dauert, bis sie zurückreflektiert werden. Bis zu einer Entfernung von wenigen Hundert Metern funktioniert das recht zuverlässig. Das alles reicht aber noch nicht zum autonomen Fahren: Der Bordcomputer muss die Daten mit einer detaillierten 3D-Karte der Umgebung abgleichen und benötigt zusätzliche Informationen über den Standort von Fußgängerwegen und Verkehrszeichen, um seinen Fahrweg zu berechnen. Zudem muss der Computer Objekte wie Menschen, Fahrradfahrer oder Tiere identifizieren, um zu berechnen, wohin sie sich bewegen. Das erledigen neuronale Netzwerke.

Neuronale Netze verbessern sich selbst

Neuronale Netzwerke suchen Antworten auf Fragen wie „Wann ist ein Pferd eigentlich ein Pferd?“. Dazu werden die Netzwerke mit einer Unmenge an Pferdebildern gefüttert, aus denen sie lernen, die markanten Pferdemerkmale wie Mähne, Schweif, Nase, Beine oder Hufe herauszufiltern. Haben die Entwickler ihre Netzwerke intensiv trainiert, können sie das für den Straßenverkehr möglicherweise gefährliche „Pferd“ im Prinzip so zuverlässig erkennen wie wir. Die größten Durchbrüche zur Identifizierung von Objekten haben Forscher mit Convolutional Neural Networks (CNN) erzielt (siehe rechts unten). Die zentrale Rechenoperation „Convolution“ (Faltung) legt dabei einen Filter über ein Quadrat aus Pixeln. Der Faltungsfilter vergleicht die Pixel in der Mitte des Quadrats mit den Pixeln an den Rändern und schaut, wie ähnlich die Umgebung ist.

Einen Faltungsfilter hat jeder schon einmal benutzt: Eine Bildbearbeitung wie Gimp setzt sie zum Absoften oder Schärfen von Fotos ein. Diese Operation wiederholen CNNs mehrere Male, wobei sie sich immer weiter vom eigentlichen Bildinhalt entfernen und mit jedem weiteren Filtervorgang auf eine höhere Abstraktionsebene kommen: Aus Pixeln werden Merkmale wie Linien, Bögen und Kanten, daraus ergeben sich wiederum Augen, Nasen und Beine. Das Ziel des Filters ist es, diese Merkmale möglichst deutlich herauszuarbeiten. Das CNN lässt dazu Dutzende oder gar Tausende Filter parallel durchlaufen und lernt, welche für einen bestimmten Objekttyp gut funktionieren. Am Ende der Filterprozesse arbeitet das neuronale Netzwerk mit immer gröberen Strukturen, bis es in der letzten Phase eine Entscheidung trifft: klar, ein Pferd.

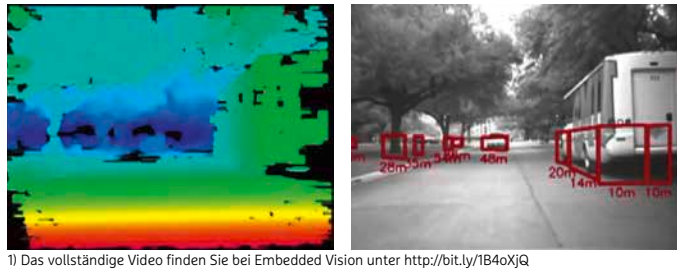
Führende CNN-Experten arbeiten an Universitäten wie Stanford, doch sie landen meist schnell bei den großen Serviceanbietern Google, Facebook und Microsoft, die ihre gigantischen Bildersammlungen von CNNs durchkämmen und katalogisieren lassen. So verkündeten Forscher der der Uni Stanford und der Yahoo Labs im →

FOTOS V. O.: EMBEDDED VISION; VELODYNE LIDAR; STANFORD

INFO

Autofahren auf 3D-Sicht

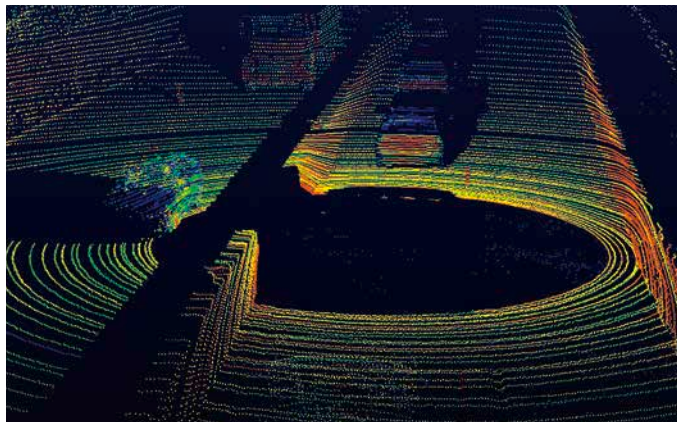
In autonomen Autos misst eine Stereokamera, wie weit die Hindernisse im Nahbereich bis 30 Meter entfernt sind. Ein Mess-System wie hier von Texas Instruments[1)] berechnet daraus die exakten Abstände.



1) Das vollständige Video finden Sie bei Embedded Vision unter http://bit.ly/1B4oXjQ

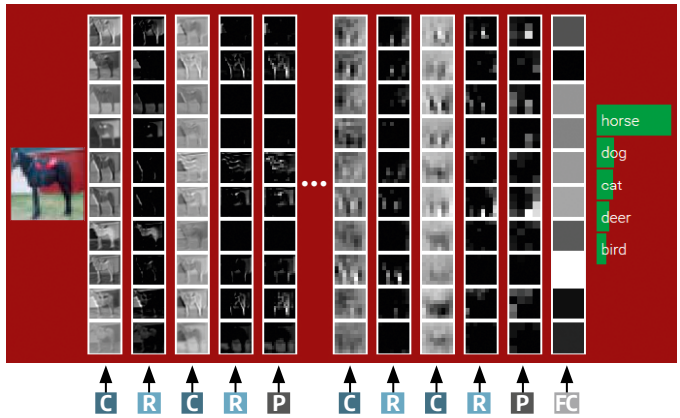
Straßenverkehr mit Lasermessung erkennen

Per LiDAR-Messung erfasst das Auto die weitere Umgebung in einer 360-Grad-Ansicht. Dazu misst es, wie lange der ausgesandte Laserimpuls braucht, bis er wieder zum LiDAR zurückreflektiert wird.



Neuronale Netzwerke identifizieren Objekte

Das Netzwerk unten stößt zehn parallele Vorgänge in mehreren sich wiederholenden Schritten an, um ein Objekt zu identifizieren. Es arbeitet sich von den Details (Linien, Punkte) zu immer gröberen und größeren Strukturen vor (Beine, Kopf), bis es das Pferd erkennt.



C Convolution Layer: Ein Fotofilter hebt jeweils die Umgebung eines Pixels hervor und macht typische Merkmale sichtbar.

R ReLu-Layer: Rectified Linear Units aktivieren etwa anhand eines Schwellwerts die wichtigen Bereiche (helle Stellen in der R-Schicht).

P Pooling-Layer: Die Auflösung wird reduziert, um die wichtigen Bereiche zu verstärken. Mit dem Resultat startet der nächste Zyklus.

FC Fully Connected Layer: Abschließend wird das Ergebnis mit denen der anderen Vorgänge verglichen, bevor es zur Entscheidung kommt.

Februar einen Durchbruch bei der Gesichtserkennung: Ihr CNN identifiziert Gesichter aus jedem Winkel, auch wenn sie teilweise verdeckt sind. Dazu haben sie es mit 200.000 Bildern gefüttert, auf denen Gesichter zu sehen sind. Hinzu kamen 20 Millionen Fotos ohne ein Gesicht als Gegencheck. Das CNN verarbeitete in 50.000 Durchgängen jeweils Schübe von 128 Bildern ehe das Training abgeschlossen war. Facebook hat sogar verkündet, dass sein DeepFace-Netzwerk in 97,25 Prozent aller Fälle ein Gesicht korrekt erkennt und damit nur wenige Prozentpunkte hinter dem Schnitt eines Menschen liegt.

Diese Erfolgsmeldungen sollte man relativieren. „Auch Neuronale Netze haben Schwierigkeiten, wenn sie einzelne Objekte in einem komplexen Bild identifizieren müssen", meint Dr. Mark Asbach, der über Gesichtserkennung promoviert hat und Projektleiter beim Fraunhofer Institut war. Aktuell beschäftigt er sich mit dem Projekt Pixolus, das Objekterkennung in einer App einsetzt. Findet sich ein Objekt nur in einem kleinen Ausschnitt, muss das neuronale Netz mitunter Millionen Ausschnitte pro Bild bewerten. „Dabei explodiert die Zahl der möglichen Fehlentscheidungen", so Asbach.

Umgekehrt erzielen CNNs ziemlich schnell gute Resultate, wenn sie zu einem bestimmten Zweck trainiert werden und dafür nur eine begrenzte Anzahl von Parametern wichtig ist. Forscher der Rutgers Universität in New Jersey haben ein neuronales Netz aufgebaut, das Maler und ihre Malstile identifiziert. In 60 Prozent der Fälle konnte das Netzwerk den Künstler eines Bildes korrekt erkennen und in rund der Hälfte aller Fälle ihren Malstil einordnen. Microsoft hat Ende April unter **how-old.net** ein neuronales Netzwerk online gestellt, das versucht, das Alter von Personen zu schätzen (siehe rechts). Leider überzeugen die Ergebnisse noch nicht. Damit das geschieht, müsste Microsoft eine Feedback-Funktion einbauen, denn neuronale Netze werden besser, je mehr man sie trainiert.

Spezial-Chips für Smartphones kommen

Künftig werden CNNs nicht auf die großen Webdienste beschränkt bleiben. Ist das rechenintensive Training abgeschlossen, läuft ein CNN auf jedem Rechner. Dazu passt, dass die Erkennungsfunktion Teil von Programmiersprachen wie etwa Wolfram Alpha wird (siehe rechts). Software-Entwickler können solche Bausteine in ihre Programme einbauen, ohne dass sie etwas von der Materie verstehen müssen. Qualcomm, Marktführer bei Mobilprozessoren, will künftig sogar spezielle Hardware-Bausteine für seine Snapdragon-Chips anbieten, die auf die stark parallelisierten Rechenoperationen von CNNs optimiert sind. Sie dürften in den nächsten Jahren auf neuen Smartphone-Modellen ein so selbstverständliches Bauelement werden wie heutzutage Signalchips zur Audio- oder Videoumwandlung.

Die Entwicklung hin zu immer besseren CNNs verläuft dramatisch – vor wenigen Jahren waren sie noch nicht der favorisierte Typ unter den neuronalen Netzen, heute stellen sie ihre Speerspitze dar. Mark Asbach fürchtet, dass künftig nur große Webdienstleister bessere neuronale Netzwerke für neue Anwendungszwecke bauen, während einzelne Forscher sich das nicht mehr leisten können: Die Anzahl der für das Training nötigen Bildbeispiele wird schlicht immer größer. Den nächsten Schritt sieht er in neuronalen Netzwerken, die wirklich selbstlernend sind: „Die heutigen CNNs könnten in ein paar Jahren von einer neuen Methode abgelöst werden, die keinen menschlichen Experten erfordert, der Trainingsbeispiele aussucht. Konkrete Lösungsvorschläge zu diesem Unsupervised Learning sind heute aber noch nicht in Sicht." Damit lässt sich abschließend auch nicht die Frage beantworten, ab wann Computer so gut sehen können wie ein sechsjähriges Kind, das alles über das Sehen gelernt hat, was es zu lernen gibt

testtechnik@chip.de

FOTOS V. O.: PIXOLUS (2); NIKOLAUS SCHÄFFLER; BENJAMIN HARTLMAIER; NGUYEN, YOSINSKI, AND CLUNE

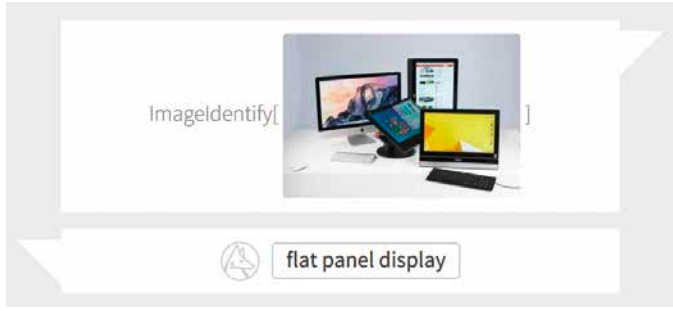
INFO

App mit neuronalem Netzwerk

Objekterkennung läuft auch auf Smartphones. Die Gratis-App von Pixolus erkennt Zählerstande anhand eines Fotos, wertet sie aus und stellt den Verbrauch über einen längeren Zeitraum dar.

Wolfram: Eine Computersprache lernt Sehen

Neuronale Netzwerke und ihre Bildanalysen sind in speziellen Programmiersprachen integriert. So erlaubt die Wolfram Language, ein Bild hinter einem Befehl zu laden und gibt dann das Ergebnis aus.



Ich sage dir, wie alt du bist

Microsoft hat ein neuronales Netz online gestellt, welches das Alter von Personen schätzt. Beim Selbstversuch in der Redaktion griff das Netz aber doch um ein paar Jahre daneben.

Die seltsamen Fehler neuronaler Netzwerke

Ein Forscherteam aus verschiedenen amerikanischen Universitäten hat gezeigt, dass auch gut trainierte neuronale Netze immer wieder dumme, für Menschen unverständliche Fehler machen.