Computer, die sehen und verstehen

Bernd Neumann
Fachbereich Informatik

Dass Computer die Bilddaten einer Kamera auswerten können, dürfte heute viele nicht mehr überraschen. In der industriellen Produktion ergreifen Roboter kameragelenkt Werkstücke, in der Medizin werden Röntgenbilder durch Computerprogramme aufbereitet, auf Autobahnen halten Fahrzeuge unter Kamerakontrolle ihre Spur, Überwachungssysteme erkennen Gesichter, etc. Für viele dieser Anwendungen dient das Sehen beim Menschen als großes Vorbild, und das Forschungsgebiet innerhalb der Informatik, das sich damit befasst, heisst deshalb auch Computersehen oder Maschinensehen (engl. Computer Vision). Nach mehr als vier Jahrzehnten Forschung werden die komplexen Prozesse, mit denen aus den Pixeln von Kamerabildern Aussagen über die Bildinhalte abgeleitet werden, in beträchtlichem Detail verstanden, und man kann sagen, dass Computer heute vielerlei eingeschränkte Sehaufgaben wahrnehmen können, für die beispielsweise das Erkennen einzelner Objekte charakteristisch ist.

Kann der Computer also im Prinzip schon sehen wie ein Mensch? Anhand eines Bildes aus dem Stummfilm "The Navigator" (Buster Keaton, 1924) kann der Unterschied verdeutlicht werden.

Sehen heißt Stummfilme verstehen können
(Szene aus Buster Keatons "The Navigator")

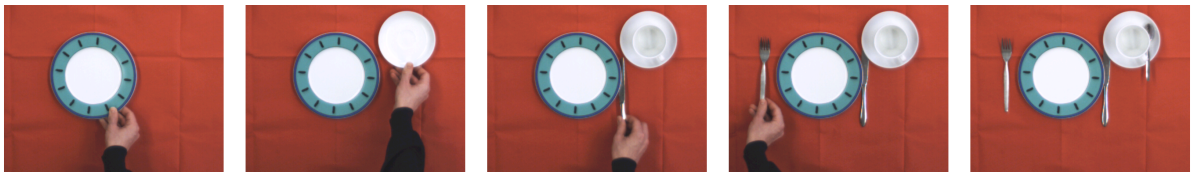
Ein Computer könnte nach dem heutigen Stand der Technik unter günstigen Umständen die wesentlichen Objekte im Bild erkennen (Person, Kessel, Kelle, Ei, Teller, ...), und dazu eine inhaltliche Beschreibung generieren, etwa "Person hält Ei auf Kelle über Kessel". Ein Mensch dagegen ist darüberhinaus in der Lage, die Szene im Kontext zu verstehen, Absichten zu erschließen, also z.B. "Person bereitet Mahlzeit in Kantine", das Missverhältnis von Kessel

und Ei zu erkennen und letztlich die Komik der Situation zu erahnen. Offenbar greifen Menschen dazu auf vielfältiges Alltagswissen zurück, das sie im Laufe ihres Lebens erfahren haben, z.B. über physikalische Zusammenhänge, typisches Verhalten von Personen, örtliche Gegebenheiten, etc.
Computer sind noch weit davon entfernt, Stummfilme zu "verstehen" wie Menschen (und ggf. an der richtigen Stelle zu lachen). Aber damit wird eine Vision beschrieben, auf die aktuelle Forschungen zum Computersehen hinarbeiten. Das Projekt CogVis ("Cognitive Vision Systems") ist das erste EU-geförderte Projekt, in dem sich Forscher mit Computersehen in erfahrungsbasierten räumlichen, zeitlichen und aufgabenbezogenen Kontexten befassen. Im Rahmen von CogVis arbeitet eine Forschungsgruppe des Autors im Arbeitsbereich Kognitive Systeme am Fachbereich Informatik der Universität Hamburg mit sechs weiteren universitären Forschungsgruppen aus Stockholm, Leeds, Genua, Zürich, Ljubljana und Tübingen zusammen. Die Projektdauer ist drei Jahre, das Projektvolumen beträgt insgesamt rund 4 Mio. Euro. Die Hamburger Aufgaben betreffen grundlegende Fragen zur Rolle übergeordneter Zusammenhänge beim Computersehen:

- Wie kann Wissen über alltägliche Vorgänge für Computeranwendungen repräsentiert werden?
- Wie kann dieses Wissen zur inhaltlichen Deutung von Szenen verwendet werden?
- Wie kann Wissen dieser Art aus Erfahrungen gewonnen werden?

Die folgenden Abschnitte geben einen Einblick in die Arbeiten der Hamburger Gruppe nach knapp zwei Jahren Projektlaufzeit.

Als Beispielszenario wird eine Tischdeckszene untersucht, in der eine Person Teller, Untertassen, Tassen, Bestecke und andere Utensilien auf einen Tisch platziert. Die untenstehende Bildsequenz zeigt fünf Bilder eines von einer festen Kamera aufgenommenen Videos von insgesamt 20 sek Dauer. Das Tischdecken ist ein experimentell leicht realisierbares aber aus der Sicht des Computersehens durchaus nichttriviales Beispiel für einen Vorgang, in dem räumliche und zeitliche Zusammenhänge, sowie Absichten eine Rolle spielen.



Fünf Bilder aus einem Tischdeck-Video illustrieren Teilvorgänge

Eine typische Aufgabe für den Computer ist zu erkennen, dass ein Teilvorgang, etwa das Platzieren eines Tellers, Bestandteil eines komplexeren Vorgangs, des Platzierens eines Gedecks, und dies wiederum Bestandteil eines Tischdeckvorgangs sein kann. Aus der Beobachtung eines Tellerplatzierens kann dann also (mit gewisser Unsicherheit) auf die Absicht der Person geschlossen werden, den Tisch zu decken. Umgesetzt in eine Forschungsaufgabe bedeutet dies, eine generische Datenstruktur zu entwickeln, mit der Vorgänge in der Gestalt von räumlich und zeitlich in Beziehung stehenden Objekten repräsentiert werden können. "Generisch" bedeutet hier, dass dieselben Repräsentationsformen auch für Vorgänge in ganz anderen Domänen benutzt werden können, z.B. für eine Müllabfuhrszene. Auch dort spielen mehrere Objekte in koordinierten Teilvorgängen eine Rolle, und das Erkennen einer konkreten Müllabfuhrszene basiert meist

auf der Wahrnehmung weniger charakteristischer Aspekte (manchmal reicht schon das Geräusch), basierend auf unserem Vorwissen über typische Müllabfuhrvorgänge.

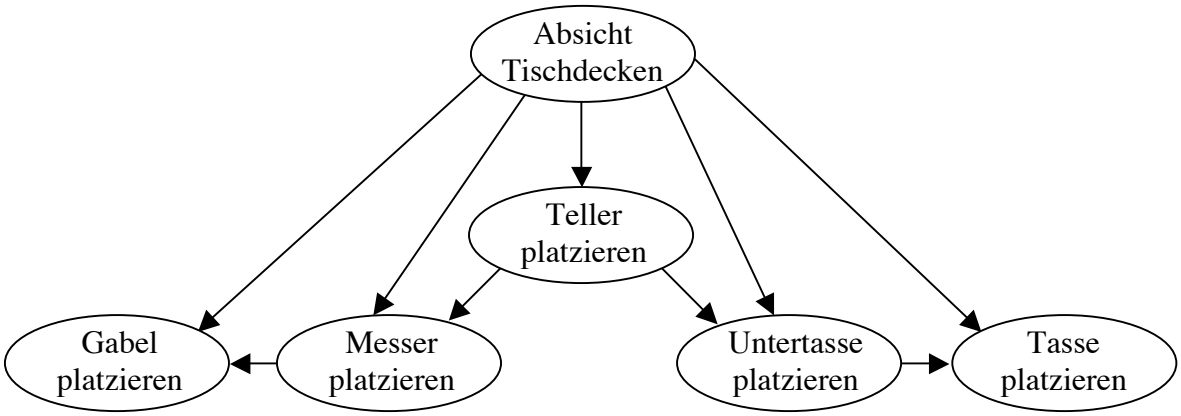
Straßenszene mit Vorgängen aus dem Alltag

Der in Hamburg entwickelte Lösungsansatz stützt sich auf zwei methodische Säulen aus der Künstlichen Intelligenz: Wissensrepräsentation mit einer Beschreibungslogik, sowie probabilistisches Schließen mit Bayesschen Netzen. Eine Beschreibungslogik erlaubt es, Objekte anhand ihrer Eigenschaften und Beziehungen zueinander logisch fundiert zu beschreiben. Das ermöglicht Schlussfolgerungsverfahren, mit denen z.B. widersprüchliches Wissen aufgedeckt oder implizites Wissen abgeleitet werden kann. Ein Tischdeckvorgang kann beispielsweise so präzise beschrieben werden, dass bei vollständiger Kenntnis der Teilvorgänge automatisch auf den übergeordneten Zusammenhang des Tischdeckens geschlossen werden kann.

name:	scene-place-cover
parents:	:is-a scene-agent-activity
parts:	pc-tt :is-a scene-table-top
	pc-tp1 :is-a scene-transport
	with (tp-obj :is-a scene-plate)
	pc-tp2:is-a scene-transport
	with (tp-obj :is-a scene-saucer)
	pc-tp3 :is-a scene-transport
	with (tp-obj :is-a scene-cup)
	pc-cv :is-a scene-cover
time marks:	pc-tb, pc-te :is-a timepoint
constraints:	pc-tp1.tp-ob = pc-cv.cv-pl
	...
	pc-tp3.tp-te ≥ pc-tp2.tp-te
	pc-tb ≤ pc-tp3.tb
	pc-te ≥ pc-cv.cv-tb

Beschreibung von Tischdecken für einen Computer. Unter "parts" werden Teilvorgänge, unter "constraints" räumliche und zeitliche Beschränkungen für die Teilvorgänge spezifiziert, z.B. durch Bezugnahme auf Anfangszeitpunkte (tb) und Endzeitpunkte (te).

Leider kann diese logische Schlussfolgerungsmöglichkeit beim Computersehen in aller Regel nicht ausgenutzt werden, denn visuell erfaßte Informationen sind normalerweise unvollständig und unsicher. Fehlende Informationen müssen durch "kontrollierte Halluzination" (ein Ausdruck des New Yorker Fachkollegen John Kender) auf der Basis von Erfahrungen und allgemeinem Wissen ergänzt werden. Dieses unsichere Geschäft erfordert die Berücksichtigung von Wahrscheinlichkeiten, insbesondere Verfahren zum Verrechnen vielfacher, statistisch korrelierter Evidenz. Erst durch die Entwicklung von Bayesschen Netzen in den letzten 10 Jahren ist es möglich geworden, Vorhersagen aus einem hochdimensionalen Verbund von Zufallsereignissen mit effektiven Verfahren abzuleiten. Die Grafik zeigt - stark vereinfacht - ein Bayessches Netz für Tischdecken. Das Netz repräsentiert mögliche Teilvorgänge als Knoten und direkte statistische Abhängigkeiten als Verbindungen (die zahlenmäßigen Wahrscheinlichkeitswerte sind nicht gezeigt). Werden nun einzelne Teilvorgänge beobachtet, etwa das Platzieren eines Tellers und einer Untertasse, so kann die Wahrscheinlichkeit weiterer Vorgänge auf der Basis dieser Evidenz berechnet werden und z.B. auf die Absicht des Akteurs, den Tisch zu decken, geschlossen werden.



Struktur eines Bayesschen Netzes gibt statistische Abhängigkeiten wieder

Die Verbindung von wahrscheinlichkeitsbasiertem Schließen mit logikbasierter Wissensrepräsentation im Zusammenhang mit beobachtbaren Vorgängen ist einer der innovativen Aspekte der Hamburger Arbeiten. Dadurch wird es möglich, die dritte zentrale Forschungsfrage des Projektes anzugehen: Wie kann man das zum Sehen erforderliche Alltagswissen erlernen? Die Relevanz dieser Fragestellung soll an einer (noch) futuristischen Anwendung verdeutlicht werden, in dem der Computer die Rolle eines Assistenten im Haushalt übernimmt. Das Assistenzsystem wird vom Hersteller mit einer gewissen Basisintelligenz aber ohne Alltagserfahrungen ausgeliefert und muss nun in der Lage sein, neue Zusammenhänge mithilfe eines Lehrers zu erlernen oder selbständig zu entdecken. Selbst wenn man auf Haushaltsroboter verzichten kann, sind maschinelle Modelle für das Lernen aus Beobachtungen von grundlegendem Interesse für zukünftige intelligente Systeme, z.B. autonome Systeme in unbemannten Weltraummissionen.

In den Hamburger Forschungen konnte bisher nur ein kleiner Ausschnitt dieser Problematik bearbeitet werden, der aber bereits wichtige Aspekte berührt. Als vereinfachte Aufgabe wurde untersucht, wie ein Computer Regeln für Gedeckformationen entdecken kann. Dazu werden Lernbeispiele (s. Bild) zur Verfügung gestellt, die ein Gedeck, teilweise auch zusätzliche Objekte, in typischer Anordnung zeigen. Der Computer gewinnt daraus eine erfahrungsbasierte konzeptuelle Repräsentation einer Gedeckformation, aus der typische Ausprägungen ableitbar sind. Dies ermöglicht - im Kleinen - den Schluss von Teilen aufs Ganze, z.B. vom Ort des Tellers auf den Ort eines - vielleicht gerade verdeckten - Messers.

Die Verallgemeinerung auf einfache beobachtbare Vorgänge, etwa das Tischdecken, scheint möglich.

Lernaufgabe für den Computer: Regelmäßigkeiten in Objektanordnungen entdecken

Angesichts des langfristigen Forschungsziels, mit Computern Bilder ähnlich wie Menschen verstehen zu wollen, sind die beschriebenen konkreten Untersuchungen natürlich kleine Schritte, die sich aber in ein langjähriges Forschungsprogramm in Hamburg einreihen. Forschungen zum Computersehen haben hier bereits vor mehr als 30 Jahren begonnen. Damals ging es darum, bewegte Objekte in Videosequenzen überhaupt erst abgrenzen zu können. (Die dazu verwendete Beispielszene eines Taxis, das von der Schlüterstraße in die Bieberstraße einbiegt, findet sich heute noch als Testszene in wissenschaftlichen Veröffentlichungen aus aller Welt.) Daran lässt sich ermessen, wieviele Fragen sich hinter der uns so selbstverständlichen Fähigkeit zu sehen auftun, und welch langer Atem trotz weltweit betriebener Forschungen erforderlich ist, um wirkliche Fortschritte zu erzielen. Wann Computer über Buster Keaton werden lachen können, ist jedenfalls noch nicht abzusehen.