Bilder in unseren Köpfen

Dr. Gabriele Peters, Lehrstuhl Graphische Systeme, Fachbereich Informatik

Täglich gehen wir in unserer Umwelt mit realen Objekten um. Dabei sind wir zum Beispiel in der Lage, unsere Kaffeetasse zu erkennen, auch wenn wir sie niemals zuvor unter exakt derselben Beleuchtung, aus exakt demselben Blickwinkel oder aus derselben Entfernung gesehen haben, wie gerade heute morgen am Frühstückstisch. Lange Zeit ging man davon aus, dass ein explizites, dreidimensionales Modell des Objektes in unseren Köpfen existieren muss, das während des Erkennungsvorgangs mental rotiert wird, bis der Vergleich zwischen dem zu erkennenden, realen Objekt und der Ansicht des mentalen Modells eine genügend große Übereinstimmung auf aufweist. Vom technischen Standpunkt aus gesehen wäre die Rotation eines dreidimensionalen Modells im Computer (und der anschließende Vergleich des künstlich erzeugten Bildes mit einem Bild des realen Objektes) eine sehr effiziente Methode zur Objekterkennung und hätte erhebliche Vorteile bei der Manipulation von Objekten. Das erfordert jedoch zunächst das Vorhandensein eines solchen Modells. Und hier liegt die Krux. Die Informationsquelle unseres Gehirns ist zweidimensional, nämlich das Bild auf unserer Netzhaut. Wie kann es möglich sein, basierend auf einer zweidimensionalen Informationsquelle, dreidimensionale Objekte zu erkennen und ihre Orientierung im Raum zu bestimmen?

Gabriele Peters, geb. 1968, seit 2001 Mitarbeiterin am Lehrstuhl Graphische Systeme des Fachbereichs Informatik der Universität Dortmund, Foschungsschwerpunkte: Verständnis unseres Gehirns, insbesondere des visuellen Systems Kommunikation zwischen Mensch und Maschine. Ausbildung zur Mathem.-Techn. Assistentin bei der Hoesch AG in Dortmund, Studium der Mathematik an der Ruhr-Universität Bochum. 1996 – 2001 Mitarbeiterin am Institut für Neuroinformatik der Ruhr-Universität Bochum, 1998/1999 Forschungsaufenthalt an der Akademie der Wissenschaften Prag, 2002 Promotion auf dem Gebiet der Objekterkennung an der Universität Bielefeld.

Die dritte Dimension

Unsere Erkennungsleistung unter nicht-konstanten Bedingungen (Abb. 1) ist deshalb so erstaunlich, weil jedes Objekt in Abhängigkeit vom Blickwinkel, von der Beleuchtung, der Entfernung oder teilweisen Verdeckungen die unterschiedlichsten Erregungsmuster auf unserer Netzhaut – und letztlich in unserem Gehirn – hinterlassen kann. Trotzdem sind wir in der Lage, zu erkennen, dass diese unterschiedlichen Muster von ein und demselben Objekt hervorgerufen werden. Diese Erkennung ist eine Leistung unseres Wahrnehmungssystems, das dazu eine *interne Repräsentation* des Objektes benötigt. Wie eine solche Repräsentation beschaffen sein kann und wie wir sie uns aneignen, sind Fragen, die die unterschiedlichsten wissenschaftlichen Fachbereiche, etwa die Philosophie, die Biologie und die Psychologie, aber auch die Physik und die Mathematik schon sehr lange beschäftigen. Die Informatik als relativ junge Wissenschaft – und hier insbesondere die Neuroinformatik – widmet sich ebenfalls diesem Grenzbereich zwischen Hirnforschung und Computerwissenschaft. In diesem Artikel soll nun einem Teilproblem der Wahrnehmung dreidimensionaler Objekte, nämlich der Frage nach der Aneignung und Beschaffenheit von *ansichtsunabhängigen* Objektrepräsentationen, nachgegangen werden. Das sind Repräsentationen, die die Erkennung eines Objektes unabhängig von dem Blickwinkel, aus dem es sich dem Betrachter gerade darbietet, erlauben.

Von Affen und Menschen

Schon seit geraumer Zeit wird die Vorstellung von mental rotierten Objektmodellen nicht mehr ernsthaft vertreten. Vielmehr gibt es eine überwältigende Fülle von Verhaltensstudien und physiologischen Experimenten, die die Annahme stützen, dass lediglich einige ausgewählte Ansichten, die in geeigneter Weise miteinander verbunden werden, ausreichen, um Wahrnehmungsleistungen zu erbringen. In typischen Verhaltensstudien, die die These einer rein *ansichtsbasierten* Repräsentation dreidimensionaler Objekte untersuchen, werden Versuchspersonen (Menschen oder Affen) in einer Trainingsphase einige Ansichten eines unbekannten Objektes präsentiert. In der darauffolgenden Testphase werden neue Ansichten desselben Objektes dargeboten, und sowohl die Zeit, die verstreicht, bis die Versuchsperson das Objekt anhand der neuen Ansicht erkennt, als auch die Fehlerrate während der Erkennung werden gemessen. Als Ergebnis solcher Studien hat sich beispielsweise ergeben, dass sowohl die Antwortzeit als auch die Fehlerrate im Allgemeinen nicht linear vom kürzesten Abstand zwischen Trainings- und Testansicht abhängen (wie man es bei einem rotierten 3D-Modell erwarten würde), sondern dass vielmehr ein Zusammenhang zwischen der Erkennungsleistung und den zweidimensionalen *Merkmalsdeformationen in der Bildebene* besteht [1]. Außerdem

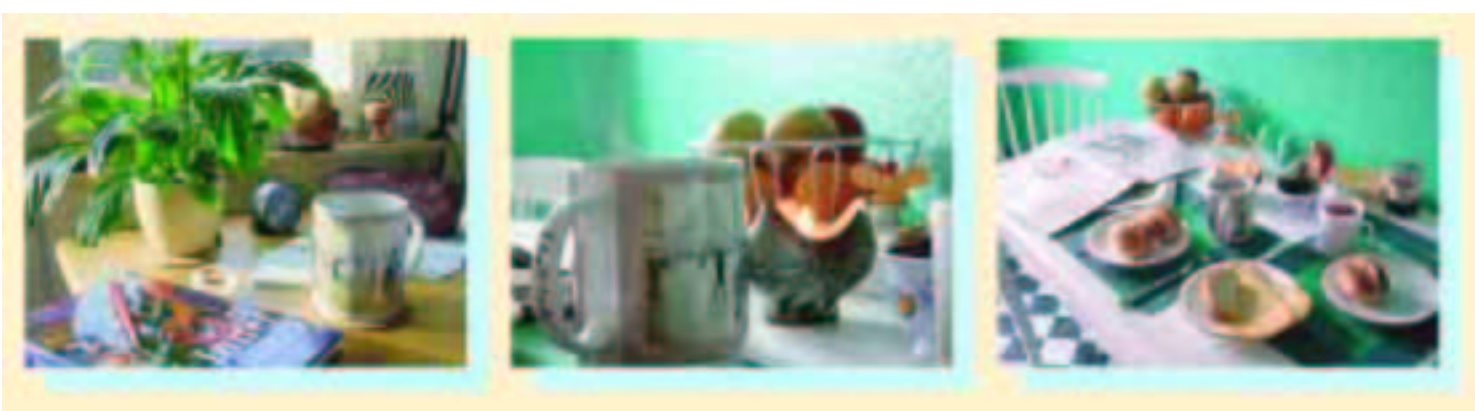


Abb. 1:
Auf jedem dieser drei Bilder erscheint dieselbe Kaffeetasse in unterschiedlichen Größen, unter verschiedenen Blickwinkeln und unter anderen Beleuchtungen. Trotzdem ist der Betrachter fähig, sie als identisches Objekt zu erkennen.

wurde herausgefunden, dass Erwachsene in der Lage sind, allein von dargebotenen, statischen Ansichten eines Objektes auf dessen dreidimensionale Form zu schließen.

Auch physiologische Experimente mit Affen stützen die These einer ansichtsbasierten Objektwahrnehmung. In typischen Experimenten werden Zellableitungen im inferioren *Temporalkortex* (IT) von Affen durchgeführt, während diese Erkennungsaufgaben durchführen müssen. In diesem Teil des Gehirns konnten spezifische Aktivitäten während der Objekt- und Gesichtserkennung festgestellt werden (Abb. 2).

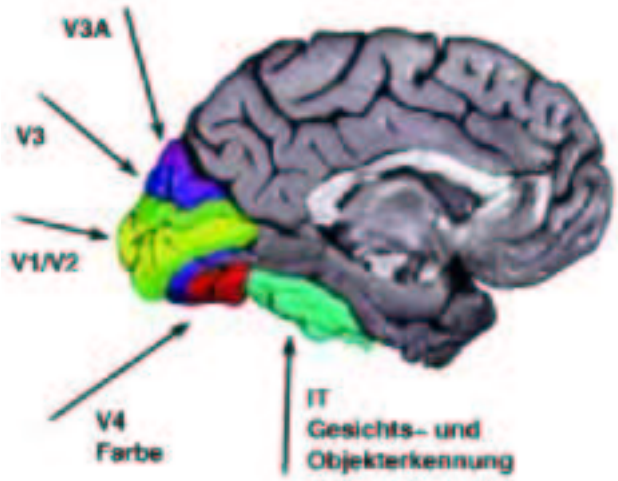


Abb. 2: Farbig markiert sind die Bereiche des Kortex, in denen die Verarbeitung visueller Information stattfindet. Der türkisfarbene Bereich markiert den inferioren Temporalkortex, der eine wichtige Funktion bei der Erkennung von Objekten und Gesichtern erfüllt. (Abbildung leicht verändert aus [7].)

Es gibt Neuronengruppen im IT, die selektiv auf lediglich einige, ausgewählte Ansichten eines Objektes reagieren, während andere Ansichten desselben Objektes geringe oder keine Reaktion auslösen [3, 4]. Wird das Objekt von einer der bevorzugten Ansichten wegrotiert, so werden die Antworten dieser Zellen schwächer. Dabei wird eine Distanz von etwa 45° zur optimalen Ansicht angegeben, bevor die Reaktion der Neurone um die Hälfte reduziert wird. Außerdem hat man herausgefunden, dass Primaten die Fähigkeit zur Verallgemeinerung von Referenzansichten zu unbekannten, bisher noch nicht gesehenen Ansichten eines Objektes besitzen. Affen können zum Beispiel unbekannte Testansichten in einem 120°-Intervall zwischen zwei Trainingsansichten erkennen, während die Erkennungsleistung außerhalb dieses Intervalls begrenzt ist [2].

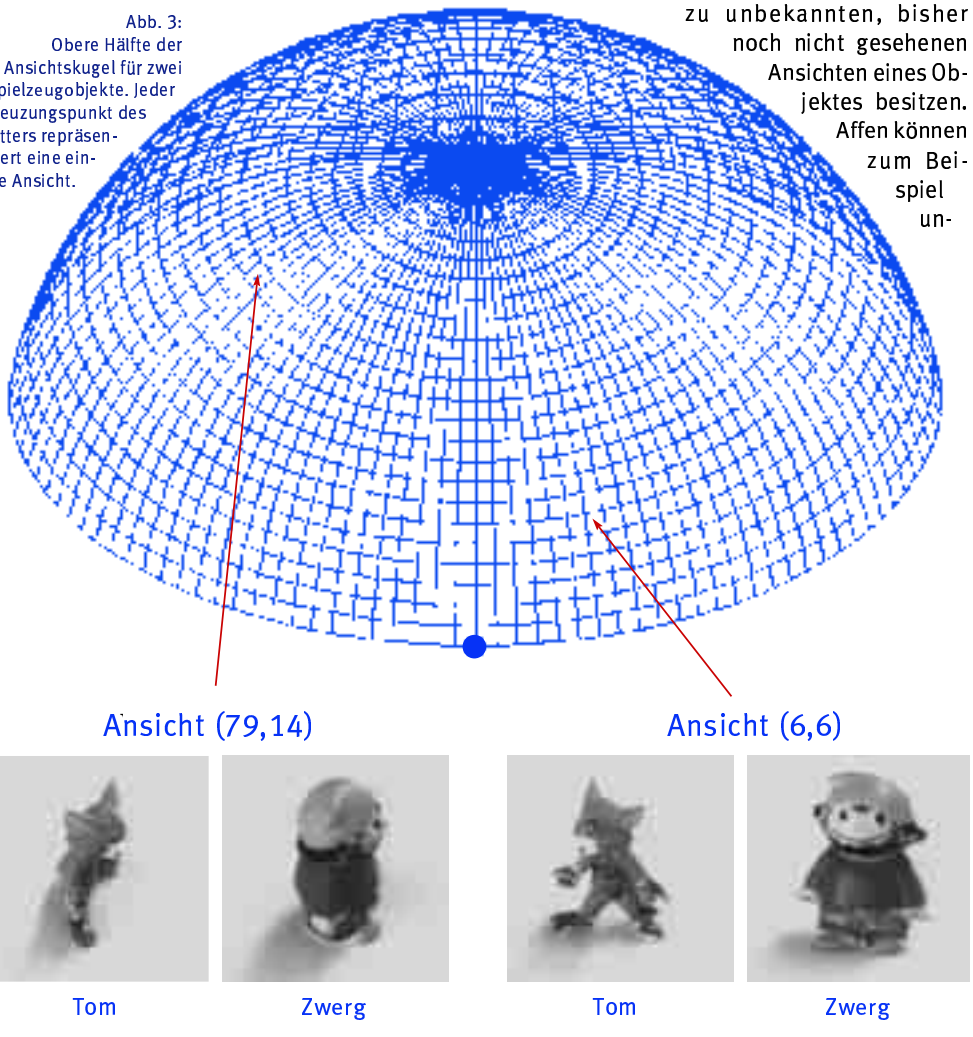


Abb. 3: Obere Hälfte der Ansichtskugel für zwei Spielzeugobjekte. Jeder Keuzungspunkt des Gitters repräsentiert eine einzelne Ansicht.

Von Objekten zu Bildern …

Um nun die These einer ansichtsbasierten Objektwahrnehmung zu überprüfen, ist die Fülle von Einzelergebnissen der Hirnforschung in ein geschlossenes Gesamtmodell integriert worden. Dabei boten Fragen, die anschließend in Computersimulationen bearbeitet wurden, Orientierung, wie etwa, ob die Wahrnehmung von dreidimensionalen Objekten ansichtsbasiert interpretiert werden kann, wie viele Ansichten für die Repräsentation eines Objektes benötigt werden, wie groß das Gebiet der Verallgemeinerbarkeit ist, innerhalb dessen man von vorgegebenen Beispielansichten auf unbekannte Ansichten schließen kann, und wie Strategien zur Kombination von bekannten zu unbekannten Ansichten aussehen können. Um Antworten auf diese Fragen mit Hilfe von Computersimulationen zu erhalten, war es zunächst notwendig, eine geeignete Datenbasis zu wählen. Diese besteht aus einer großen Anzahl von Ansichten von Spielzeugobjekten, und zwar decken 2500 in gleichen Abständen verteilte Ansichten jeweils die obere Ansichtshemisphäre eines Objektes ab (Abb. 3).

Jede dieser Ansichten wird einer automatischen Vorverarbeitung unterzogen, an dessen Ende eine Repräsentation einer einzelnen Objektansicht in Form eines Graphen steht (Abb. 4). Ein Graph ist eine Datenstruktur, die aus Knoten (in diesem Falle Bildpunkte) und Kanten, die die Knoten verbinden, besteht. Für die Repräsentation einer Ansicht sind die Knoten des Graphen mit einem Zahlenvektor etikettiert, der die lokale Bildumgebung des Knotens, d.h. ein *zweidimensionales* Objektmerkmal, beschreibt. Diese Art der Repräsentation einer Ansicht ermöglicht einen berechenbaren Vergleich zwischen verschiedenen Objektansichten.

Spärliche Objektrepräsentation

Vergleicht man nun jede Ansicht mit benachbarten Ansichten auf der Ansichtskugel, so erhält man einen sie umgebenden Bereich *ähnlicher* Ansichten, genannt *view bubble*. Die Anwendung eines geschickten Algorithmus ermöglicht es dann, möglichst wenige *view bubbles* auszuwählen, die jedoch die gesamte Ansichtskugel komplett abdecken. Dabei hängt die Größe und Anzahl der ausgewählten *view bubbles* von der vorab gesetzten Ähnlichkeitsschwelle des Bildvergleiches ab (Abb. 5). Je höher die Ähnlichkeitsschwelle für den Vergleich der Ansichten gewählt wird, desto feiner fällt die Partitionierung der Hemisphäre aus. Derartige Abdeckungen der Hemisphäre ermöglichen nun die Repräsentation eines dreidimensionalen Objektes durch einige, wenige Ansichten: nämlich durch die Zentralansichten und jeweils vier Granzansichten der *view bubbles* einer Abdeckung (Abb. 6). Die Repräsentation eines Objektes in Form einiger, weniger Ansichten stellt eine enorme Datenreduktion dar. Sie kann jedoch nur dann praktikabel sein, wenn sie noch genügend Information über das Objekt beinhaltet, um mit ihr Wahrnehmungsaufgaben lösen zu können.

Morphen unbekannter Ansichten

Um den Informationsgehalt einer solch spärlichen Repräsentation zu testen, ist für eine große Anzahl von Ansichten, die nicht in der Objektrepräsentation vorhanden sind, eine künstliche Version aus den Ansichten der Repräsentation erzeugt worden. Dies geschieht mit Hilfe eines sogenannten *Morphing*-Algorithmus. Diese Bezeichnung wurde von dem Begriff „Metamorphose“ hergeleitet, der die allmähliche Umwandlung eines Objektes in ein anderes Objekt beschreibt. Der Begriff „Morphing“ wird für eine Animationstechnik verwendet, die zum Beispiel auch bei der Produktion von Science-Fiction-Filmen zum Einsatz kommt. Dabei werden Quellansichten eines Objektes in eine gewünschte Zielansicht desselben oder eines anderen Objektes deformiert. Für unseren Anwendungsfall werden zu diesem Zweck aus den Grauwerten der Bildpunkte (Pixel) der gegebenen Quellansichten und ihren gegebenen Positionen auf der Ansichtskugel die Grauwerte der Pixel der gesuchten, unbekannten Ansicht berechnet. Dieses so erzeugte, gemorphte Bild kann dann mit der Originalansicht verglichen werden, die zwar nicht zur Objektrepräsentation gehört, aber noch in der Ausgangsdatenbasis vorhanden ist (Abb. 7).

Wie zu erwarten war, liefern feinere Partitionierungen der Ansichtskugel genauere Rekonstruktionen unbekannter Ansichten. Für eine vernünftig erscheinende, mittlere Partitonierung eines Objektes, bei der die gespeicherten Ansichten einen Abstand von über 30° haben, ist der Rekonstruktionsfehler jedoch mit weniger als 5 % Abweichung von der Originalansicht schon erstaunlich gering. Dies zeigt, dass eine Objektrepräsentation, die aus lediglich einigen ausgewählten Ansichten des Objektes besteht, genügend Information enthält, um nicht gespeicherte Ansichten rekonstruieren zu können.

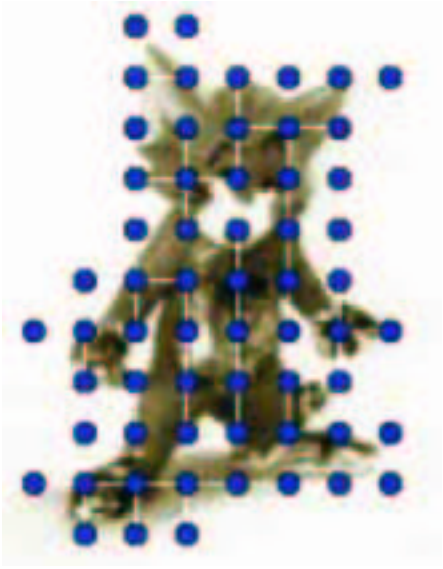


Abb. 4:
Jede Ansicht eines Objektes wird durch einen Graphen repräsentiert, dessen Knoten mit Beschreibungen zweidimensionaler, lokaler Objektmerkmale etikettiert sind.

Abb. 5:
Eine view bubble wird durch ein auf die Ansichtshalbkugel projiziertes Rechteck dargestellt. Die Zentralansichten der view bubbles sind durch Punkte gekennzeichnet. Die Zahlen geben die Anzahl der view bubbles der jeweiligen Partitionierung an.

Abdeckung der Ansichtshalbkugel mit view bubbles

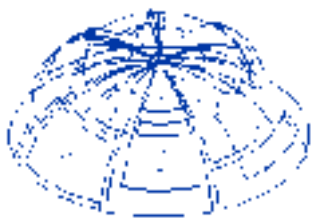
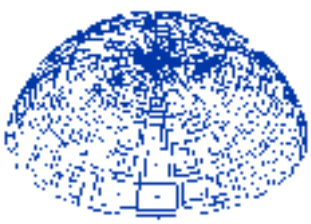
Objekt \ Ähnlichkeitsschwelle	0.75	0.85	0.95
 Tom	 6	 29	 289
 Zwerg	 4	 26	 225

Abb. 6:
Die hier dargestellten Ansichten bilden eine vollständige Repräsentation des Objektes „Tom“. Sie wurden mit der niedrigsten Ähnlichkeitsschwelle 0.75 ermittelt (vergleiche mit Abbildung 5).

... zurück zu Objekten

Die Rekonstruierbarkeit unbekannter Ansichten mit Hilfe einer spärlichen Objektrepräsentation sagt allerdings noch nichts über die Fähigkeit aus, Wahrnehmungsaufgaben zu erfüllen. Zu diesem Zweck mussten sich die automatisch gelernten Objektrepräsentationen in einer Wahrnehmungsaufgabe bewähren, in der eine unbekannte Ansicht des Objektes vorgegeben wurde und das künstliche System die Pose, d.h. die genaue Position der Ansicht auf der Ansichtskugel, angeben sollte.

Dazu wird die Testansicht mit den Zentralansichten aller *view bubbles* einer Repräsentation verglichen. Diejenige Zentralansicht, welche die größte Ähnlichkeit zur Testansicht aufweist, stellt eine grobe Schätzung der gesuchten Pose dar. Diese grobe Schätzung kann daraufhin noch verfeinert werden, indem mit Hilfe des eben beschriebenen Morphing-Algorithmus Ansichten innerhalb der geschätzten *view bubble* erzeugt werden. Ein weiterer Vergleich mit diesen künstlichen Ansichten liefert dann die geschätzte Pose der Testansicht. In Abbildung 8 sind einige Ergebnisse der Posenschätzung dargestellt. Die Hemisphären sind hier in der Aufsicht abgebildet. Die grünen Quadrate bezeichnen die Ansichten, die in der Objektrepräsentation gespeichert sind, die blauen Punkte stellen die Ansichten dar, deren Pose geschätzt werden soll. Die roten Kreise sind die geschätzten Positionen. Man erkennt, dass die Schätzung bereits bei einer mittleren Anzahl von Ansichten in der Repräsentation (Ähnlichkeitsschwelle 0.85) sehr gute Ergebnisse liefert. Die mittlere Abweichung der geschätzten Pose zur Testansicht liegt hier beim Objekt „Tom“ bei 0.8° für 30 geschätzte Posen. Mit bloßem Auge kann man so gerade eben eine Abweichung von 4° zwischen zwei Ansichten erkennen.

Weißes Rauschen

Da dieses Resultat so überzeugend ist, wurde die Qualität der Testbilder stark verschlechtert, indem zusätzliche Störungen zu den Grauwerten der einzelnen Pixel addiert wurden. Auf diese derart *verrauschten* Bilder wurde der Schätzalgorithmus abermals angewendet. Selbst damit kann das System umgehen. Zehn verrauschte Ansichten, deren Posen mit einer Re-

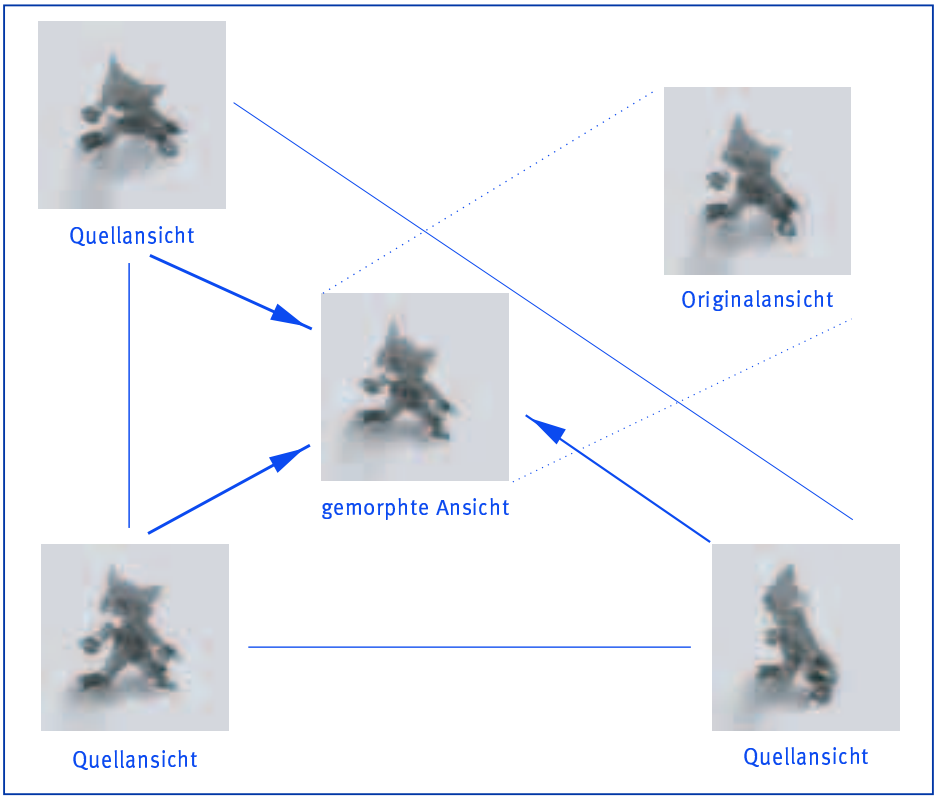


Abb. 7:
Aus den drei Quellansichten, die in der Objektrepräsentation vorhanden sind, kann die unbekannte Ansicht „gemorpht“ werden. Diese künstlich erzeugte Version ist zu vergleichen mit der Originalansicht.

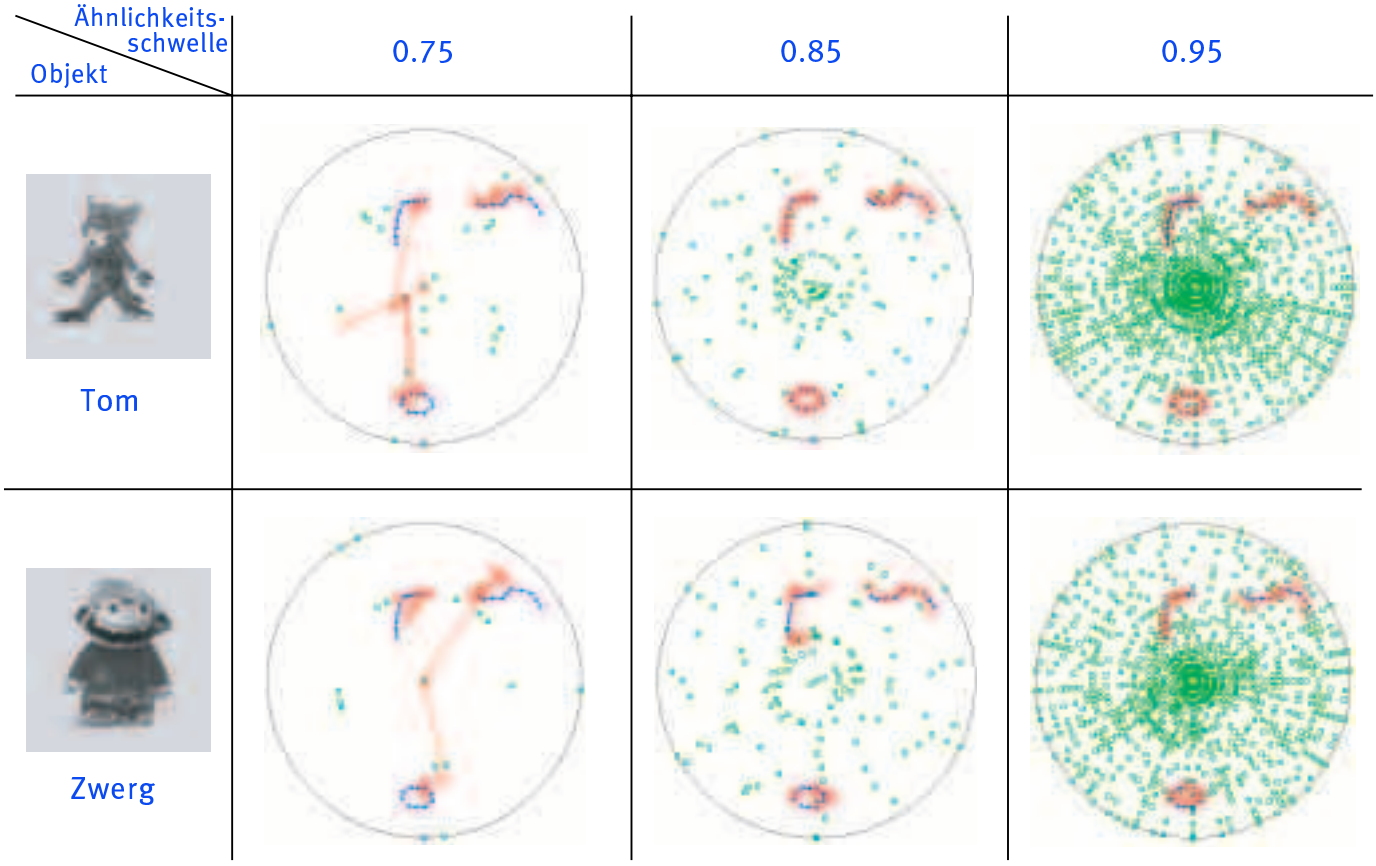


Abb. 8: Für drei verschiedene Partitionierungen der Ansichtshemisphäre (parametrisiert durch die Ähnlichkeitsschwelle des Bildvergleichs beim Erzeugen der Objektrepräsentation) sind hier für zwei Objekte Ergebnisse (rot) der Posenschätzung von jeweils 30 Testansichten (blau) dargestellt.

Abb. 9: Die erste Zeile zeigt eine Testsequenz von Objektansichten die mit weißem Rauschen versehen wurde. In der zweiten Zeile sind die geschätzten Posen abgebildet. Bis auf einen gekennzeichneten Ausreißer sind alle Ansichten recht gut ihren korrekten Positionen auf der Ansichtskugel zugewiesen worden.

präsentation geschätzt wurden, die 130 Ansichten des Zwerges enthält, sind in Abbildung 9 dargestellt. Der mittlere Schätzfehler liegt in diesem Beispiel bei etwa 10°.

Fazit

Zusammenfassend kann man sagen, dass die These von internen Objektrepräsentationen, die ohne explizite, dreidimensionale Information auskommen, sondern lediglich aus einigen zweidimensionalen Bildern bestehen, die zur Ausführung von Wahrnehmungsfunktionen in geschickter Weise miteinander verknüpft werden, durch diese Simulationen bestätigt werden kann. Detailliertere Ergebnisse können in [5, 6] nachgelesen werden. Auch in Zukunft soll diese interessante Schnittstelle zwischen Informatik und Biologie weiterverfolgt werden, denn sehr viele Fragen müssen noch beantwortet werden, bis auch ein künstliches System in der Lage ist, unsere Kaffeetasse unter den verschiedensten Bedingungen zu erkennen.

Kontakt: peters@ls7.cs.uni-dortmund.de, Ruf: (0231) 755-6122

Literatur

[1] F. Cutzu and S. Edelman. Canonical Views in Object Representation and Recognition. Vision Research, 34: 3037–3056, 1994.

[2] N. K. Logothetis, J. Pauls, H. H. Bülthoff, and T. Poggio. Evidence for Recognition based on Interpolation among 2D-Views of Objects in Monkeys. Invest. Ophthalmol. Vis. Sci. Suppl., 34: 1132, 1992.

[3] N. K. Logothetis, J. Pauls, and T. Poggio. Shape Representation in the Inferior Temporal Cortex of Monkeys. Current Biology, 5(5): 552–563, 1995.

[4] D. I. Perrett, P. A. J. Smith, D. D. Potter, A. J. Mistlin, A. S. Head, A. D. Milner, and M. A. Jeeves. Visual Cells in the Temporal Cortex Sensitive to Face View and Gaze Direction. In: Proceedings of the Royal Society of London, B(3), pages 293–317, 1985.

[5] G. Peters and C. von der Malsburg. View Reconstruction by Linear Combination of Sample Views. In T. Cootes and C. Taylor (editors) Proceedings of the 12th British Machine Vision Conference (BMVC:2001), volume 1, pages 223–232, Manchester, UK, September 10–13, 2001.

[6] G. Peters, B. Zitova, and C. von der Malsburg. How to Measure the Pose Robustness of Object Views. Image and Vision Computing, 20(4): 249–256, 2002.

[7] S. Zeki. Inner Vision – An Exploration of Art and the Brain. Oxford University Press, 1999.