Zeitschrift: Helvetica Physica Acta
Band: 61 (1988)
Heft: 1-2

Artikel: Computer Vision und künstliche Intelligenz
Autor: Ade, Frank
DOI: https://doi.org/10.5169/seals-115931

Helvetica Physica Acta
Vol. 61 (1988)

0018-0238/88/020036-06$1.50+0.20/0

COMPUTER VISION UND KÜNSTLICHE INTELLIGENZ

Frank Ade
Institut für Kommunikationstechnik
Fachgruppe Bildwissenschaft
ETH Zürich

1 Definitionen

Das Gebiet 'Künstliche Intelligenz' umfasst die Bemühungen, Computer zu befähigen, Probleme zu lösen, bei denen ihm der Mensch zur Zeit noch überlegen ist. Die Problemlösung erfolgt dabei vornehmlich mit symbolischen Methoden.

Das Gebiet 'Computer Vision' umfasst den Komplex von Wissen und Methoden, der notwendig ist, um - ausgehend von einem oder mehreren zweidimensionalen Bildern - mit dem Computer eine explizite, bedeutungsvolle, symbolische Beschreibung der dem Bildmaterial zugrundeliegenden Realität zu erzeugen. Wenn diese Realität eine Szene, d.h. ein Ausschnitt aus der dreidimensionalen Welt ist, spricht man gelegentlich auch genauer von dreidimensionaler Computer Vision oder von Szenenanalyse.

2 Computer Vision als Angewandte Künstliche Intelligenz

Historisch gesehen sind Computer Vision und Künstliche Intelligenz von den Anfängen an miteinander verbunden. Viele der allgemeinen Techniken und Methoden der Künstlichen Intelligenz wurden bei Fragestellungen aus dem Gebiet der Computer Vision entwickelt. Deshalb wird sie häufig als ein Schlüsselgebiet der Künstlichen Intelligenz bezeichnet.

Das Verstehen der visuellen Umwelt ist für den Menschen so mühelos , dass er diese Leistung meist unbewusst vollbringt. Es hat sich aber als äusserst schwierig herausgestellt, die gleiche Aufgabe mit Computern zu lösen. Damit ist gemäss der obigen Definition die Zugehörigkeit der Computer Vision zum Gebiet der Künstlichen Intelligenz gegeben. Computer Vision *ist schwierig,* vor allem aus folgenden Gründen:

- Bilder sind Projektionen eines Ausschnitts aus der dreidimensionalen Wirklichkeit auf eine zweidimensionale Mannigfaltigkeit. Dabei geht unwiederbringlich Information über die Szene verloren, vor allem Tiefeninformation, und damit verknüpft, Information über die Nachbarschaftsverhältnisse zwischen den Objekten der Szene. Um eine rechnerinterne dreidimensionale Rekonstruktion der Wirklichkeit zu durchführen zu können, müssen neben den Bilddaten in massiver Weise 'vernünftige' (heuristische) Annahmen über die Wirklichkeit mitverwendet werden.
- Eine weitere, andersgeartete Quelle des Informationsverlustes ist das von den bildgebenden Sensoren eingeführte Bildrauschen, dem ebenfalls durch geeignete Massnahmen begegnet werden muss. Bevor irgendeine informationsextrahierende Operation auf dem Bild durchgeführt werden kann, muss eine Regularisierung, i.a. eine Glättung durch Faltung z.B. mit einer zweidimensionalen Gaussfunktion erfolgen.
- Mehrere physikalische Prozesse und Eigenschaften wirken zusammen, um einem Bildpunkt eine bestimmte Helligkeit, bzw. Farbe zu geben: direkte und indirekte Lichteinstrahlung, die Reflexionseigenschaften des Körpers, Schattenwurf, atmosphärische Absorption etc.. Ein eindeutiger

Rückschluss auf die Einzelursachen ist ohne Vorwissen und/oder plausible Annahmen nicht möglich.

- Die Computer Vision muss sehr grosse Mengen von Eingabedaten sehr schnell verarbeiten. Gängige Bildgrössen bei der Computer Vision sind 512 x 512 'Pixel' (Bildelemente) mit 8 Bit Intensitätsauflösung (256 Abstufungen). Beim menschlichen visuellen System werden solche Datenmengen 'real time' verarbeitet. Auch bei der Computer Vision gibt es Aufgabenstellungen mit 'real time'-Anspruch, insbesondere in der Robotik und bei Überwachungsaufgaben. Die Verarbeitung von Bildsequenzen ist gegenwärtig der am schnellsten wachsende Zweig der Computer Vision. Hier sind offensichtlich die Anforderungen an Speicherplatz und Verarbeitungsgeschwindigkeit noch höher.
- Um die Interpretation eines Bildes korrekt durchzuführen, wird viel Vorwissen benötigt. Dazu gehört Wissen über Objekte und Prozesse im betrachteten Weltausschnitt, Wissen über die Prozesse der Bildverarbeitung und Bildinterpretation sowie über Ablaufkontrolle. Dieses Wissen ist stark strukturiert, d.h. nicht nur eine Ansammlung von Einzelfakten. Langzeitwissen über ganze Klassen von Bildern muss getrennt gehalten werden vom Kurzzeitwissen, das sich auf das aktuelle Bild bezieht.

3 Ein prototypisches Computer-Vision-System (CVS)

Motivationen für die Beschäftigung mit Computer-Vision-Systemen: Eine wichtige treibende Kraft bei der Entwicklung von Computer-Vision-Systemen waren sicher Bedürfnisse, wie sie bei praktischen Anwendungen zutage traten, etwa der Wunsch, den Menschen von der ermüdenden Interpretation grosser Bildmengen zu befreien, wie sie in der Medizin, der Biologie und der Fernerkundung anfallen oder das Bestreben, Roboter mit visuellem Feedback auszustatten, um die bei blinden Robotern erforderlichen hohen absoluten Positioniergenauigkeiten vermeiden zu können. Auch das automome Landvehikel (ALV), an dem vielerorts gearbeitet wird, benötigt visuellen Feedback.

Während diese Arbeitsrichtungen, die von einer sehr schnell wachsenden Zahl von Gruppen gepflegt werden, grundsätzlich zu spezifischen Problemlösungen führen, besteht unabhängig davon auch ein grosses, rein wissenschaftliches Interesse daran, universelle Prinzipien der Verarbeitung visueller Information aufzudecken, die gleicherweise für das menschliche visuelle System und für Computer-Vision-Systeme gelten würden. Dafür sind auch Ergebnisse aus der Neuroanatomie, der Neurophysiologie und der Kognitiven Psychologie von grosser Bedeutung.

Stand der Entwicklung von Computer-Vision-Systemen: Die Entwicklung eines Computer-Vision-Systems, das nicht auf eine bestimmte, stark eingeschränkte Aufgabe zugeschnitten ist, sondern einem gewissen Anspruch an Allgemeinheit genügt, ist ein schwieriges und zeitaufwendiges Unterfangen. Deswegen haben sich weltweit nur vielleicht zwei Dutzend Arbeitsgruppen an diese Aufgabe herangewagt. Obwohl an einigen dieser Systeme bereits seit mehr als zehn Jahren gearbeitet wird, kann von keinem behauptet werden, dass es ausgereift sei. Dennoch sind von diesen Arbeiten immer wieder richtunggebende Impulse ausgegangen.

Die verschiedenen demonstrierten Systeme weisen zwar Unterschiede auf, aber es scheint sich doch ein gewisser Konsensus darüber herauszubilden, welche Komponenten ein solches System aufweisen muss, und in welchen Etappen die Verarbeitung der visuellen Eingabedaten ablaufen sollte.

Grundsätzlicher Ablauf einer Szenen-Interpretation: Das umseitige, notwendigerweise stark vereinfachte Schema soll dazu dienen, klar zu machen, was man unter einem Computer-Vision-System versteht und einige seiner Charakteristika zu erläutern.

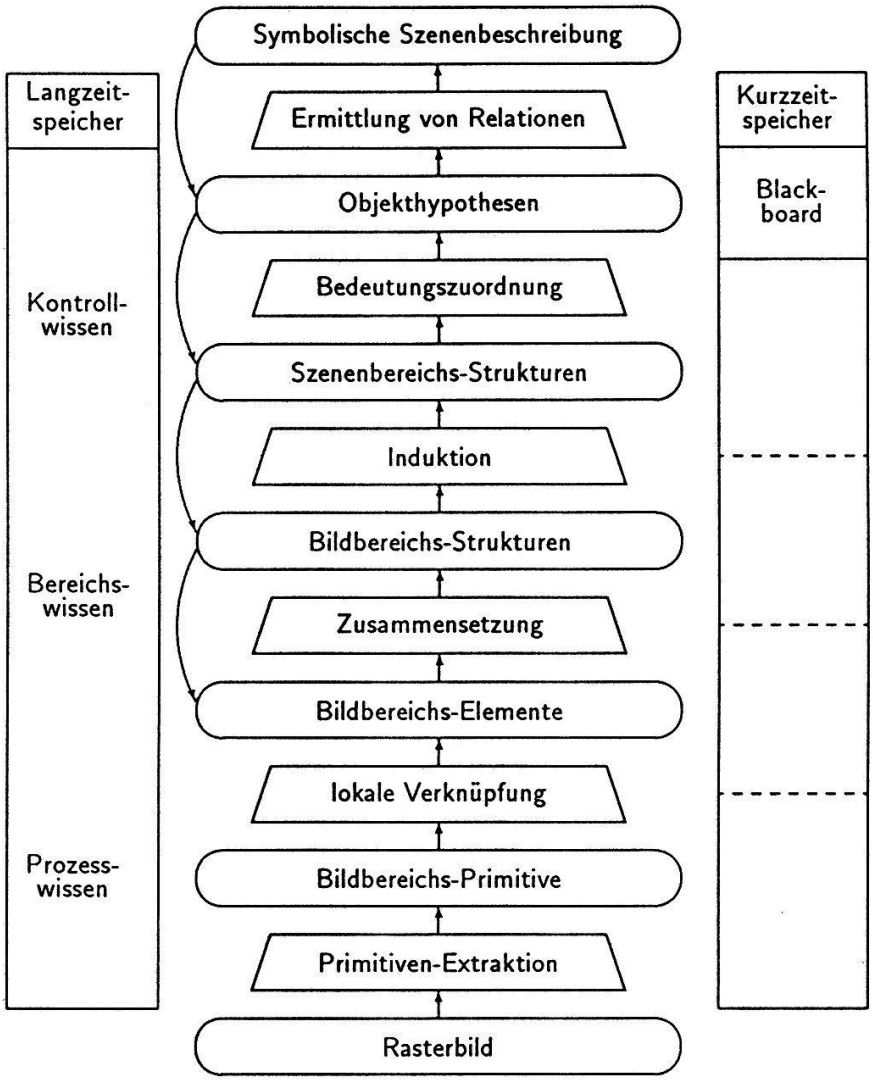

Fig.1 Schema eines Computer-Vision-Systems

Die verwendeten Bilder werden in digitaler Form eingegeben, d.h. sie sind sowohl in den beiden räumlichen Dimensionen, wie auch in der Intensität quantisiert. Aus den Bildern werden in einem ersten Schritt Primitive extrahiert, die geeignet erscheinen, als Bausteine beim Aufbau komplexerer Substrukturen und Strukturen zu dienen. Gewöhnlich sind dies Kantenpunkte bzw. kleine homogene Flächen. Kantenpunkte werden zu Kantenstücken, diese zu Konturen zusammengebaut. Auf einer bestimmten Stufe erfolgt der Übergang zu dreidimensionalen Strukturen, die aufgrund der bis dahin erarbeiteten zweidimensionalen Substrukturen erschlossen werden. Dabei wird allgemeinstes Wissen über das Aussehen physikalischer Körper im dreidimensionalen Raum verwendet. Erst dann beginnt die Deutung dieser Körper als Objekte im Rahmen bestimmter Kontexte, die bestimmte Erwartungen erzeugen. Die symbolische Beschreibung wird durch die Beschreibung der Beziehungen zwischen den erkannten Objekten abgeschlossen.

Modularität: Die Aufarbeitung eines Bildes geschieht in einer Kette von Schritten. Jeder Schritt wird von einem spezialisierten Modul durchgeführt. Daneben müssen Kontrollmoduln vorhanden sein, die diese Aktivitäten koordinieren und spezielle Speicher für Zwischenresultate, die von den Arbeits- und Kontrollmoduln begutachtet werden können. Je nach Resultat können weitere Aktivitäten gestartet werden. Somit weisen CVS eine modulare Struktur auf.

bottom-up/top-down: Die Arbeitsrichtung zu Beginn der Prozesskette ist vorwiegend bottom-up (auch als data-driven, forward chaining, Vorwärtsableitung bezeichnet). Dies ist besonders dann unproblematisch, wenn die Eingabedaten störungsarm sind. Falls das Eingabebild Störungen oder Mehrdeutigkeiten aufweist, müssen Zielvorgaben von höheren Stufen erfolgen, die die auf unteren Stufen laufenden Prozesse mithilfe intermediärer Resultate zu erfüllen suchen. Dann spricht man von einem top-down Vorgehen (auch als goal-driven, backward chaining oder Rückwärtsableitung bezeichnet). Da ja der visuelle Input unvermeidbar vieldeutig ist, spielt diese zielgetriebene Arbeitsweise eine grosse Rolle. Ein Computer-Vision-System muss bidirektional arbeiten können, d.h. die Arbeitsrichtung wird 'opportunistisch' top-down oder bottom-up gewählt.

Methoden aus der Künstlichen Intelligenz in einem Computer-Vision-System: Methoden aus dem Arsenal der Künstlichen Intelligenz lassen sich überall in einem CVS aufweisen:

- Beim Verknüpfen von Kantenpixeln zu Konturen werden Suchverfahren eingesetzt. Die 'Güte' der Kanten wird heuristisch (z.B. aufgrund von Gradientenstärke und Krümmung) beurteilt.
- Wissen über die Welt kann durch Semantische Netze dargestellt werden. Darin sind Objekte durch Instanzen (Realisierungen) von Konzepten (Frames) repräsentiert, deren Attribute explizit gespeichert sind, oder von Konzepten und/oder Superkonzepten (Generalisierungen), mit denen sie in einer Taxonomie verknüpft sind, durch Vererbung erhalten werden.
- Wissen in Form von Regelmengen ('if A then B') wird bei Gruppierungsprozessen eingesetzt.
- Von grosser Bedeutung beim Erkennen von Objekten und Objektrelationen ist das Matching (Mustervergleich) zwischen Ganzheiten, die vom aktuellen Bild, und solchen, die aus der Modellbasis (Langzeitwissen) abgeleitet wurden. Es kann sich um Matching von Symbolen, von Symbolketten oder von stark strukturierten Graphen handeln.
- Blackboard-Strukturen dienen der Buchhaltung über aktuelle Objekthypothesen und regeln den Zugang für die verschiedenen Moduln bezüglich Lesen, Modifikation und Eintrag.

Endresultat: Symbolische Beschreibung der Szene: Einfachste symbolische Beschreibungen sind schon als Resultat der Primitiven-Extraktion vorhanden. Jeder extrahierte Kantenpunkt besitzt dann eine Anzahl von Attributen. Nach Verknüpfen einer Reihe von Kantenpunkten zu Konturstücken sind neue, als Einheit ansprechbare Entitäten entstanden. Auf der einen Seite führt dies zu einer Datenreduktion, da im ganzen Bild relativ wenige dieser 'höheren' Entitäten vorhanden sind, auf der anderen Seite wird die Struktur dieser Entitäten immer komplexer. Dieser Prozess setzt sich fort bis zur endgültigen symbolischen Szenenbeschreibung.

Wenn wir den Informationsgehalt der symbolischen Szenenbeschreibung mit dem des Eingabebilds vergleichen, sehen wir, dass er geringer geworden ist. Aber die Information im Eingabebild war nur implizit vorhanden, nicht direkt zugänglich und verwertbar. Die symbolische Szenenbeschreibung dagegen ist explizit. Sie enthält nichts Überflüssiges mehr, sondern genau das, was ein 'intelligenter Agent' (Mensch oder Maschine) braucht, um über die Dinge nachzudenken, über sie zu kommunizieren oder in einer bestimmten Situation adäquat zu handeln.

4 Parallelen zwischen dem menschlichen visuellen System und einem Computer-Vision-System

Wohl in wenigen Forschungsbereichen hat man sich so bereitwillig von Lösungen der Natur inspirieren lassen wie in der Computer Vision. Hier einige Punkte, wo sich Vergleiche aufdrängen:

Eingabe des Bildes durch einen Array diskreter Sensoren: Es gibt einen quantitativen Unterschied: Das menschliche Auge weist 100 Millionen Sensoren auf, eine CCD-Kamera typischerweise 1/4 Million (Tendenz steigend). Es gibt Unterschiede in der Farbtüchtigkeit, in der

Empfindlichkeit und in der Anpassungsfähigkeit an verschiedene Grundhelligkeiten. Unmittelbar im Augenhintergrund kommt es bereits zu einer starken hardwaremässig realisierten Datenreduktion. Der Sehnerv enthält nur noch 1 Million Fasern.

Die Verwendung verschiedener Auflösungen: Die Dichte der Sensoren nimmt im menschlichen Auge von der Fovea ausgehend nach aussen hin ab. Eine solche Anordnung in der Sensoreinheit eines CVS verwenden zu wollen, würde beim heutigen Stand der Technik zu grossen Schwierigkeiten führen. Stattdessen wird das ganze Bild mit hoher und über die Bildfläche konstanter Auflösung aufgenommen. Aus diesem Urbild werden durch Zusammenfassung und Mittelung von Helligkeitswerten gröbere (schlechter aufgelöste) zusätzliche Bilder erzeugt. Durch Iteration dieses Vorgangs entsteht eine 'Auflösungs-Pyramide'. Deren Nutzung, vor allem auf den unteren Stufen des Prozesses der Bildanalyse, ist aktueller Forschungsgegenstand.

Extraktion von Primitiven: Beim menschlichen visuellen System erfolgt die Extraktion von Primitiven, z.B. Kantenpunkten parallel über das ganze Gesichtsfeld, beim CVS meist noch sequentiell. Gewisse parallele Rechner für Bildverarbeitung mit SIMD-Architektur (Single Instruction Stream, Multiple Data Stream) können effizient in paralleler Weise die dabei erforderliche Faltungsoperation rechnen und könnten somit Modul in einem CVS werden.

Neural nets: Das menschliche Gehirn, von dessen Gesamtmasse etwa 1/4 dem visuellen System zugerechnet wird, enthält 10^{12} Neuronen. Jedes Neuron ist Ziel, bzw. Ursprung von je ca. tausend Signalleitungen. Diese stellen Verbindungen nicht nur innerhalb eines Moduls her, sondern auch zwischen ihnen. Solche Befunde führen immer wieder zu neuen Entwürfen künstlicher neuraler Netzwerke mit der Hoffnung, sie in Computer-Vision-Systemen einsetzen zu können.

5 'Computing for Vision'

Obwohl die Leistungen des menschlichen visuellen Systems diejenigen heutiger CVS noch weit übertreffen, hat man bisher keinen stichhaltigen Grund gefunden, die Arbeitshypothese *Sehen ist Rechnen* aufzugeben. Der noch bestehende Unterschied wird als nicht grundsätzlich angesehen, sondern als Auswirkung der stark verschiedenen Hardware.

Die verschiedenen Stufen eines CVS stellen verschiedene Anforderungen an die Rechner-Unterstützung:

Extraktion der Bildprimitiven (Low level): Zur Glättung verrauschter Bilder und zur Extraktion von Kanten werden in grossem Umfang Faltungsoperationen eingesetzt, d.h. es wird massives 'number crunching' benötigt. Die Operationsfolge ist für alle Punkte eines Bildes gleich. Deshalb ist es vorteilhaft, sie 'geographisch' parallel durchzuführen. Im menschlichen visuellen System ist diese Art von Parallelismus von der Retina bis zum visuellen Kortex implementiert.

Eine zweite Art von Parallelismus, ein 'funktionaler' Parallelismus, liegt dann vor, wenn dieselben Eingangsdaten parallel verschiedenen Prozessoren zur Verarbeitung übergeben werden, die aus ihnen verschiedene Resultate extrahieren. Als Beispiel sei hier die parallele Untersuchung des visuellen Inputs auf Kanten verschiedener Richtung genannt, wie sie vom menschlichen visuellen System bekannt ist. Auch die Realisierung dieses Parallelismus wäre sehr erwünscht.

Schliesslich gibt es auch noch den 'pipeline'-Parallelismus, oder 'temporalen' Parallelismus, der dadurch zu charakterisieren ist, dass Daten verschiedene Verarbeitungsstationen durchströmen, die alle gleichzeitig aktiv sind. Diese Art von Parallelismus wäre vor allem bei Echtzeit-Bildverarbeitung erwünscht, wo ständig Bilddaten an den Eingängen des Systems anliegen.

Gruppierung (Intermediate level) und Interpretation (High Level): In diesen System-Teilen sind die Entitäten bereits symbolisch dargestellt, wobei anfangs ihre Zahl noch erheblich ist, während weiter oben vor allem ihre interne Struktur komplexer wird und die Verbin-

dungen zu anderen Entitäten dargestellt werden müssen (Semantische Netze). Hier wird 'symbol crunching' verlangt, d.h. die effiziente Manipulation stark strukturierter symbolischer Objekte (Matchen von Strukturen, Einfügen, Abändern und Löschen von Komponenten). Ebenso sollte das Bewerten und Verwalten verschiedener Hypothesen möglich sein.

Diesen Anforderungen kann gegenübergestellt werden, wie verschiedene Rechnerarchitekturen sie befriedigen können.

v.Neumann-Architektur: Computer mit v.Neumann-Architektur sind wegen ihrer sequentiellen Arbeitsweise weder für Low-Level-Aufgaben noch für Symbolverarbeitung ideal.

Zweidimensionale SIMD-Arrays: Unter diesem Namen seien hier alle Architekturen zusammengefasst, mit denen sich bei der Bildverarbeitung ein massiver geographischer Parallelismus realisieren lässt (ein Prozessor pro Pixel). Die Zahl der Prozessoren, mit denen ein gegebener Prozessor kommunizieren kann, ist i.a. relativ niedrig. Diese Architektur ist sehr gut geeignet für Low-Level-Prozesse, z.B. Faltungen, nicht jedoch für Symbolverarbeitung.

LISP-Maschinen: LISP, die klassische Sprache der Künstlichen Intelligenz, ist geeignet, höhere Sprachen zur Manipulation komplexer Datenstrukturen wie Semantischer Netze etc. zu implementieren. Es hat sich auch gezeigt, dass LISP durch zugeschnittene Computer-Architekturen effizient gemacht werden kann ('LISP-Maschinen'). Deren Stärke ist 'symbol crunching'.

PROLOG-Maschinen: Auch PROLOG-ähnliche Sprachen werden dazu verwendet, höhere Sprachen zur Manipulation komplexer Datenstrukturen zu implementieren. Japan bemüht sich, im Rahmen des '5th Generation Computer Project', effiziente, durch PROLOG inspirierte Maschinen zu entwickeln, deren Architektur echten Parallelismus ermöglichen wird (parallele Verfolgung von Alternativen und Konjunktionen in logischen Programmen). Der Prototyp einer solchen 'Parallel Inference Machine' (PIM) mit ca. 100 Sub-Rechnern wird 1988 fertiggestellt sein.

Connection Machine: In den letzten Jahren ist die Connection Machine bekannt geworden. Sie weist mehr Prozessoren (65K) als die leistungsfähigste der SIMD-Maschinen auf. Ihr hervorstechendstes Entwurfsmerkmal ist jedoch ihre Flexibilität, die darauf beruht, dass im Prinzip jeder Prozessor mit jedem anderen kommunizieren kann und dass die Konfigurierung dieser Verbindungen programmierbar ist. Einerseits ist die Connection machine als SIMD-Maschine konfigurierbar, andererseits wird die Implementation immer neuer Vision-Algorithmen auf ihr bekannt, die weit über die Low-Level-Vision hinausgehen. Darüber hinaus können mit ihr auch Semantische Netze dargestellt werden. Jedem Prozessor wird ein Konzept zugeordnet, die Verbindungen zwischen den Prozessoren stellen Relationen zwischen den Konzepten dar.

Mit der an sich ungeeigneten v.Neumann-Architektur ist man erstaunlich weit gekommen. Insbesondere ist es gelungen, mit ihr Computer-Vision-Systeme prinzipiell zu demonstrieren.

Ein mit heute diskutierten Multi-Computer-Konzepten im Einklang stehender Entwurf eines CVS würde jeden Prozess-Modul durch auf ihn zugeschnittene Hardware optimal realisieren. Diese Menge von heterogenen Subrechnern müsste in einer kohärenten Architektur unter einem Leit-Rechner integriert sein.

Die Connection machine lässt bereits eine andere Entwicklungsmöglichkeit erahnen. Die Konfigurierbarkeit der Kommunikationspfade zwischen den Prozessoren wird es wahrscheinlich erlauben, jede in einem CVS benötigte Modul-Architektur zu simulieren. Zusätzlich müsste noch die Möglichkeit geschaffen werden, eine sehr grosse Connection machine in Moduln zu partitionieren, die sich einerseits bei ihren spezifischen Tätigkeiten nicht gegenseitig stören, aber dennoch miteinander als Einheiten kommunizieren können.

Diese abschliessenden Bemerkungen lassen die Vorhersage zu, dass uns die Entwicklung von Computer-Vision-Systemen noch einige Dezennien beschäftigen wird.