

Kapitel 9

Sprach- und Bilderkennung

Sensoren erfassen physikalisch messbare Größen unserer Umwelt und liefern damit Signale, die von digitalen Rechnern automatisch verarbeitet und ausgewertet werden können (vgl. Kapitel 1). Die Signale können beispielsweise mit einem Mikrofon oder einer Digitalkamera aufgenommen werden. Eine mögliche Auswertung besteht darin, ein gesprochenes Wort zu erkennen, den Sprecher zu identifizieren oder Objekte in einem Bild zu identifizieren und zu lokalisieren, die auf einem Förderband liegen. Bei der Sprach- und Bilderkennung steht die Klassifikation und Analyse von Sprach- und Bildsignalen im Mittelpunkt.

9.1 Verarbeitung von Signalen

9.1.1 Digitale Sprach- und Bildsignale

Ausgangspunkt für die Sprach- und Bilderkennung sind Signale, die entsprechend vorverarbeitet werden müssen. Handelsübliche Computer verarbeiten Sprach- und Bildsignale nur in digitalisierter Form. *Analog-Digital-Wandler* (A/D-Wandler) konvertieren kontinuierliche Signale in diskrete Signale.

Digitalisierung Bei der *Digitalisierung* spielen zwei Größen eine wichtige Rolle: die *Abtastfrequenz* und die *Quantisierung* der Abtastwerte.

Die Qualität des digitalen Signals wird durch den *Signal-Rausch-Abstand* (engl. Signal to Noise Ratio, SNR) in Dezibel (dB) gemessen. Die *Abtastfrequenz* ist durch das Abtasttheorem vorgegeben (vgl. Kapitel 1), welches theoretisch die fehlerfreie Rekonstruktion des kontinuierlichen

Signals aus dem diskreten Signal sicherstellt. Nach dem Abtasttheorem muss die Abtastfrequenz mindestens gleich der doppelten Grenzfrequenz des zu digitalisierenden Signals sein (siehe Kapitel 7, 8). Wird die Abtastfrequenz zu niedrig gewählt, treten Artefakte auf (engl. Aliasing), die auch im Alltag zu beobachten sind: Zum Beispiel flimmern in Filmen gezeigte PC-Monitore oder die Speichenräder einer Kutsche scheinen sich rückwärts zu drehen. Beide Effekte sind auf eine Verletzung des Abtasttheorems zurückzuführen.

Für die Quantisierung gibt es kein zum Abtasttheorem vergleichbares theoretisches Ergebnis. Die Quantisierungskennlinie gibt an, welches Intervall im Kontinuierlichen auf welchen diskreten Signalwert abgebildet wird. Die Quantisierungskennlinie wird in aller Regel so gewählt, dass der mittlere quadratische Fehler zwischen kontinuierlichem und diskretem Signal minimiert wird. Dieser Fehler wird auch als *Quantisierungsrauschen* bezeichnet. Der Verlauf der Kennlinie hängt entscheidend von der statistischen Verteilung der Abtastwerte ab. Quantisierungsstufen sollten in Bereichen eng beieinander liegen, in denen viele Abtastwerte auftreten. Bei gleichverteilten Abtastwerten, ist die Quantisierungskennlinie sinnvollerweise linear [8].

In der Sprach- und Bildverarbeitung werden die Signale häufig transformiert, um annähernd gleichverteilte Signalwerte zu erhalten, die dann mittels einer linearen Quantisierungskennlinie quantisiert werden können. In der Sprachverarbeitung wird das Signal meist logarithmisch transformiert. Für die Lösung vieler Erkennungsaufgaben ist die Quantisierung nicht sehr wesentlich. In vielen industriellen Anwendungen werden für die Erkennung Binärbilder mit nur zwei Grauwerten verwendet.

□ *Beispiel:*

Sprachsignale in CD-Qualität werden mit 44 kHz abgetastet und mit 16 bit quantisiert. Gängige Bildgrößen für die Bildanalyse sind 256×256 , 512×512 und 1024×1024 . Die Quantisierung bei Bildern ist oft auf 8 bit pro Farbkanal beschränkt. In vielen medizinischen Anwendungen ist jedoch eine Quantisierung mit 16 bis 18 bit erforderlich. Moderne Grafikkarten wie beispielsweise die GeForce 6800 unterstützen bereits 16 bit pro Farbkanal.

- *Hinweis:* Sprach- und Bilddaten haben einen hohen Speicherbedarf, was einfache Zahlenbeispiele belegen. Ein 1024×1024 -Bild mit 16 bit Quantisierung benötigt bereits 2 MB. Der hohe Speicherbedarf macht also die Entwicklung und den Einsatz leistungsfähiger Kompressionsverfahren erforderlich. Für die Kompression von Sprachsignalen ist der MP3-Standard etabliert (siehe Kapitel 3). Einzelbilder werden häufig

unter Verwendung JPEG und Bildfolgen mittels MPEG[12] komprimiert (vgl. Kapitel 3).

Die Sprach- und Bilderkennung erfordert die Berechnung von Merkmalen aus den vorverarbeiteten Signalen, die einerseits eine deutliche Datenreduktion nach sich ziehen und andererseits die Trennung unterschiedlicher Klassen erlauben sowie die Analyse des Signals ermöglichen. Auf die Merkmalsextraktion folgt der Klassifikationsschritt. Die wesentlichen Module eines Systems für die Sprach- und Bilderkennung nach Niemann [8] sind in Bild 9.1.1 zusammengefasst.

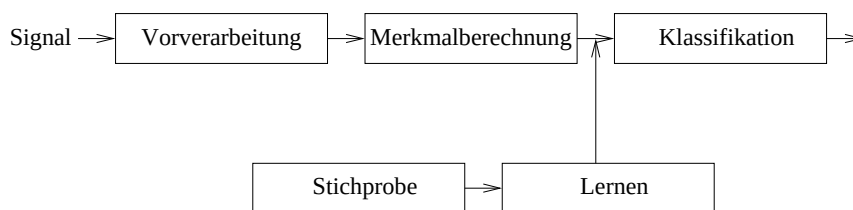


Bild 9.1 Module eines Klassifikationssystems (nach [8])

9.1.2 Merkmalberechnung — Beispiele aus Sprach- und Bildverarbeitung

Automatische Verfahren der Sprach- und Bilderkennung beruhen auf Merkmalen.

Merkmale. Merkmale sind für die Zugehörigkeit zu einer Klasse charakteristisch und eignen sich für die Klassifikation und Analyse.

Gute Merkmale zeichnen sich dadurch aus, dass sie für jede Musterklasse einen kompakten Bereich im Merkmalsraum einnehmen. Generell gilt die Faustregel, dass ein Klassifikator nur so gut sein kann wie seine Merkmale. Entsprechend vielfältig sind die Merkmale, die in der Literatur zu finden sind [8, 11, 13]. Merkmale werden sowohl aus dem Signal im Zeit- bez. Ortsbereich als auch aus dem Signal im Frequenzbereich berechnet, d.h. nach der Fouriertransformation des Signals (vgl. Kapitel 7). Oftmals folgt die Berechnung der Merkmale auf eine *Segmentierung* des Signals (vgl. Kapitel 7).

Etablierte Verfahren zur Merkmalberechnung im Zeitbereich berechnen beispielsweise:

Energie. Die Energie eines Sprachsignals f_k , $k = 1, \dots, N$ ist definiert durch:

$$E = \sum_{k=1}^N f_k^2 \quad (9.1)$$

Nulldurchgänge. Die Anzahl der Vorzeichenwechsel im Sprachsignal fasst das Merkmal Nulldurchgänge ZC zusammen:

$$ZC = \#\{k \mid f_{k-1} = 0 \vee f_k \cdot f_{k-1} < 0; k = 2, 3, \dots, N\} \quad (9.2)$$

Lineare Vorhersage. Bei der linearen Vorhersage (engl. linear predictive coding, LPC, vgl. Kapitel 3). werden einzelne Abtastwerte des Sprachsignals durch eine Linearkombination der n Vorgänger approximiert.

$$f_i = \sum_{k=1}^n c_k f_{i-k} \quad (9.3)$$

Die Gewichte $c_1, \dots, c_n$ der Linearkombination können mit der Methode der kleinsten Fehlerquadrate berechnet werden und bilden den n -dimensionalen Merkmalvektor. Statt der Signalwerte können in die lineare Vorhersage auch andere Merkmalsequenzen einfließen wie beispielsweise die Werte des Fourier Spektrums [11].

In der deutschen Sprache gibt es etwa 40 Phoneme, die sich am Besten im Frequenzbereich voneinander unterscheiden lassen. Zur Berechnung der Merkmale im Frequenzbereich wird die Kurzzeit-Fourieranalyse angewendet. Dabei wird das Sprachsignal in äquidistante Bereiche (engl. Frame) unterteilt. Für jeden Frame wird die *diskrete Fouriertransformation* berechnet (vgl. Bild 9.1.2). In der Praxis hat sich bewährt, spektrale Merkmale alle 10 ms in einem Zeitfenster von 25 bis 80 ms zu berechnen. Für jedes Zeitfenster wird ein Merkmalvektor mit 10 bis 50 Komponenten berechnet. Praktisch relevante Merkmale sind das Spektrum, das logarithmierte Spektrum, Spektrogramme, das Cepstrum oder die Mel-Cepstrum-Koeffizienten sowie diskrete Ableitungen. Einen umfassenden Überblick über Merkmale für die Spracherkennung gibt [11].

Merkmalfolgen für die Spracherkennung. Merkmale für die Spracherkennung werden üblicherweise im Frequenzbereich berechnet. Für jedes Sprachsignal wird dabei nicht nur ein einzelner Merkmalvektor berechnet, sondern eine Folge von Merkmalvektoren.

□ *Beispiel:*

In Bild 9.2 ist ein Sprachsignal dargestellt. Das Sprachsignal wird in Zeitfenster zerlegt. Für jedes Zeitfenster wird ein Kurzzeitspektrum berechnet und daraus ein Merkmalvektor abgeleitet.

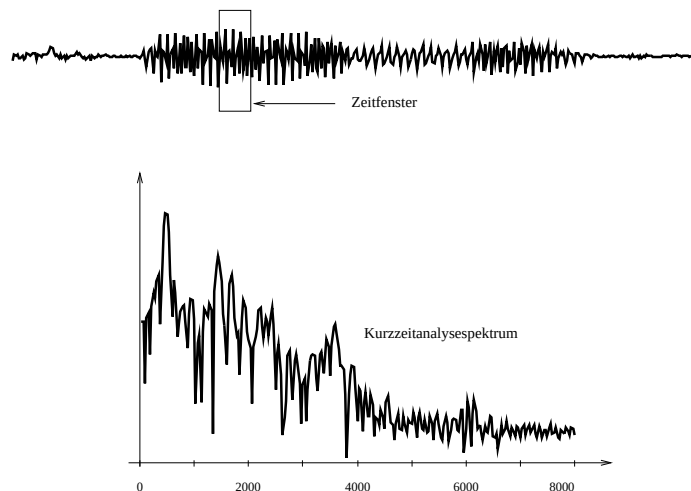


Bild 9.2 Vom Sprachsignal zum Merkmal

In der Bilderkennung werden häufig geometrisch motivierte Merkmale verwendet, die aufbauend auf eine Segmentierung des Bildes berechnet werden wie Eckpunkte, Kanten, Texturen, Regionen und deren Flächen. Alternativ zu den geometrischen Kenngrößen kommen auch Merkmale zum Einsatz, die keine Segmentierung voraussetzen und die direkt aus den Intensitäten der einzelnen Bildpunkte berechnet werden. Beispiele hierfür sind Grauwert- oder Farbhistogramme oder das Ergebnis einer Hauptachstentransformation des Bildvektors [13].

Merkmalmengen für die Bilderkennung. Für die Bilderkennung sind Merkmale üblicherweise geometrische Kenngrößen oder sie werden direkt aus den Intensitäten berechnet. Zu einem Bild kann eine Merkmalmenge berechnet werden. Dabei ist oft unbekannt, welche Merkmale zum Objekt oder zum Objekthintergrund gehören. Die Kardinalität der Merkmalmenge variiert mit den Bildern.

- *Beispiel:* Bild 9.3 zeigt ein Beispiel für eine Aufnahme, in der die einzelnen Objekte erkannt werden sollen. In dem Bild wurden mittels eines Eckendetektors Punktmerkmale berechnet, die für eine spätere Klassifikation und Lokalisierung verwendet werden können.

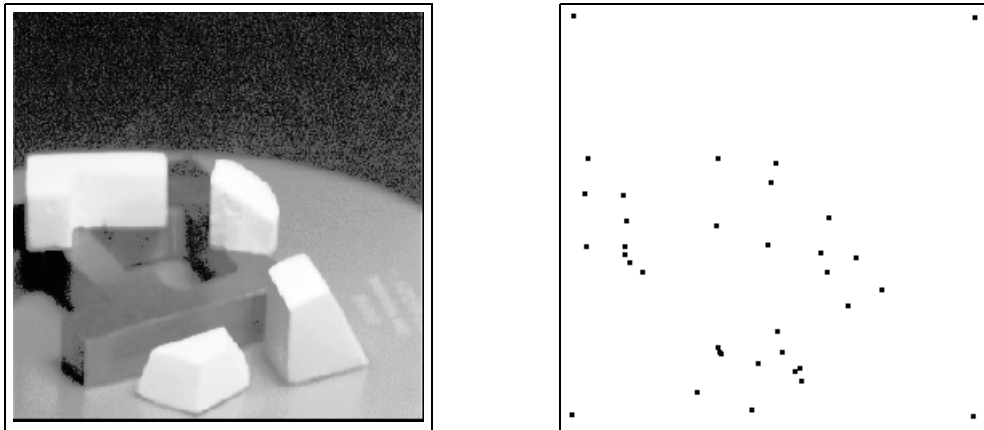


Bild 9.3 Vom Bild zu Punktmerkmalen

9.1.3 Klassifikation und Lokalisierung

Klassifikation. Bei der Klassifikation wird ein Merkmal c , das aus dem ursprünglichen Signal f berechnet wird, einer Musterklasse κ aus einer endlichen Menge diskreter Musterklassen zugeordnet oder zurückgewiesen, falls es nicht klassifizierbar ist.

In der Bildererkennung ist die Identifikation eines Objekts in einer Szene ein Beispiel für ein Ergebnis der Klassifikation. Eine Rückweisungsklasse ist für viele Anwendungen wichtig, da Bilder oft Objekte enthalten, die nicht bekannt sind, oder Wörter gesprochen werden, die nicht zum Wortschatz gehören.

Die Klassifikation erfolgt unter Zuhilfenahme einer *Entscheidungsregel*. Jede Entscheidungsregel wird mit einer Kostenfunktion assoziiert, die jeder Entscheidung Kosten zuordnet.

Bayes-Klassifikator. Ein wichtiges theoretisches Ergebnis der Mustererkennung besagt, dass die Bayes-Entscheidungsregel die Wahrscheinlichkeit für eine Fehlklassifikation minimiert.

Bei der Entscheidung für eine Klasse wird beim *Bayes-Klassifikator* die a-posteriori Wahrscheinlichkeit maximiert. Es wird also die Musterklasse κ gewählt, bei der die bedingte Wahrscheinlichkeit am größten ist, mit der die Klasse κ beim Auftreten des Merkmals c zutrifft:

$$\kappa = \operatorname{argmax}_{\lambda} p(\lambda|c) = \operatorname{argmax}_{\lambda} p(\lambda)p(c|\lambda) \quad . \quad (9.4)$$

Der Bayes-Klassifikator setzt die Kenntnis der a-posteriori Wahrschein-

lichkeiten voraus. Die benötigten Dichtefunktionen können mittels parametrischer und nicht-parametrischer Schätzverfahren aus repräsentativen Stichproben berechnet werden [8]. Ein sehr einfach zu realisierender und häufig angewendeter Klassifikator ist der *Nächster Nachbar Klassifikator* (NN-Klassifikator). Hierbei wird im Rahmen einer Lernphase eine repräsentative Menge klassifizierter Merkmalvektoren berechnet.

Nächster Nachbar Klassifikator. Der NN-Klassifikator entscheidet sich für die Klasse, der das Merkmal angehört, das den geringsten Abstand zum beobachteten Merkmal hat.

Eine umfassende Beschreibung der Theorie der Klassifikatoren findet man in [8].

- *Beispiel:* Das Klassifikationsergebnis für die in Bild 9.3 gezeigten Objekte könnte für bestimmte Anwendungen mit den Objektklassen 1, 2, ..., 20 lauten: In der Szene befinden sich vier polyedrische Objekte der Objektklassen 2, 7, 9 und 10.

Klassifikatoren der Mustererkennung erwarten im Regelfall einzelne Merkmalvektoren. In der Spracherkennung wird für jedes Signal nicht nur ein Merkmal, sondern eine Folge von Merkmalen berechnet. Bei der Erkennung basiert die Klassenzuordnung auf einer Merkmalfolge und nicht auf einem einzelnen Merkmalvektor. Die Merkmalfolgen können dabei unterschiedliche Länge haben, denn eine Äußerung kann je nach Sprechgeschwindigkeit unterschiedlich lange dauern. Die Skalierung der Äußerung ist zudem nichtlinear, da Wörter unterschiedlich gedehnt werden. Neben der Ähnlichkeit einzelner Merkmale muss bei der Klassifikation von Sprachsignalen beispielsweise mittels des NN-Klassifikators zusätzlich eine geeignete Zuordnung der Merkmale aus dem Referenz- und dem beobachteten Signal berechnet werden. Die Berechnung der optimalen Zuordnung erfolgt in diesem Fall unter Verwendung der Dynamischen Zeitverzerrung (engl. Dynamic Time Warping, DTW), das auf dem Konzept der Dynamischen Programmierung beruht.

DTW eignet sich für die einfache, sprecherabhängige Erkennung. In komplexeren Spracherkennungssystemen wird heute mit *Hidden Markov Modellen* (HMM) ein wahrscheinlichkeitstheoretischer Ansatz genutzt.

Hidden Markov Modell. Ein HMM ist ein endlicher überführender stochastischer Automat, dessen Zustände die einzelnen Laute in Folge repräsentieren. In ihm werden zwei stochastische Prozesse gekoppelt, die sowohl die Variationen in den Lauten (Ausgabewahrscheinlichkeiten) als auch die unterschiedlichen zeitlichen Verzerrungen (Übergangswahrscheinlichkeiten) modellieren (s. Bild 9.4).

HMMs erlauben die Anwendung des Bayes-Klassifikators, wobei in die a-priori Wahrscheinlichkeiten des Sprachmodells (s. 9.2.4) einfließen. Bei der Erkennung entscheidet sich der Klassifikator für das wahrscheinlichste Wort.

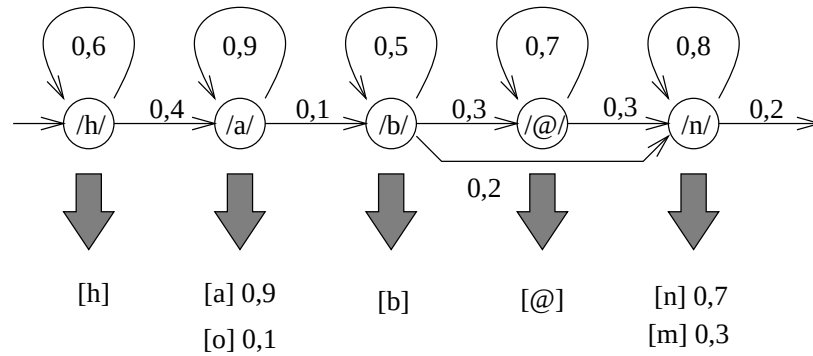


Bild 9.4 Einfaches HMM für das Wort „haben“.

Nicht nur in der Spracherkennung, sondern auch in der Bildverarbeitung und im Rechnersehen reicht häufig ein einzelner Merkmalvektor nicht zur Lösung der Klassifikationsaufgabe aus. Für ein Objekt können markante Punkte wie Ecken oder Kanten berechnet werden, die für den Vergleich mit Modellen der erkennbaren Objekte herangezogen werden. Eine weitere Schwierigkeit resultiert aus der Tatsache, dass die Bilder für die Klassifikation Abbildungen der dreidimensionalen Welt sind. Ein optisches System und eine Projektion bilden die Objekte in die zweidimensionale Bildebene ab.

Um bestimmte Objekte z.B. mit einem Roboterarm greifen zu können, muss neben der Klassifikation auch die Lokalisierung durchgeführt werden.

Lokalisierung. Bei der Lokalisierung wird die Lage des Objektes berechnet, d. h. die Rotation und Translation bezüglich eines Referenzkoordinatensystems.

Die Rotationswinkel und die Komponenten des Translationsvektors sind kontinuierliche Variablen. Die Lokalisierung entspricht damit einem Regressionsproblem:

$$\hat{\theta} = \operatorname{argmax}_{\theta} p(\theta) p(c|\theta) \quad (9.5)$$

wobei θ die zu schätzenden Lageparameter bezeichne. Zur Lösung dieses Problems können bekannte Verfahren der numerischen Optimierung eingesetzt werden [4].

9.1.4 Analyse

Für viele Anwendungen reicht die Klassenzuordnung und die Lokalisierung nicht aus, sondern es muss eine Analyse des Musters durchgeführt werden. Das Analyseziel wird dabei durch die konkrete Anwendungsdomäne definiert. Eine einfache Klassifikation in Form eines Worterkenners reicht sicher nicht aus, um eine natürlichsprachliche Anfrage an ein Zugauskunftssystem zu interpretieren. Es müssen die Uhrzeit, der Abfahrts- und der Zielort erkannt und entsprechend verstanden werden. Bei unvollständigen Anfragen muss das System in der Lage sein, entsprechende Rückfragen zu stellen. Ähnlich verhält es sich bei Analysebeispielen in der Bildverarbeitung. In der Radiologie werden beispielsweise Computertomographieaufnahmen zur Diagnose unterschiedlicher Krankheiten eingesetzt. Es sollen Läsionen ebenso erkannt und lokalisiert werden wie Ablagerungen in den Gefäßen oder Blutungen. Ziel ist hierbei die vollständige Beschreibung des Krankheitsbildes.

Analyse. Bei der Analyse von Sprache und Bildern werden die mit Sensoren aufgenommenen Signale in eine symbolische Beschreibung transformiert.

9.2 Spracherkennung und -verarbeitung

Ziel der **Sprachverarbeitung** ist es, Strukturen und Prozesse der menschlichen Sprachpraxis mithilfe expliziter Verarbeitungsmodelle nachzubilden und technisch verfügbar zu machen. Insbesondere geht es darum, wie sprachliche Äußerungen in Strukturen überführt werden können, die ihren Inhalt im Kontext repräsentieren, und die für vielfältige weitere Verwendungen geeignet sind. Grundlagen der Sprachverarbeitung sind die Konstruktion einschlägiger Algorithmen und Datenstrukturen sowie Untersuchungen der Komplexität (s. [10], 2.3, 8.1.3).

Die Spracherkennung i.e.S. befasst sich mit der *Identifikation* linguistischer Einheiten in einem Sprachsignal. Diese erfolgt üblicherweise in einem mehrstufigen Prozess, wobei auf jeder Ebene der Verarbeitung Schritte der *Segmentierung* und *Klassifikation* auszuführen sind.

Eine derart komplexe Aufgabe ist nur zu bewältigen, wenn man sie in leichter überschaubare Teilaufgaben zerlegt, hier in zwei Phasen, eine *zeit-orientierte*, die direkt auf dem Sprachsignal arbeitet, und eine *struktur-orientierte*, die auf abstrakten Objekten wie Silben, Wörtern,

Konstituenten, Sätzen und satzübergreifenden Äußerungen operiert.¹

Die von Spracherkennern i.e.S. geleistete Zuordnung einer Folge von Wortformen zu einem Sprachsignal beinhaltet schon einen Interpretationsschritt. Die „Verschriftlichung“ von Sprachsignalen führt zu einer wichtigen Klasse von Anwendungen, den Diktiersystemen, die heute bereits als kommerzielle Produkte verfügbar sind. Mit zunehmender Komplexität der Anwendungen reicht dies jedoch nicht aus; es kommt auf eine formale, maschinell verarbeitbare Darstellung des kontextbezogenen Inhalts des Gesagten oder Geschriebenen an, z.B. für

- inhaltsbezogene Suche in sprachlichen Korpora,
- Auskunft- und Verhandlungsdialoge mit Informationssystemen,
- Steuerung elektronischer Geräte, etwa am Arbeitsplatz, in Kfz oder im Haushalt.

Dies setzt eine *Operationalisierung* und *Simulation* der menschlichen sprachlichen Kommunikation voraus, also der Interpretation und Erzeugung sprachlicher Mitteilungen mit formalen Mitteln. M.a.W.: Es geht um eine Nachbildung des Sprachhandelns mit dem Instrumentarium formaler Sprachen und statistischer Verfahren.

9.2.1 Modellierung des sprachlichen Kommunikationsprozesses

Schon seit der Antike wird die menschliche Sprache als Werkzeug angesehen, mittels dem *einer* dem *anderen etwas* mitteilt (Platon). Dies wird ergänzt durch die moderne Auffassung, dass Sprachhandeln primär auf die Verständigung innerhalb gemeinsamen Handelns zielt: „meaning as use“ (Wittgenstein). Demnach ist die Sprache ein Medium, das zugleich drei verschiedene, aber miteinander verknüpfte Funktionen erfüllt: Kommunikativ verwendete sprachliche Äußerungen dienen dazu, Intentionen auszudrücken, Sachverhalte mitzuteilen und Beziehungen zwischen Re-departnern herzustellen.

Die begrifflichen Voraussetzungen der computerlinguistischen Modellierung können v.a. auf den Anfang des 20. Jh. entstandenen „Strukturalismus“ (F. de Saussure) zurückgeführt werden. Dieser thematisiert

¹Für alle sprachwissenschaftlichen und computerlinguistischen Termini sei verwiesen auf [1]; dort finden sich weitere Hinweise auf Originalpublikationen und weiterführende Literatur. Das z.Zt. modernste und umfassendste deutschsprachige Lehrbuch der Computerlinguistik ist [2].

nicht den Aspekt des kommunikativen Handelns, sondern weitaus eingeschränkter die Sprache isoliert als „System“ von Zeichen, wobei zwischen der Sprache („*le langage*“) als abstrakter allgemeiner Sprachfähigkeit, den verschiedenen Sprachen („*les langues*“) und dem Vollzug des Sprechens („*la parole*“) unterschieden wird. Das sprachliche Zeichen selbst ist im Prinzip willkürlich, ebenso wie die in ihm getroffene Zuordnung zwischen Lautbild („der Bezeichnung“) und Begriff („dem Bezeichneten“). Weiterhin war die Theorie der Generativen Transformationsgrammatik (N. Chomsky) sehr einflußreich. Formale Kalküle und die mit ihnen begründete operative Logik dienen hier als Basis der linguistischen Modellierung; sie ist damit der theoretischen Grundeinheit des Satzes, der Wahrheitswertsemantik und der Behauptung, dass sprachliche Zeichen eine außersprachliche Realität abbilden, verhaftet. Konstitutiv ist die Idee einer mehrstufigen Übersetzung von Sätzen in eine „Logische Form“; es wird also nur die o.g. Funktion der Mitteilung von Sachverhalten erfaßt.

Bei jeder Reduktion der Umgangssprache auf eine formale ist der Tatsache Rechnung zu tragen, dass sie eine Reihe von Eigenschaften aufweist, die in formalen Sprachen von vorneherein gar nicht vorkommen: Sie verfügt über ein reichhaltiges Repertoire an semantisch weitgehend fixierten *Grundelementen* (Wortschatz) und generellen *Ausdrucksmit-teln* (grammatische Formen und Funktionen, Referenzmechanismen, etc.) und kennt eine Vielfalt von Äußerungsformen. Die Ausdrucksmittel der „natürlichen“ Sprachen sind primär auf die aktuelle Herstellung von *Sachbezügen* angelegt, d.h. auf die *situationsabhängige* Charakterisierung *sinnlich wahrnehmbarer* Gegenstände und Sachverhalte der Umwelt, die sie dann auf eine objektivierte Darstellung *abstrakter* Beziehungen und *verallgemeinerter* Aussagen zu übertragen gestattet. Zugleich dient sie auch als *Metasprache* für sich selbst (Selbstreferenz).

Ein zentrales Problem der Sprachverarbeitung ist die **Ambiguität** auf allen sprachlichen Ebenen, d.h. die Mehrdeutigkeit von Wortbedeutungen, von grammatischen Konstruktionen, von Bedeutungen zusammengesetzter Ausdrücke und der pragmatischen Deutung von Redeabsichten (Intentionen). Große Schwierigkeiten birgt auch die *Vagheit* sprachlicher Äußerungen, sowie ihre Möglichkeiten der *elliptischen Verkürzung* und des *metaphorischen Sprachgebrauchs*.

9.2.2 Besonderheiten der Sprachverarbeitung

In der Sprachverarbeitung wird das „Sprechen-Können“ („*knowing how*“), die Sprachkenntnis, als „Sprach-Wissen“ („*knowing that*“) repräsentiert,

d.h., sie wird zur Verarbeitung formalisierten linguistischen *Wissens*. Das linguistische Wissen selbst ist in Abstraktionsschichten unterteilt, wodurch eine Organisation in Verarbeitungsebenen bestimmt wird.

Der zeit-orientierte Teil der Verarbeitung wird oft auch als **Signal-Symbol-Transformation** bezeichnet: Zunächst müssen Signale in Sprachlaute darstellende Zeichenfolgen transformiert werden, die geeignet zu segmentieren sind, so dass man *diskrete, sequentiell angeordnete* phonetische Einheiten erhält, denen dann Wortformen zugeordnet werden (s. 9.2.3, 9.2.4). Dabei schließt jede Identifikation eine *Interpretation* ein, was schon aus der Definition des Phonems ersichtlich ist:

Phoneme sind aus dem Sprachsignal abstrahierte lautliche Segmente mit potentiell bedeutungsunterscheidender Funktion.

Das Ziel der strukturorientierten Phase besteht darin, die symbolischen Darstellungen von Äußerungen nun wiederum zu segmentieren zur Bestimmung ihrer sinntragenden Teile. Auf die Signal-Symbol-Transformation folgend werden Beschreibungen

- der *Anordnung* in grammatischen Einheiten: **Syntax**, einschliesslich **Phonologie** und **Morphologie** (Wortgrammatik),
- des *Inhalts* („Sinn und Bedeutung“): **Semantik**, und
- des *Gebrauchs*: **Pragmatik**

erzeugt. Dabei können von einer Verarbeitungsebene Teilergebnisse an die nächste weitergegeben werden, so bald sie vorliegen, so dass auch Rückkopplungen möglich sind und die Verarbeitung insgesamt eine gewisse Parallelität erreicht.

9.2.3 Verarbeitungsschritte auf Sprachsignalen

Um eine angemessene Analyse von Sprachsignalen durchführen zu können, ist es notwendig, ihrer Entstehung nachzugehen. Der menschliche Sprechapparat, der mit dem Atmungssystem verbunden ist, verfügt über eine Vielzahl von Möglichkeiten zur Lautbildung. Die Stimmbänder vibrieren und erzeugen Töne, deren Höhe wie bei einem Saiteninstrument durch die Länge und Spannung der Stimmbänder bestimmt wird. Der Klang dieser Töne wird beeinflusst durch Bewegungen von Zunge, Unterkiefer und Lippen, womit wir die Form des Stimmtrakts verändern. Der Stimmtrakt ist ein Resonanzraum, dessen harmonische Resonanzen bestimmte Obertöne verstärken und andere vermindern. Eine weitere Veränderung des Resonanzraums wird durch das Gaumensegel bewirkt, das die

Nasenhöhle öffnet und schließt. Darüberhinaus können wir durch das Strömen von Luft durch Mund- und Rachenhöhle Geräuschaute erzeugen. Die Vielfalt der menschlichen Sprachlaute resultiert aus der Kombination dieser verschiedenen Prozesse, die durch einen komplizierten Bewegungsablauf koordiniert werden: stimmhafte (Vokale wie a, e) und stimmlose (Zischlaute wie s, z) Laute, konsonantische Verschlusslaute (Plosive, wie p, k) und Nasale (wie m, n). Mit dem *source-filter*-Modell hat G. Fant eine Modellierung der Sprachsignalerzeugung entwickelt, die die wesentlichen Komponenten des Artikulationsapparats abbildet (s. Abb. 9.4). Zur Bestimmung der Merkmale der Laute hat sich in erster Näherung das „Linear Predictive Coding“ (s. 9.1.2) als geeignetes technisches Verfahren erwiesen.

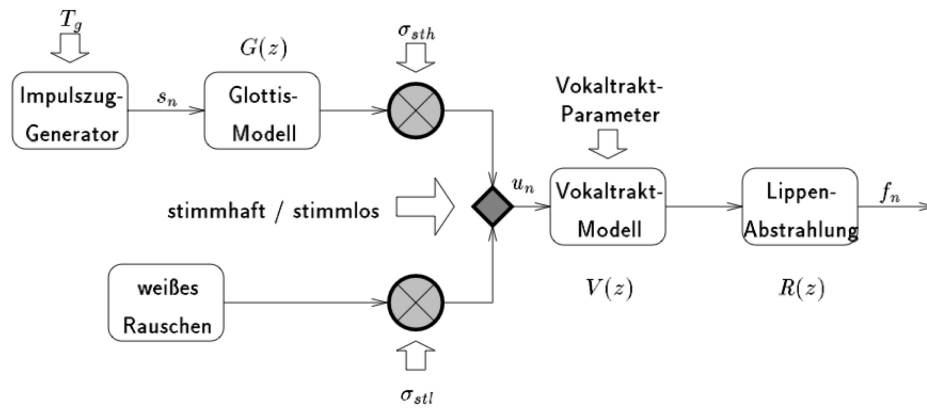


Bild 9.4 Das *source-filter*-Modell der Spracherzeugung nach G. Fant (aus [11], S. 33)

Spracherkennung kann *sprecherunabhängig* oder, mit weitaus höherer Genauigkeit, *sprecherabhängig* erfolgen, wobei im letzteren Fall die Charakteristika der Sprecherstimme in einer vorgeschalteten Trainingsphase erfasst werden. Dies ist aber in vielen Fällen, etwa bei Auskunftssystemen, nicht praktikabel. Einen Mittelweg bilden *sprecheradaptive* Systeme, die sich allmählich an die Sprecherstimme anpassen.

Der kontinuierliche Strom von Lauten wird zunächst durch ein Mikrofon aufgenommen, welches ein analoges elektromagnetisches Signal liefert, das in einem ersten Verarbeitungsschritt digitalisiert wird. Für die Güte der Spracherkennung ist die Aufnahmequalität des Signals von entscheidender Bedeutung. Das digitalisierte Signal wird dann unter Berücksichtigung der Eigenschaften des menschlichen Hörsystems (s. [6], 5.2) in eine Reihe unterschiedlicher Frequenzbänder zerlegt und normalisiert, u.a.

zum Angleichen der Lautstärke. Die menschliche Sprachwahrnehmung ist *kategorial*, d.h. im Kontinuum erzeugter Laute werden zwischen 20 und 60 Lautklassen unterschieden. Für die weitere Verarbeitung erfolgt eine Segmentierung in kleine Abschnitte (*Frames*) (s. 9.1.2). Darin werden bestimmte Merkmale gesucht, z.B. Frequenzwechsel und Energiedifferenzen (Stille). Insgesamt handelt es sich um einen vieldimensionalen Merkmalsraum, in dem die Lauterkennung eine komplexe Musterklassifikationsaufgabe darstellt (s. 9.1.3). Um der enormen Variationsbreite der akustischen Realisierung Rechnung zu tragen, werden verschiedene Transformationen auf den Signaldaten durchgeführt; z.B., um die Effekte unterschiedlicher Sprechgeschwindigkeit oder Wechselwirkungen mit Nachbarlauten zu berücksichtigen.

9.2.4 Kontinuierliche Spracherkennung

Eine besondere Schwierigkeit der Worterkennung in fließender Rede besteht darin, dass Wortgrenzen nicht markiert sind. Zur Erkennung der aus Lautkombinationen gebildeten Wörter wird Wissen aus Phonetik, Phonologie und Wortbildung einbezogen — Laute können nicht in beliebiger Reihenfolge erscheinen, sondern es gibt eine Vielfalt von Einschränkungen der möglichen Kombinationen, die von Eigenschaften des Artikulationsapparats bis zu Bildungsregeln von Silben und Wörtern reichen. Da eine unübersehbare Zahl von Parametern — u.a. umgebungsspezifische, sprecherspezifische, anatomische, linguistische — die Formung von Sprachsignalen beeinflusst, ist jede Klassifikation mit Unsicherheit behaftet. Daher ist es unumgänglich, statistische Verfahren einzusetzen, die mit Hilfe von Hand klassifizierter Sprachdaten trainiert werden. Hierzu gehören auch künstliche neuronale Netze, die in der Spracherkennung mit großem Erfolg eingesetzt werden. Einen schematischen Überblick über den gesamten Prozess der Spracherkennung gibt Abb. 9.5, die auch die für die einzelnen Stufen jeweils relevanten wissenschaftlichen Disziplinen benennt.

Die akustischen Wahrscheinlichkeiten verbinden die Folge der normierten Merkmalvektoren mit den im Lexikon des Erkenners festgelegten Wortformen, die als Folge von Phonemen beschrieben sind. Das Aussprachelexikon selbst wird zur Minimierung der Redundanz baumförmig angelegt. Die Wahrscheinlichkeitsverteilungen der Merkmalsvektoren für die Realisierung der Phoneme werden üblicherweise durch stochastische endliche Automaten, die Hidden-Markov-Modelle (HMM, s. 9.1.3), modelliert. Zur Bestimmung der „akustischen“ Wahrscheinlichkeit einer Wortform nutzt

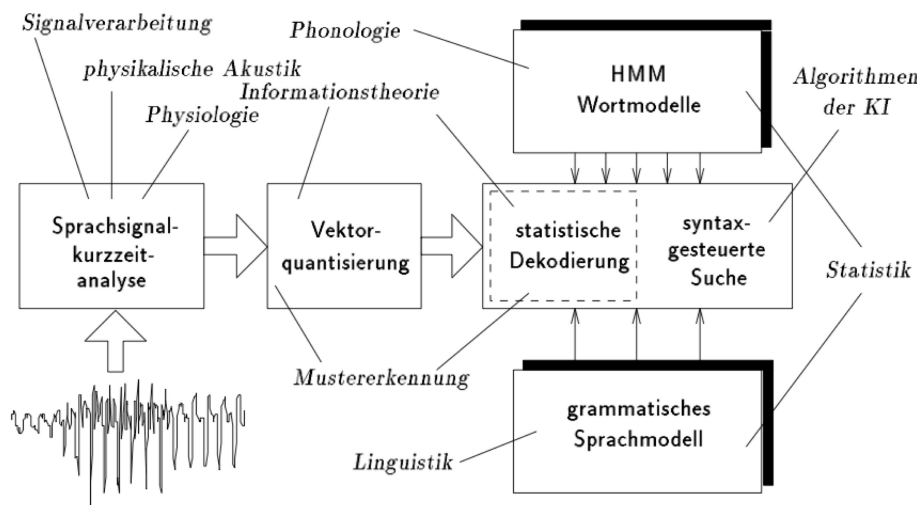


Bild 9.5 Automatische Spracherkennung als interdisziplinärer Prozess (aus [11], S. 15)

man die dynamische Programmierung (s. 9.1.3).

Diese wird komplementiert durch die „linguistische“ Wahrscheinlichkeit von Wortfolgen, die in einem *Sprachmodell* festgehalten ist. Bisher ist es noch nicht gelungen, ausschliesslich regelbasierte Sprachmodelle zu entwickeln, die erfolgreich zur Erkennung eingesetzt werden können. Daher geht man im einfachsten Fall davon aus, dass sich die Wahrscheinlichkeit für einen Satz als Produkt der bedingten Wahrscheinlichkeiten der Wortformen ergibt, aus denen er besteht. Da diese durch Training mit einem Korpus von Sätzen bestimmt werden, muss man sich, um hinreichend viele Beispiele vorzufinden, auf kleine Wortnachbarschaften beschränken, üblicherweise Paare (Bigramme) oder Tripel (Trigramme). In der neueren Forschung wird versucht, diese rein korpusbasierte Vorgehensweise durch den Einbezug grammatischer Regeln zu ergänzen. Als Mass für die Schwierigkeit des Erkennungsprozesses dient die *Perplexität*.

Die **Perplexität** hält fest, wie viele Wortformen im Mittel in Frage kommen, wenn die Vorgänger bekannt sind.

Die Erkennungsaufgabe ergibt sich damit als ein äußerst komplexes Such- bzw. Optimierungsproblem: Es ist aus den ermittelten Merkmalsvektoren unter Ausnutzung der beiden stochastischen Modelle die gesprochene Wortfolge zu ermitteln. Sie wird approximiert durch die Folge von Wortformen, die das Produkt aus akustischer und linguistischer Wahrscheinlichkeit maximiert. Zur Lösung unter Echtzeitbedingungen dienen

effiziente Sucherverfahren wie die *Viterbi-Suche* und die daraus abgeleitete *Strahlsuche* (*beam search*) und die *A*-Suche*. Im Idealfall besteht also das Ergebnis aus einer Folge mit einem Qualitätsmass versehener Wortformen, die die beste Hypothese zu dem gegebenen Sprachsignal darstellt; bei größeren Wortschätzen hat man stattdessen mit sog. Worthypothesengraphen zu tun, die in gleichen Zeitintervallen erkannte bewertete Worthypothesen in einem Netzwerk zusammenfassen.

Eine Besonderheit der gesprochenen Sprache besteht darin, dass sie Informationen auf verschiedene Weisen codieren kann, insbesondere auch durch die Intonation. Hierzu die prosodische Analyse, die sich mit segmentübergreifenden Ereignissen in Sprachsignalen befasst und diesen funktionale Eigenschaften zuordnet. Hierbei werden u.a. vier Arten von grenzmarkierenden Merkmalen zur Gliederung einer Äußerung („*Gliederungssignale*“) unterschieden: Intonation, Pausen, lexikalische Merkmale, Konstruktionswechsel (Satzbrüche und Neuansätze) und Wiederholungen von Wörtern oder Satzteilen. Außer Phrasengrenzen zeigt die Betonung auch unterschiedliche Äußerungsmodi und emotionale Aspekte in Sprachsignalen an. Weiterhin kommen in Dialogen auch *Sprecher-* und *Hörersignale* vor, die durch sog. Kontaktwörter realisiert werden und die eine Vergewisserungs-, Bestätigungs- oder Ablehnungsfunktion haben.

9.2.5 Linguistische Verarbeitung: Lexikon und Syntax

So bald eine Wortfolge — entweder direkt oder aus dem Worthypothesengraphen extrahiert — vorliegt, kann die erste Stufe der strukturorientierten Phase ausgeführt werden, die die sog. *grammatischen Abbildung* realisiert.

Die **grammatische Abbildung** bestimmt für jede zulässige Zeichenkette grammatische Relationen, sodass jeder wohlgeformten sprachlichen Äußerung eine — oder im Fall der Ambiguität mehrere — strukturierte Beschreibungen zugeordnet werden können, die ihren Sinn repräsentieren.

Zunächst werden die als diskrete Einheiten vorliegenden Wortformen in Wortstamm und Endungen zerlegt und ihnen eine *Kategorie* zugeordnet. Dann werden die kategorisierten Einheiten in sog. *Konstituenten* gruppiert und diesen dann spezifische und wohlunterschiedene *grammatische Funktionen* zugeschrieben, z.B. Subjekt, Prädikat oder Objekt.

Allerdings müssen bei gesprochener Spracheingabe die in der primär text-

orientierten Computerlinguistik getroffenen idealisierten Annahmen angesichts der heute praktisch erreichten Wortakkuratheit von bestenfalls 80% relativiert werden. Die beste erkannte Wortkette bzw. der beste Pfad durch einen Worthypothesengraphen wird Fehler erhalten, so dass eine die gesamte Äußerung überspannende syntaktische Zerlegung nicht erwartet werden kann. Daher arbeitet man in der Praxis mit hybriden symbolisch-stochastischen Verfahren und konzentriert sich auf die Erkennung gut erkannter Fragmente („Chunk Parsing“), die dann unter Berücksichtigung semantisch-pragmatischer Kriterien weiter verarbeitet werden.

Lexikon und Morphologie.

Ein **Lexikon** ist eine um linguistische Annotationen angereicherte Zusammenstellung des Wortschatzes einer Sprache.

Lexika für die maschinelle Sprachverarbeitung benutzen als abstrakte Basiseinheit das *Lexem* (i.a. der Wortstamm); zu jedem Lexem gibt es einen Lexikoneintrag, der Daten für die verschiedenen sprachlichen Ebenen enthält:

- *Morphologie*: Daten zur Wortformenbildung;
- *Syntax*: wortspezifische Daten (z.B. die Wortkategorie wie Nomen, Verb, etc.) einschließlich spezifischer Beschränkungen zum Vorkommen in grammatischen Konstrukten (sog. Valenzen);
- (*Lexikalische*) *Semantik*: Daten zur „wörtlichen“ semantischen Interpretation.

Die **Morphologie** ist die Lehre von den Klassen und Formen der Wörter mit Beugung (*Flexion*), Ableitung (*Derivation*) und Zusammensetzung (*Komposition*).

Hierfür haben sich Verarbeitungsmodelle auf der Basis endlicher überführender Automaten bewährt. Ein Lexemlexikon gestattet die am wenigsten redundante Darstellung des Wortschatzes, denn die wortspezifischen Daten müssen nur einmal, eben beim Lexem, gespeichert werden; die wortformenspezifischen Daten liefert die morphologische Komponente (s. [5]).

Der erste Schritt bei der Realisierung der *grammatischen Abbildung* ist die Identifikation der Wörter einer Äußerung als grammatische Objekte, d.h. die Zuordnung von grammatischen Beschreibungen (Wortkategorie — z.B. Verb, Nomen, Adjektiv, Adverb, Präposition — und Ausprägung der syntaktischen Merkmale — Person, Numerus, Genus, etc.) zu den Wörtern. Dies als eigene Verarbeitungsstufe voranzustellen hat einen

technischen Grund: Ein Lexikon als separate Komponente ist leichter handhabbar.

Grammatische Strukturanalyse. Darauf aufbauend erfolgt die Zerlegung von Wortketten in ihre Teile sowie die Zuordnung struktureller Beschreibungen, die die grammatischen Eigenschaften und Beziehungen dieser Teile untereinander darstellen. Eine Grammatik wird allgemein als ein System von Kategorien, Funktionen und Relationen verstanden, das mit Hilfe von Einschränkungen — aufbauend auf Anordnung und syntaktischen Merkmalen der Wörter — eine Klasse von Abhängigkeitsstrukturen zwischen sprachlichen Objekten definiert. Als analytisches Instrument dienen Systeme von *Kombinations-* und *Verwendungsregeln*.

Erstes Teilziel der Analyse (auch engl. „*Parsing*“ von lat. „pars orationis“ = Redeteil) ist die Zerlegung eines Satzes in eine Struktur von grammatischen *Konstituenten*, auch „Phrasen“ genannt, z.B. Verbalphrase, Nominalphrase, Präpositionalphrase (s. [7, 2]). Die Grundbeziehung dieser Zerlegung, die *Konstituenz*, ist eine Teil-Ganzes-Relation, die jeweils festlegt, *wie* systematisch aus sprachlichen Einheiten, beginnend mit den Wörtern, immer größere Einheiten zusammengesetzt sind. Für eine bestimmte Sprache werden die Kategorien der Konstituenten, die Anordnung ihrer Teile und deren Stellung im Ganzen durch eine Konstituentenstrukturgrammatik vorgeschrieben. Die über die Zerlegung gewonnene Strukturinformation wird üblicherweise durch *Baumdiagramme* dargestellt, wobei die Wurzel des Baums mit der Kategorie „Satz“ markiert ist und seine Blätter die Wörter des Satzes in ihrer zeitlinearen Abfolge sind.

Die Grundlagen der algorithmischen Durchführung der grammatischen Zerlegung liefert die Theorie der formalen Sprachen, in der Bildungsgesetze für Zeichenketten untersucht werden. Chomsky hat, ausgehend von unbeschränkten Regelsystemen, durch immer stärkere Einschränkung der Form der Regeln eine Hierarchie von Klassen formaler Sprachen definiert: als mächtigste Turing-äquivalente (mit Turing-Maschinen berechenbare), dann kontextsensitive, kontextfreie und reguläre Sprachen (Definitionen s. [10], Kap. 2.1). Von besonderer Bedeutung für die Anwendung auf natürliche Sprachen ist die *kontextfreie* Familie, da der überwiegende Teil der Strukturen natürlicher Sprachen mit kontextfreien Mitteln elegant und hinreichend effizient beschrieben werden kann. Dies hat einen offensichtlichen Grund: Die einfachste uniforme Möglichkeit zur Beschreibung von Konstituentenstrukturen besteht darin, zuzulassen, dass die Muster, welche die Aufeinanderfolge von Elementen beschreiben, selbst rekursiv sind — genau dadurch ist aber die Klasse der kontextfreien formalen

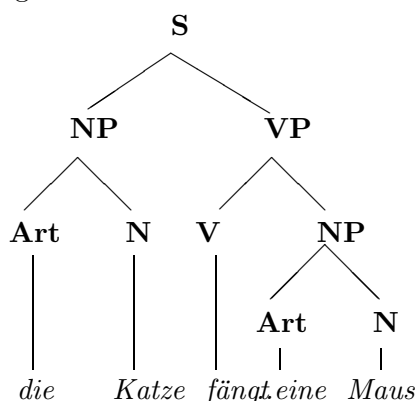
Sprachen konstituiert.

Jedes Verfahren, das zu einer Zeichenkette mindestens einen korrespondierenden Strukturbaum bestimmt, muss die Regeln in einer bestimmten Reihenfolge anwenden. Die Folge dieser Schritte wird *Ableitung* der Zeichenkette genannt. Beginnt man die Ableitung mit dem sog. Startsymbol der Grammatik — in unserem Fall S für „Satz“ — und wendet dann die entsprechenden Regeln für die davon abhängigen Kategoriensymbole an, bis man bei den „terminalen“ Symbolen, den Wörtern, angelangt ist, so bezeichnet man eine solche Ableitung als „*top-down*“. Beginnt man die Ableitung bei den terminalen Symbolen, also ganz „unten“ im Baum bei den Wörtern und wendet die Regeln von rechts nach links an, so spricht man von einer „*bottom-up*“-Ableitung. Orthogonal dazu unterscheidet man „*Tiefe-zuerst- (depth first)*“ und „*Breite-zuerst- (breadth first)*“ Ableitungen, je nachdem, ob die Expansion in Teilbäume in die Tiefe — jeweils ein Pfad im Baum über mehrere Niveaus — oder in die Breite — zuerst alle Knoten auf einem Niveau — erfolgt. Durch Kombination dieser Dimensionen sind auch andere Ordnungen der Ableitungsschritte möglich; im Strukturbaum selbst wird über die Reihenfolge abstrahiert.

□ *Beispiel:* In der folgenden einfachen kontextfreien Grammatik bedeuten die Zeichen S: Satz, NP: Nominalphrase, VP: Verbalphrase, V: Verb, Art: Artikel, N: Nomen. Die Lexikoneinträge seien der Einfachheit halber auch in Regelform notiert; der senkrechte Strich steht für „oder“.

S	→	NP VP	N	→	<i>Katze</i> <i>Maus</i>
VP	→	V NP	V	→	<i>fängt</i>
NP	→	Art N	Det	→	<i>die</i> <i>eine</i>

Für den Satz *Die Katze fängt eine Maus* wird der folgende Konstituentenstrukturbaum abgeleitet:



Mit kontextfreien Grammatiken tritt jedoch das Problem auf, dass für

jede Merkmalskombination² jeweils eigene Kategoriensymbole und mit diesen passende Regeln definiert werden müßten, was die Übersichtlichkeit und Handhabbarkeit von Grammatiken sehr schnell begrenzt.

- *Beispiel:* (Fortsetzung) Das Nomen „Katze“, hat das Genus-Merkmal Feminin, der Artikel „die“ (im vorliegenden Fall) das Genus Feminin, den Numerus Singular, den Kasus Nominativ, und steht in der dritten Person. Damit kann unterschieden werden, dass „die Katze“ eine zulässige NP ist, „das Katze“ aber nicht. Im Falle der Übereinstimmung („die Katze“) kann man diese Eigenschaften an die ganze NP vererben. Wir bräuchten dann aber statt eines NP-Symbols mehrere, in unserem Fall also ein Symbol für NP-feminin-singular, etc., wodurch auch die Anzahl der Regeln drastisch erhöht würde.

Eine elegante Lösung dieses Problems bieten die sog. *constraint-basierten* oder *Unifikations-Grammatikformalismen*, mit denen in nahezu allen modernen Sprachverarbeitungssystemen gearbeitet wird.

In **constraint-basierten Grammatikformalismen** treten an die Stelle atomarer Kategoriensymbole komplexe *Merkmalstrukturen*. Merkmalstrukturen sind endliche Mengen von Merkmalen, deren jedes aus einem Merkmalsnamen und einem Merkmalswert besteht, wobei letzterer entweder ein atomares Symbol oder selbst wieder eine Merkmalstruktur ist.

Hierbei kommt eine Schlüsselrolle für die Charakterisierung wohlgeformter Äußerungen einer Menge von *Restriktionen* („*Constraints*“) auf Merkmalen zu. Letztere werden oft als Systeme von Gleichungen, die Übereinstimmung von Merkmalen ausdrücken, formuliert und mit Hilfe der fundamentalen Operation der *Unifikation* gelöst³

Zur Anordnung verschiedener Lesarten entsprechend ihrer Plausibilität im Fall der syntaktischen Mehrdeutigkeit dienen *Probabilistische Kontextfreie Grammatiken*.

In **Probabilistischen Kontextfreien Grammatiken** ist jeder Regel wird ein Wahrscheinlichkeitswert zugeordnet, der aus dem Training mit einem Korpus von Sätzen gewonnen wurde.

Ein derartiges hybrides System eignet sich in Verbindung mit der bottom-up-Analysestrategie auch gut zur Erkennung grammatischer Fragmente in fehlerbehafteten Spracherkenner-Ausgaben.

Über die bloße Konstituentenstruktur hinaus ist aber noch festzustel-

²Im linguistischen Zusammenhang werden unter **Merkmalen** linguistisch relevante Eigenschaften von phonologischen, morphologischen, syntaktischen oder semantischen Einheiten verstanden.

³Für ausführliche Darstellungen sei z.B. auf [7, 2] verwiesen.

len, *wie* diese innerhalb eines Satzes funktioniert: Die *Dependenz* als zweite syntaktische Grundrelation geht über das Kriterium der bloßen Form hinaus. Sie bezeichnet Abhängigkeitsverhältnisse sprachlicher Elemente durch eine Vorkommensrelation, d.h. aus dem Vorkommen eines Elements kann auf das anderer Elemente geschlossen werden. Es geht dabei um den Zusammenhang zwischen *Konstituententypen funktionalen* (oder *relationalen*) Begriffen wie Subjekt oder Objekt. Bei der *Dependenz* kommt dem Hauptverb als grammatischem Prädikat eine zentrale Position zu: Es hat die Eigenschaft, bestimmte nominale Einheiten in seiner Umgebung zuzulassen bzw. zu verlangen und dabei Bedingungen für die Ausprägung ihrer Merkmale festzulegen („*Valenz*“).

9.2.6 Linguistische Verarbeitung: Semantik und Pragmatik

Semantische Interpretation. Eine Äußerung kann verschiedenartige Funktionen erfüllen; daher müssen folgende Bedeutungsaspekte unterschieden werden (s. [9]):

- Die *innersprachliche wörtliche*, Bedeutung, die sich vor allem auf die Wortbedeutung abstützt;
- die *referentielle* Bedeutung, die auf die Kommunikationssituation Bezug nimmt.

Die *kommunikative* Bedeutung, zu deren Darstellung Sprechakte und Interaktionsregeln herangezogen werden müssen, und die *soziale, emotive und rhetorisch-stilistische* Bedeutung, die aus dem soziokulturellen Umfeld zu erklären ist, werden der Pragmatik zugeordnet.

Die **semantische Analyse** einer Äußerung untersucht ihre im Sprachsystem angelegte Bedeutungsstruktur, insbesondere die Darstellung der innertextlichen *funktionalen* und *logischen Relationen* der Äußerungselemente zueinander — unabhängig von situativen und kommunikativen Faktoren und von ihrem Bezug auf Wissen über die Welt.

Ihre Ziele sind

- der Aufbau formaler Ausdrücke zur Darstellung von Bedeutungsstrukturen (*Semantikkonstruktion*) unter Berücksichtigung verschiedener Lesarten,

- die *semantische Resolution* zur Bestimmung des aktuellen semantischen Werts (Auflösung lexikalischer Mehrdeutigkeiten, die Aufdeckung semantischer Anomalien) und
- die *semantische Auswertung* als Bestimmung der „relevanten Äußerungsinformation“ mit Hilfe von Schlußfolgerungsmechanismen.

Somit fällt ihr sowohl eine *Bewertungs-* und *Auswahl-* als auch eine *Repräsentationsfunktion* zu (vgl. [9]).

Die *lexikalische Semantik* befaßt sich mit der Darstellung der Bedeutung der Wörter der Sprache und den Beziehungen zwischen diesen sowie mit der Frage, auf welche Objekt- und Ereignistypen sich Wörter beziehen. Die einfachste Form, solche Beziehungen darzustellen, bietet die Organisation eines Lexikons als sog. *Thesaurus*, wobei als Grundbeziehung die Synonymie herangezogen wird. Synonyme Wörter werden in Gruppen zusammengefasst. Mit der Berücksichtigung weiterer Beziehungen wie Antonymie und Hyponymie (engere vs. weitere Bezeichnung) entstehen sog. semantische Wortnetze wie WordNet [3].

Hinzu kommen semantische Einträge im Lexikon, die darauf beruhen, dass die Bedeutung eines Wortes die Bedeutungen anderer Wörter einschränkt, die mit ihm „zusammengehen“ können. Solche *Selektionsrestriktionen* werden oft als Muster dargestellt, mit denen die Kompatibilität semantischer Attribute (z.B. „physisches Objekt“) in bestimmten syntaktischen Strukturen überprüft werden kann. Eine zentrale Rolle spielen hier die semantischen Lexikoneinträge der Verben (s.o. *Dependenz*), die üblicherweise aus einer „Logischen Form“, gegeben als Prädikat-Argumentstruktur, bestehen. Für die Argumente sind *thematische Rollen* wie AGENS, INSTRUMENT, DATIV, etc. angegeben, in die dann die Repräsentationen für Subjekt, Objekt, etc. einzusetzen sind.

- *Beispiel:* (Fortsetzung) Dem Handlungsverb „fangen“ wird ein dreistelliges Prädikat zugeordnet, dessen erste Argumentstelle die AGENS-Rolle trägt, die den belebten Ausführenden einer Handlung bezeichnet, hier das Subjekt, „die Katze“. Die DATIV-Rolle als zweites Argument bezeichnet einen belebten Beteiligten, der durch die Handlung betroffen ist: das Objekt, „die Maus“. Das dritte Argument, durch die INSTRUMENT-Rolle beschrieben, bliebe leer, denn es wird nichts über ein Hilfsmittel zur Ausführung der Handlung ausgesagt.

Den Aufbau logischer Formen für ganze Äußerungen als Darstellungen ihrer wörtlichen Bedeutung leistet die *kompositionale Semantik*.

Die **kompositionale Semantik** bestimmt, wie sich die Bedeutungen komplexer sprachlicher Objekte aus denen ihrer Komponenten zusammensetzen.

Im einfachsten Fall besteht die Schnittstelle zwischen syntaktischer und semantischer Analyse darin, dass mit jeder Grammatikregel eine Kompositionsregel zum Aufbau der logischen Form der Mutterkonstituente aus den logischen Formen ihrer Töchter assoziiert wird. Eine besondere Rolle spielen bei der kompositionalen Semantik Verfahren zur Referenzauflösung (v.a. Bestimmung von Referenzobjekten für Pronomina), Mechanismen der Negation und Quantifikation und die Auflösung des Bereichs dieser Operatoren. Im Unterschied zur Logik verfügt die Umgangssprache neben „für alle“ und „für einige“ über eine Vielzahl weiterer Quantoren, teilweise in Verbindung mit Vagheit, z.B. „die meisten“, „manche“, etc., die sämtlich grammatisch als Artikel erscheinen — und nicht als Operatoren über Aussagen.

Wenn wir unser obiges Beispiel mit einem weiteren Satz fortsetzen, „Sie frisst sie“, so ergibt sich die Aufgabe, für beide Personalpronomina die jeweiligen Referenzobjekte zu identifizieren: das erste „sie“ bezieht sich auf die Katze, das zweite auf die Maus — und eben nicht umgekehrt. Dass normalerweise das Opfer gefressen wird und nicht der Täter, kann nur durch Rückgriff auf formalisiertes Weltwissen erschlossen werden.

Eine Semantikkonstruktion auf der Ebene einzelner Sätze reicht jedoch nicht aus: Es wird eine semantische Repräsentation gebraucht, die ganze Diskurse umfasst und die mit jedem Redebeitrag schrittweise erweitert wird. Wenn wir die Referenzauflösung in unserem Beispiel „Sie frisst sie“ angehen wollen, benötigen wir einen Zugriff auf die logische Form des vorhergehenden Satzes, um feststellen zu können, von welchen Referenzobjekten, den sog. Diskursreferenten, überhaupt die Rede ist. Ein in der Sprachverarbeitung vielfach genutzter Ansatz geht auf die *Diskursrepräsentationstheorie* nach Kamp u.a. (s. [2, 9]) zurück. Die durch sie definierten Diskursrepräsentationsstrukturen (DRSen) besitzen zwei Komponenten: die mit einer Äußerung eingeführten Diskursreferenten sowie die logischen Formen. Durch Verschachtelung solcher Strukturen können Repräsentationen für ganze Diskurse aufgebaut werden.

Pragmatik und Diskurs. Damit ein Hörer eine Äußerung über das rein wörtliche Verstehen hinaus auch die zugrundeliegende Absicht des Sprechers erkennen kann, muss er sie in Beziehung zum bisherigen Dialogverlauf und seinem eigenen Hintergrundwissen setzen. Hierzu ist eine Klassifikation der Sprachhandlungen (*Sprechakte*) erforderlich. Äußerungen sind neben den Sprecherabsichten von weiteren pragmatischen Fak-

toren wie Präsuppositionen (implizite Voraussetzungen) abhängig. Eine Präsupposition für die Aussage „Die Katze steht auf“ wäre z.B., dass sie zuvor gelegen hat.

Pragmatische Analyse: Ihre wesentliche Aufgabe der pragmatischen Analyse besteht in der Erkennung und Analyse der *Intentionen* oder Redeabsichten des Sprechers.

Hierfür und um die daraus abgeleiteten Ziele erfüllen bzw. ihre Erfüllung überprüfen zu können, müssen Schlußfolgerungen über Sprechakte ausgeführt werden. Diese Interpretationsprozesse finden normalerweise in dynamischen Kontexten statt, zu denen die Dialogsituation, der Diskursbereich, sowie Präferenzen des Sprechers gehören. Dies heißt, dass wir mit einer komplexen *Planungsaufgabe* konfrontiert sind, zu der sprachliche Handlungen, d.h. Sprechakte, und Aktionen in dem jeweiligen Diskursbereich beitragen. Hierzu sind allgemeine Verfahren der Schlußfolgerung und Handlungsplanung anzuwenden, die jenseits des Bereichs der Sprachverarbeitung i.e.S. liegen.

9.3 Bilderkennung

9.3.1 Segmentierung

Bilder beinhalten üblicherweise nicht nur ein einzelnes Objekt, sondern mehrere Objekte, die nicht notwendigerweise alle bekannt sind. Zur Erkennung einzelner Objekte in Mehrobjektszenen ist es notwendig, Merkmale für die einzelnen Objekte zu berechnen und auf der Basis dieser Merkmale die Objekte zu separieren. In aller Regel handelt es sich bei den Merkmalen um geometrische Primitiva, die in einem Segmentierungsschritt berechnet werden.

Segmentierung. Bei der Segmentierung eines Bildes wird die ursprüngliche Matrix von Intensitätswerten in geometrische Primitiva transformiert, welche die Klassifikation und Lokalisierung von Objekten ermöglichen und vereinfachen sollen.

□ *Beispiel:*

Typische Beispiele für Segmentierungsergebnisse sind markante Punkte wie beispielsweise Eckpunkte polyedrischer Objekte (siehe Bild 9.3) oder Konturen von Tumorzellen wie in Abbildung 9.6 dargestellt.

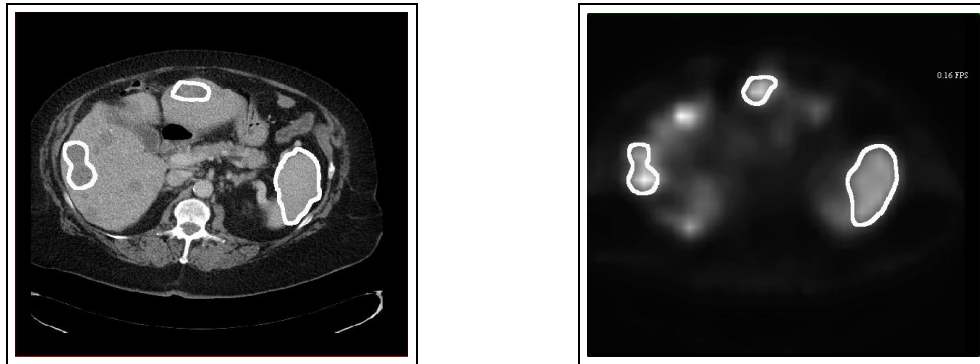


Bild 9.6 Beispiele für segmentierte Tumorzellen in einer Computertomographiaufnahme (links) und einem Positronen-Emissionstomogramm (rechts)

9.3.2 Projektionsmodelle

Die Verarbeitung und Analyse von Bildern muss sich mit dem Problem auseinander setzen, dass die verfügbaren Bilder Projektionen der realen 3D-Welt sind und dass die Blickrichtung auf das Objekt sowie die photometrischen Rahmenbedingungen variieren. Zur Lösung vieler Problemstellungen der Bildverarbeitung und des Rechnersehens ist die Kenntnis oder zumindest die Berechnung der Projektionsparameter zwingend erforderlich. Die Projektionsparameter hängen vom gewählten Projektionsmodell ab. Im Rechnersehen werden die folgenden Projektionsmodelle eingesetzt: *Parallelprojektion*, *schwach perspektivische Projektion*, *paraprospektivische Projektion*, *perspektivische Projektion* [4]. Bild 9.7 veranschaulicht die unterschiedlichen Projektionsmodelle. Homogene Koordinaten erlauben, die Projektionen als lineare Abbildungen zu beschreiben.

□ *Beispiel:*

Bei der *Parallelprojektion* wird ein 3D-Vektor durch Weglassen der z -Koordinate auf einen 2D-Bildpunkt abgebildet.

- *Beispiel:* Die *schwach perspektivische Projektion* bildet alle Punkte orthogonal auf eine Hilfsebene ab, die durch den Schwerpunkt aller zu projizierenden Punkte geht und parallel zur Bildebene ist. Die Punkte der Hilfsebene werden danach perspektivisch in die Bildebene projiziert. Da alle Punkte denselben Abstand zur Bildebene haben, ist diese Abbildung nach wie vor linear.

□ *Beispiel:*

Die perspektivische Projektion erfordert die Division der x - und y -Koordinate durch die z -Koordinate. Diese nichtlineare Abbildung führt

dazu, dass für den Fall der perspektivischen Projektion viele Problemstellungen des Rechnersehens äußerst schwierig zu lösen sind.

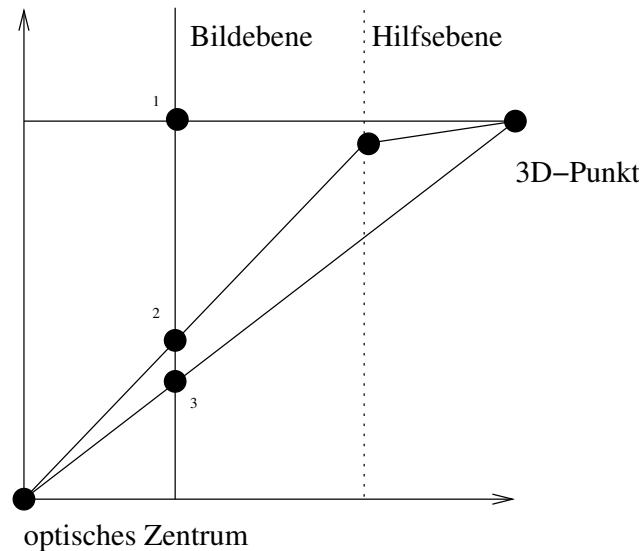


Bild 9.7 Parallelprojektion (1), schwach perspektivische Projektion (2) und perspektivische Projektion (3)

Neben der Projektionsabbildung gibt es noch weitere Freiheitsgrade, die für die Bildanalyse von Bedeutung sind: Die Kamera kann im Dreidimensionalen rotiert und verschoben werden. Charakterisiert wird diese Abbildung durch sechs Parameter: drei Rotationswinkel und einen 3D-Verschiebungsvektor. Diese Parameter werden als *extrinsische Kameraparameter* bezeichnet und ändern sich mit der Kameraposition. Unverändert bei Rotation und Translation bleiben die *intrinsischen Kameraparameter*. Die Abbildung der projizierten Punkte in das Koordinatensystem des verwendeten CCD-Chips werden über die intrinsischen Kameraparameter definiert. Diese haben im allgemeinen fünf Freiheitsgrade. Beispiele sind der Winkel, den die Koordinatenachsen einschließen oder das Längenverhältnis der Bildpunkte im Falle nicht-quadratischer Pixel.

Kamerakalibrierung. Unter Kamerakalibrierung wird die Schätzung der ex- und intrinsischen Kameraparameter verstanden. Durch die Verwendung homogener Koordinaten können viele Kalibrierprobleme auf der Basis linearer Schätzverfahren gelöst werden. Steht für die Kalibrierung kein präzise vermessenes Kalibriermuster zur Verfügung, dann handelt es sich um ein Problem der Selbstkalibrierung.

9.3.3 Modellbasierte Objekterkennung

Objekterkennung. Die Objekterkennung zerfällt in die Teilschritte Objektidentifikation und die Objektlokalisierung.

Die Lokalisierung ist für die Erkennung nicht notwendig, wenn Merkmale eingesetzt werden, die *lageinvariant* sind. Die Grundidee der modellbasierten Objektidentifikation und -lokalisierung ist der Vergleich der beobachteten Bilder mit Modellen von Objekten, die in einer Datenbasis abgelegt sind. Die Erkennung und Lokalisierung basiert auf der Optimierung einer Zielfunktion, welche die Ähnlichkeit von Modell- und Bilddaten und bewertet. Hierzu ist notwendig, dass für alle zu erkennenden Objekte *Modelle* vorliegen. Die Generierung der Objektmodelle kann manuell, unter Verwendung von CAD-Daten oder mittels Trainingsaufnahmen (Stichprobe) erfolgen. Die Merkmale, die Objekte charakterisieren, sind meist geometrischen Ursprungs. Zur Modellierung werden beispielsweise Konturlinien, Kanten, Flächen oder Eckpunkte verwendet. Der Vergleich der Modelle mit der Bildinformation erfolgt unter Verwendung dieser geometrischen Merkmale. Alternativ zur geometrischen Modellierung werden auch ansichtenbasierte Ansätze verfolgt. Dabei werden die Bilder mathematisch als hochdimensionale Vektoren aufgefasst. Die Erkennung erfolgt dann unter Verwendung von Transformationen dieser Vektoren. Einen Überblick über Modellierungsansätze für die Objekterkennung gibt [13].

Objektidentifikation. Bei der *Objektidentifikation* wird die Frage beantwortet, welche Objekte der Modelldatenbasis in einem gegebenen Bild zu sehen sind.

Die Identifikation von Objekten ist am einfachsten in Aufnahmen, die ein Objekt vor homogenem Hintergrund zeigen. Ungleich schwieriger wird das Problem der Objektidentifikation, wenn Mehrobjektszenen vorliegen, wenn die Objekte teilweise verdeckt sind oder wenn ein oder mehrere zu identifizierende Objekte vor heterogenen Hintergrund aufgenommen werden.

Die Objektidentifikation erfolgt häufig unter Verwendung hierarchischer Klassifikatoren. In der ersten Phase werden *Invarianten* als Merkmale verwendet. Bei Invarianten handelt es sich um Merkmale, die sich bei einer bestimmten Art von Transformationen nicht ändern. Ein Merkmal, das im 2D translations- und rotationsinvariant ist die Länge einer Strecke. Ein weiteres Merkmal, das sich bei Rotation, Translation und perspektivischer Projektion nicht ändert, ist das *Doppelverhältnis* (engl. *cross ratio*). Invarianten erlauben die schnelle Suche geeigneter Model-

le für die Objekterkennung. Werden keine invarianten Merkmale für die Objekterkennung verwendet, muss für jedes Objektmodell die rechenintensive Schätzung der Lageparameter durchgeführt werden.

Objektlokalisierung. Bei der *Objektlokalisierung* werden aus den Bilddaten die Lageparameter der identifizierten Objekte relativ zu einem Referenzkoordinatensystem berechnet.

Die Objektlokalisierung erfordert die Anpassung eines Modells an die Beobachtung. Die Abbildung des Objektmodells in die Bildebene ist in den Lageparametern parametrisiert. Die Schätzung der Lageparameter erfolgt durch Optimierung einer Zielfunktion wie beispielsweise (9.5).

□ *Beispiel:*

Liegt ein Objektmodell in Form von 3D-Punkten vor und werden 2D-Punkte im Bild beobachtet, dann wird die Abbildung vom Dreidimensionalen in die Bildebene durch eine parametrische Abbildung beschrieben. Wichtig für die Schätzung der Objektrotation und -translation ist die Berechnung der Punktkorrespondenzen. Es muss bestimmt werden, welche 2D-Punkte der Bildebene zum Objekt und welche zum Hintergrund gehören. Die dem Objekt zugeordneten Punkte müssen dann den korrespondierenden 3D-Modellpunkten zugeordnet werden. Erst danach kann die Berechnung der Rotations- und Translationsparameter erfolgen.

9.3.4 3D-Rekonstruktion aus Bildern

Ein Teilgebiet des Rechnersehens beschäftigt sich mit Problem der 3D-Rekonstruktion aus Bildern, die aus unterschiedlichen Blickrichtungen aufgenommen wurden. Eingabe kann hier beispielsweise ein Stereobildpaar oder eine Bildfolge sein. Die Algorithmen zur 3D-Rekonstruktion wurden in den letzten 15 Jahren entscheidend verbessert und weiterentwickelt. Die Rekonstruktionverfahren sind wesentlich vom verwendeten Projektionsmodell abhängig.

Die Rekonstruktion aus Bildpaaren bei perspektivischer Projektion basiert im Wesentlichen auf Triangulationsansätzen und der Verwendung der Epipolargeometrie oder des Bündelausgleichs [4].

Der entscheidende Durchbruch bei der Lösung der Rekonstruktionsaufgabe aus einer Bildfolge bei vorhandener Parallelprojektion gelang 1992. Tomasi [4] entwickelte einen Algorithmus, der sowohl die 3D-Struktur als auch die 3D-Bewegung berechnet. Bis auf eine Mehrdeutigkeit ist die Berechnung linear. Dieser Ansatz wurde in den Folgejahren auf die an-

deren Projektionsmodelle erweitert [4] und ist heute die Grundlage für die 3D-Rekonstruktion aus Bildfolgen.

9.3.5 Szenenanalyse

In vielen Anwendungen sollen nicht nur einzelne Objekte erkannt werden, sondern eine gesamte Szene analysiert werden. Wie in der Spracherkennung ist bei der Szenenanalyse der Kontext zu berücksichtigen, um ein konsistentes Analyseergebnis zu berechnen. Ein Schreibtisch wird nicht auf dem Telefon stehen, sondern das Telefon auf dem Schreibtisch. Diese Zusammenhänge sind bei der Szenenmodellierung zu berücksichtigen und fließen in die Modellierung ein.

Literaturverzeichnis

- [1] Bussmann, H. (Hrsgb.): *Lexikon der Sprachwissenschaft*, Kröner, Stuttgart, 3. Ausg., 2002.
- [2] Carstensen, K.-U.; u. a., (Hrsgb.): *Computerlinguistik und Sprachtechnologie*, Elsevier, München, 2. Ausg., 2004.
- [3] Fellbaum, C. (Hrsgb.): *WordNet. An Electronic Lexical Database*, The Mit Press, Cambridge, MA, 1998.
- [4] Hartley, R. I.; Zisserman, A.: *Multiple View Geometry in Computer Vision*, Cambridge University Press, Cambridge, 2. Ausg., 2004.
- [5] Heid, U.: *Morphologie und Lexikon*, in Görz, G.; u. a., (Hrsgb.): *Handbuch der Künstlichen Intelligenz*, Oldenbourg, München, 4. Ausg., 2003, S. 665–709.
- [6] Henning, P. (Hrsgb.): *Taschenbuch Multimedia*, Fachbuchverlag Leipzig im Carl Hanser Verlag, München and Wien, 3. Ausg., 2003.
- [7] Langer, H.; Naumann, S.: *Parsing natürlicher Sprache*, in Görz, G.; u. a., (Hrsgb.): *Handbuch der Künstlichen Intelligenz*, Oldenbourg, München, 4. Ausg., 2003, S. 711–738.
- [8] Niemann, H.: *Klassifikation von Mustern*, Springer, Berlin, 1983.
- [9] Pinkal, M.: *Semantikformalismen für die Sprachverarbeitung*, in Görz, G.; u. a., (Hrsgb.): *Handbuch der Künstlichen Intelligenz*, Oldenbourg, München, 4. Ausg., 2003, S. 739–782.
- [10] Schneider, U.; Werner, D. (Hrsgb.): *Taschenbuch der Informatik*, Fachbuchverlag Leipzig im Carl Hanser Verlag, München and Wien, 2004.
- [11] Schukat-Talamazzini, E. G.: *Automatische Spracherkennung: Grundlagen, statistische Modelle und effiziente Algorithmen*, Reihe Künstliche Intelligenz, Vieweg, Braunschweig, 1995.

- [12] Steinmetz, R.; Nahrstedt, K.: *Multimedia Applications*, Springer, Berlin, 2004.
- [13] Trucco, E.; Verri, A.: *Introductory Techniques for 3-D Computer Vision*, Prentice Hall, Upper Saddle River, NJ, 1998.