Technische Grenzen von ubiquitären Anwendungen im Bereich von Computer Vision

Author: Reto Strobl
Assistent: Harald Vogt
Professor: Friedemann Mattern

June 20, 2000

Abstract

Bei der Entwicklung von ubiquitären Anwendungen stösst man auf Probleme verschiedenster Schwierigkeit. Grob gesagt lassen sich diese in vier Klassen aufteilen: Absprache Probleme, technologische Probleme, ethische Probleme und algorithmische Probleme. Bei den Absprache Problemen geht es hauptsächlich darum, sich auf einen Standart zu einigen. Kunststoffbildschirme sind ein Beispiel für ein technologisches Problem. Ethische Fragen beschäftigen sich mit z.B. "big brother Fragestellungen". Bei algorithmischen Problemen kennt man schlicht und einfach keine Algorithmen, welche eine echt befriedigende Lösung finden.

In dieser Seminararbeit möchte ich auf algorithmische Probleme im Bereich Computer Vision und KI aufmerksam machen.

1 Einführung

Computer Vision Techniken werden in bei sehr vielen Ideen von ubiquitären Anwendungen vorausgesetzt. Beispiele sind Smart rooms[1], Aware Home[2], Autonomous driving vehicles, intelligente PDAs[3], HCI ("augmented desk" Projekt[16]). Computer Vision Systeme, die solch hohe Anforderungen erfüllen, sind (noch) nicht eine Selbstverständlichkeit. In den folgenden Abschnitten soll ein Bild der Problematik und des aktuellen Forschungsstandes vermittelt werden.

[1] diese können z.B. die Anwesenden Personen per Kamera erkennen

[2] ein Haus, welches alltägliche Tätigkeiten der Bewohner erkennt und diese darin Intelligent unterstützt

[3] aufgrund bestimmter Umgebungseigenschaften machen sie spontane Vorschläge, wie z.B. "nimm besser Tram 9, das geht schneller als Tram 10 dorthin wo du gehen willst"

2 Überblick der Problematik: Objekt Erkennung und Klassifizierung von Bewegungen

Zusammenhängende Flächen in einem Bild zu finden ist ein häufig anzutreffendes algorithmisches Problem ("Segmentation"). Das Ziel[4] der Segmentation ist oft die *Objekt Erkennung* ("Object recognition"). Weiter ist das *Erkennen von Bewegungen* (Gehen, Rennen, Hüpfen) von Interesse. Die dazu verwendeten Algorithmen arbeiten in verschiedenen Modellen.

Wichtige Charakteristika der Algorithmen sind "bottom up" bzw. "top down". Ein bottom up Algorithmus schliesst ausschliesslich aufgrund lokaler Information (z.B. Farbe der Pixel) auf Segmente. Ein top down Algorithmus hingegen zieht auch den Kontext in Betracht: Ein teilweise verdecktes Segment (im bottom up Fall als mehrere verschiedene Segmente erkannt) wird so als zusammenhängendes identifiziert.

Objekt Erkennung in expliziten Modellen: *Explizite Modelle* modellieren die Welt in einer klassischen 3D Struktur und versuchen dann, geometrische Formen wiederzuerkennen. Ein grosser Nachteil solcher Methoden ist die Anfälligkeit auf "noise". Beispiele dafür sind [12, 10]

Klassifizieren von Bewegungen in impliziten Modellen: Ein anderer Ansatz ist die *implizite Representation*. In einer Sequenz von Bildern werden Segmente gefunden. Ohne auf Objekte zu schliessen wird durch die Verschiebung der Segmente über die Zeit auf z.B. Schritt Zyklen einer gehenden Person geschlossen. Wärend für diesen Ansatz noise kein Problem zu sein scheint, so ist er beschränkt auf Szenen mit gleichmässigem Hintergrund, und/oder mit nur einer Person. Die einfachste Representation stammt von Freeman und Roth [13]. Dabei wird durch Veränderungen in einer Sequenz von Bildern auf Bewegungen von Menschen geschlossen.

"High-level human motion models" ist ein Versuch, Bewegungsabläufe zu erkennen, ohne die dazugehörigen Objekte genauer zu identifizieren. Es wird auf klassische Computer Vision Techniken verzichtet. In [5] wird ein System erklärt, welches auf menschliche Gangarten schliesst, lediglich durch Betrachtung von bewegenden Lichtquellen in einer Sequenz von Bildern.

Eine gute Übersicht über diese verschiedenen Ansätze sowie weiterführende Literaturreferenzen sind in [1] zu finden. Im folgenden werden einige Beispielprobleme und Lösungsansätze präsentiert, um die erwähnte Problematik zu verdeutlichen.

Der oft beschränkende Faktor liegt in der Allgemeinheit des Problems. Ein System kann nur die ihm bekannten Objekte identifizieren. Jede erlaubte Variation eines Objektes[5] muss als eigenes Objekt erkannt werden. Im Abschnitt "finding naked people" wird eine Problemstellung beschrieben, die (noch) zu allgemein ist, als dass man sie heutzutage zuverlässig lösen kann. In "Fortgeschrittene Objekterkennung" wird ein Modell eingeführt, nach welchem state

[4] überspitzt gesagt dient Segmentierung als Informations-komprimierung bzw. Aufbereitung

[5] Menschen sitzen, gehen, liegen, ...

of the art systeme arbeiten. Am Beispiel von autonom fahrenden Fahrzeugen wird gezeigt, was Computer Vision in einem beschränkten Umfeld alles kann. Im Abschnitt "Erkennen von Bewegungen" wird darauf eingegangen, wie man eine Sequenz von Bildern interpretieren kann.

3 Finding naked people

In [3] wird folgende Problemstellung untersucht, anhand deren ich auf typische Probleme von "Object recognition" Algorithmen aufmerksam machen möchte: Gegeben eine Menge von Bildern, finde diese Bilder, welche nackte Menschen zeigen.

Resultat: 57% Präzision und 43% Ausbeute erreichte das System auf einer Menge von 4954 Bildern; darunter 565 relevante.

Auf den ersten Blick erschien mir dies eher ernüchternd. Was aber wie eine einfache Fragestellung aussieht, birgt viele Schwierigkeiten, da man fast nichts a priori weiss über die zu untersuchenden Bilder. Ausserdem sind die Objekte sehr flexibel. Der Hintergrund kann beliebig sein. Mehere Körper können auftreten; womöglich verdecken sie sich gegenseitig teilweise. Glieder können sich in verschiedensten Positionen befinden. Die Aufnahmen stammen oft aus verschiedenen Kamerwinkeln und unter stark varierenden Lichtverhältnissen.

Der in [3] vorgeschlagene Algorithmus funktioniert folgendermassen:

- **Skin Filter:** Grössere Flächen werden lokalisiert, deren Farbe und Textur deren von Haut gleichen.
- **Grouper:** In diesen Flächen wird versucht, Glieder zu finden. Diese wiederum werden zu verbundenen Gliedern zusammengefügt. Je nach gefundenen Strukturen ist dann ein Bild "positiv".

Der Skin Filter untersucht die Farbwerte der Pixel, wendet Filter an und schliesst dann auf Haut oder nicht Haut. Im Grouper kommen Geometrische Algorithmen zur Anwendung. Keine Neuronalen Netze oder andere KI Methoden werden eingesetzt.

Nach genauerem Untersuchen des Problems und Vergleichen mit anderen "Information retrieval" Systemen scheint mir das Resultat recht befreidigend zu sein, ja sogar überraschend gut. Der gewählte algorithmische Ansatz erscheint mir nämlich zu stark "bottom up" und unnatürlich. Betrachten Menschen ein Bild, sehen sie eine Szene mit Objekten. Man untersucht nicht mit einer Lupe kleine Bildflecken um dann daraus auf das Grosse zu schliessen. Genau das umgekehrte ist der Fall.

4 Fortgeschrittene Objekterkennung

Idee: in einem beliebigen Videofilm sollen Menschen erkannt werden[6]. In einem weiteren Schritt soll ihre Bewegung klassifiziert werden (dazu siehe Abschnitt

[6] nur das Objekt "Mensch" soll erkannt werden, nicht um welchen Menschen es sich handelt

"Erkennen von Bewegungen"). In [1, 2] beschreibt C. Bregler ein System, welches Segmentiert, Menschen identifiziert und ihre Bewegungen interpretiert. Das System ist in der beschriebenen Implementation darauf beschränkt, einen einzelnen Menschen und seine Bewegung zu erkennen.

Breglers Ansatz besteht aus einem Schichten Modell. Auf jeder Schicht werden Hypothesen aufgestellt. Schliesslich wird auf diese Hypothesen entschlossen, welche als ganzes die höchste Wahrscheinlichkeit aufweisen.

Auf niedriger Stufe geht es um Segmentierung. Es wird auf die Zusammengehörigkeit von Pixeln nicht nur aufgrund ihrer Farbe und Textur geschlossen, sondern auch aufgrund ihrer Bewegung (bottom up) und erkannten Formen der mittleren Stufe (top down feedback). Auf der mittleren Stufe werden in den Segmenten Glieder erkannt, sowie einfache Bewegungszyklen. Auf einer hohen Stufe werden dann komplexe Bewegungen erkannt, wie z.B. Gesten, Tanzstile, usw. . "Soft commitement" ist ein grundlegendes Prinzip: Können Segmente von der niedrigsten Stufe nicht eindeutig erkannt werden, propagieren die Optionen als Hypothesen einfach zur mittleren Stufe weiter. Aus den der mittleren Stufe bekannten Gliedformen können Hypotesen eliminiert werden. Auf hoher Stufe sind Bewegungsformen bekannt, die wiederum gewisse Hypothesen ausschliessen.

Beim Erkennen von Gliedern bzw. zyklischen Bewegungen werden probabilistische Methoden eingesetzt (EM[7], Markov Schätzungen). Es gibt auch Methoden, die auf Neuronalen Netzen basieren. Solche sind in der Regel effizienter, liefern aber einerseits keine klaren Wahrscheinlichkeiten und sind andererseits schwierig zu verstehen bzw. nachzuvollziehen.

Dieser Ansatz erscheint mir vielversprechend vor allem aufgrund der Schichtenstruktur, die einen top down Ansatz ermöglicht. Das Projekt scheint mir ein wichtiger Teil der Grundlagenforschung im Gebiet der Objekt Erkennung zu sein.

5 State of the art Segmentation

J. Shi und J. Malik haben erst kürzlich einen neuen vielversprechenden Versuch zum lösen des Segmentation und Gruppierungs Problem vorgeschlagen. Dabei werden ähnlich wie bei Breglers System die Gruppierung nicht nur durch low level Kriterien (bottom up) erkannt, sondern auch durch abstraktere high level Kriterien (top down), die wiederum eine Neugruppierung auf tiefem Niveau zur Folge haben. Ihr Verfahren basiert auf dem "normalized Cut" und ist in [6] beschrieben.

In [14] wird ein Verfahren aufbauend auf normalized Cuts beschrieben und eine Fülle von experimentellen Resultaten präsentiert.

[7] Estimation Maximation Algorithmus

6 Automatische Fahrzeug Steuerung durch visuelle Systeme

Idee: Kamera Systeme beschaffen dem Computer die nötigen Informationen über Verkehr, Strassengeometrie, Position und Ausrichtung des Autos, sodass dieser in der Lage ist, die richtigen Steuerbefehle zu betätigen. Eine weitere Schwierigkeit ist der übrige Strassenverkehr. Er soll erkannt werden und es soll eine sichere Distanz eingehalten werden. Überholmanöver, wie auch Abbiegen an Kreuzungen soll sicher funktionieren. Im weiteren wird näher auf die Problematik der Strassengeometrie sowie das Erkennen und "im Auge behalten" (tracken) von Hindernissen eingegangen.

Kennt man die relative Position und Ausrichtung der Strasse in einem bestimmten Abstand vor dem Auto (z.B. 50m) ist es möglich, aus der momentanen Lage des Autos relativ zur Strasse präzise Steuerbefehle zu berechnen. Das Problem ist die Erfassung der Strassengeometrie. Dies soll auf beliebigen Strassen funktionieren; sei das eine Autobahn, eine Landstrasse (innerorts, ausserorts), oder sogar eine Schotterstrasse.

Im Rahmen des noch aktiven PATH Projektes in Californien wurde diese Problemstellung in Berkeley untersucht. Allerdings scheint dieses Teilprojekt unterdessen eingestellt zu sein.

6.1 Erkennen der Strassengeometrie

In [8] werden die Strassenmarkierungen zu Hilfe genommen. Man nimmt weiter an, dass sich die Markierung an einer gewissen Stelle befinden muss (im dynamischen Fall wird dieser Ort aus dem Momentzustand berechnet). Eine einfache Template Methode dient dann zur Erkennung der Markierung. Diese Methode hat klar den Vorteil, dass sie effizient ist, was ein kritischer Punkt für diese Art Anwendung ist. Allerdings ist sie beschränkt auf Strassen mit Markierungen.

ALVINN [9] ist ein Projekt, das die Strassengeometrie viel allgemeiner zu erkennen versucht. Es basiert auf Neuronalen Netzen. Im unterschied zu [8] liefert das Netz nicht die Strassengeometrie einem Controller, sondern direkt die Steuersignale als Output. Die Neuronalen Netze werden "on the fly" vom Autofahrer trainiert. Es wurde auf asphaltierten, einspurigen und zweispurigen Strassen, sowei Autobahnen und Schotterstrassen getestet (Resultat: 150 km erfolgreiches autonomes Fahren"). ALVINN ist nur ein "line keeping" System. Um Überholmanöver durchzuführen oder andere Autos zu erkennen muss ein weiteres Modul eingefügt werden. Ein Vorschlag ist in [15] zu finden.

6.2 Hindernisse erkennen und verfolgen

Der in [8] präsentierte Ansatz, um Hindernisse (andere Autos) zu erkennen basiert auf einem 2D Feature-Match Algorithmus, der spezielle Eigenschaften der anderen Autos kennt, sucht und findet. Um die erkannten Autos nicht aus den Augen zu verlieren, werden markante Punkte verfolgt ("corner tracking"),

sowie parallel dazu die Bounding Box. Ein Radarsystem sorgt ausserdem für Präzision in der Abstandschätzung.

In [7] wird "binocular stereopsis" verwendet, um Hindernisse zu erkennen. Der Algorithmus berechnet die "Stereo Disparität". Dies ist effizient möglich und ergibt Kandidaten für Hindernisse. Diese werden über die Zeit mit einem Kalman Filter verfolgt.

6.3 Resultate

Das PATH Projekt [8] wurde 1997 erfolgreich in San Diego demonstriert. Dabei fuhr das System autonom auf einem Stück Freeway (demo des longitudinalen Kontrollers) mit 60 mph und führte auch Überholmanöver aus[8]. Der laterale Kontroller wurde durch Parkiermanöver demonstriert. Dennoch scheint das Projekt eingestellt zu sein. Ich konnte keine weiteren Informationen dazu finden.

Ähnliche Resultate erzielten die unabhängige Gruppe an der Ohio State Universität. Doch auch dort ist das Projekt seit 1997 eingestellt.

Das ALVINN Projekt der Carnegie Mellon Universität ist trotz vielversprechender Ergebnisse ebenfalls seit 1996 eingestellt worden. Auf Anfrage beim Projektleiter gab man mir Geldprobleme als Grund an.

Bemerkenswert sind die Ergebnisse von Ernst D. Dickmanns. Bereits 1987 entwickelte er einen Prototypen, welcher auf abgesperrten Strassen längere Strecken autonom fahren konnte. 1994 gelang dies sogar auf Strasse mit öffentlichem Verkehr. Das System ist in [4] beschrieben. Es basiert hauptsächlich auf einem 4-Dim Modell des Zustandes (Raum und Zeit), das durch Computer Vision aufgebaut wird.

7 Erkennen von Bewegungen

In Breglers System (siehe "Fortgeschrittene Objekterkennung") wird die HMM (Hidden Markov Model) Technik angewendet, die vielerorts als sehr geeignet aufgefasst wird. Dabei modelliert man Bewegungen (Gehen, Rennen, Hüpfen, ...) mit Hilfe von Markovketten. Jeder Zustand der Markovkette ordnet jeder möglichen Stellung des betrachteten Objektes eine Wahrscheinlichkeit zu. So lässt sich einfach ausrechnen, wie wahrscheinlich eine Sequenz von Bildern einer "Hüpfen-Markovkette" entspricht, bzw einer "Gehen-Markovkette" usw. . Ab einer gewissen Wahrscheinlichkeit wird dann auf eine Bewegung entschieden. Durch eine iterative nichtlineare Trainingsstrategie können die Werte der Übergangsmatrix der verschiednen dynamischen Systeme – Gehen, Rennen, Hüpfen sind Beispiele von dynamischen Systemen – erlernt werden. Das Modell ist sehr detailiert und einfach erklärt in [11]. HMMs werden übrigens auch in Spracherkennungssystemen eingesetzt.

Resultate von Breglers System: Um das Interpretieren der Tätigkeiten zu

[8] über Überholmanöver-Algorithmen konnte ich keine Informationen finden

testen, wurden markierte[9] Videos eingesetzt. 90% der Tätigkeiten wurden korrekt eingestuft. Experimente auf unmarkierten Videos zeigen vielversprechende Resultate. Konkrete Zahlen konnte ich nicht finden.

Ein Grundsätzliches Problem dieses Ansatzes ist, dass wiederum nur die Bewegung an sich angeschaut wird, und nicht das Umfeld. Während Bewegungen eines einzelnen Menschen immer etwa gleich aussehen (Gangart, Gestiken, ...), so sehen dieselben Bewegungen bei vielen Menschen doch recht verschieden aus. Die Mächtigkeit von HMMs könnte man also folgendermassen darstellen:

Klassifizierbarkeit von Gesten

	1 Person	>> 1 Personen
1 Geste	sehr gut	gut
>> 1 Geste	gut	schlecht

Wie beim Erkennen von Objekten werden auch beim Klassifizieren von Bewegungen Neuronale Netze eingesetzt. Es treten auch hier dieselben Probleme und Vorteile auf wie bereits vorher erwähnt.

8 Schlussfolgerungen

Das noch ungelösste Problem im Computer Vision Bereich ist wohl "Object recognition". Isolierte Aufgaben in dem Bereich sind heutzutage durchaus realisierbar (siehe obige Beispiele). In einem integrierteren Sinn hingegen versagen die meisten Algorithmen. Es gibt beispielsweise noch keine Algorithmen, die auf einem beliebigen Bild Objekte zuverlässig identifizieren können. Zweifelsfrei ist ein kombinierter Ansatz ("bottom up" und "top down") vielversprechend. Beim Lösen von Segmentierungsproblemen hat man erst kürzlich erfolge Erzielt, was den Grundstein zur Identifizierung von Objekten sein könnte.

Im Bereich der Erkennung von Tätigkeiten – also der Interpretation von erkannten Objekten – sind vor allem Resultate bezüglich "Gangarten" erzielt worden. Die Forschung beschränkt sich noch stark auf Grundlagen. Ein Ziel (an der Stanford Universität) ist es, ein "body action coding system" zu finden; sozusagen eine Menge von Primitiv-Aktionen, aus denen sich alle Tätigkeiten von Menschen bilden lassen.

Im Bereich der Automatischen Fahrzeugsteuerung sind die grössten Projekte der 90er trotz anfänglicher Erfolge eingestellt worden. Während standart Situationen (Geradeausfahren, Überholen, Parkieren) zwar sehr gut realisierbar sind, geht man von vielen Voraussetzungen aus (Witterung[10], Markierungen, Autobahn, ...). Weiter kann eine abnormale Situation wie z.B. Schleudern sehr schnell zu komplex werden, als dass dies durch bekannte Techniken noch kontrolliert werden könnte. Eine Komerzialisierung der Systeme wird vor allem wegen zu hoher Kosten und Mangel an Bedürfnis noch eine Zeit lang Zukunftsvision bleiben.

[9] durch "Markierung" der bewegten Komponenten wird die Interpretation nicht durch falsche Segmentierung beeinflusst

[10] Keines der erwähnten Systeme funktioniert z.B. bei Nacht oder dichtem Nebel.

References

[1] Christoph Bregler. Probabilistic recognition of human actions. May 1996.

[2] Christoph Bregler. Learning and recognizing human dynamics in video sequences. In *Proc. IEEE Conf. Comp. Vision and Pattern Recognition, San Juan, Puerto Rico*, June 1997.

[3] Margaret Fleck David A. Forsyth. Finding naked people. 1996 European Conference on Computer Vision, Volume II, pp. 592-602.

[4] Ernst D. Dickmanns. Vehicles capable of dynamic vision. In *Joint Conference on Artificial Intelligence (IJCAI-97)*, August 1997.

[5] N.H. Goddard. *The Perception of Ariculated Motion: Recognizing Moving Light Displays*. PhD thesis, Dept. of Comp.Sci., Univ. Rochester, 1992.

[6] J. Malik J. Shi. Normalized cuts and image segmentation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[7] J. Weber D. Koller Q.-T. Luong J.Malik. An integrated stereo-based approach to automatic vehicle guidance. In *Int. Conference On Computer Vision*, June 1995, Boston.

[8] J. Malik J. Camillo P. McLauchlan J. Kosecka. Development of binocular steropsis vor vehicle lateral control, longitudinal control and obstacle detection. Technical report, Dep. of El.Eng. and Comp. Sciences, Berkeley, 1997.

[9] Dean A. Pomerleau. *Neural Network Perception for Mobile Robot Guidance*. Kluwer Academic Publishers, 1993.

[10] L. Concalves E.D. Bernardo E.Ursella P.Perona. Monocular tracking of the human arm in 3d. In *Proc. Int. Conf. Computer Vision*, 1995.

[11] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. IEEE Log Number 8825949, January 1988.

[12] K. Rohr. Incremental recognition of pedestrians from image sequences. In *IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn., pages 8-13, New York City*, June 1993.

[13] W. Freeman M. Roth. Orientation histograms for hand gesture recognition. In *International Workshop on Automatic Face- and Gesture-Recognition*, 1995.

[14] J. Malik S. Belongie T. Leung J. Shi. Contour and texture analysis for image segmentation. In *Perceptual Organization for Artificial Vision Systems*. K.L. Boyer and S. Sarkar, editors. Kluwer Academic Publishers, 2000.

[15] Charles E. Thorpe Todd M. Jochen, Dean A. Pomerleau. Vision guided lane transition. In *IEEE Symposium on Intelligent Vehicles*, Septemper 25-26, 1995 Detroit.

[16] Hideki Koike Yoichi Sato, Yoshinori Kobayashi. Fast tracking of hands and fingertips in infrared images for augmented desk interface. In *IEEE Automatic Face and Gesture Recognition*, March 2000.