Sehen und Verstehen: Kamerabasierte Überwachung der Fahrzeugumgebung

Zusammenfassung

In heutigen Kraftfahrzeugen kommen immer mehr Sensorsysteme zur Erkennung und Interpretation des Fahrzeugumfeldes zum Einsatz: Sie versorgen komplexe Fahrerassistenzsysteme mit den relevanten Daten und erlauben damit erst die vielfältigen Funktionen zur Unterstützung des Fahrers in den unterschiedlichsten Fahrszenarien.

Das für die Führung eines Automobils wichtigste Sinnesorgan des Fahrers ist zweifellos der Gesichtssinn (das Auge). Ohne die visuelle Erfassung des Umfelds lässt sich ein Fahrzeug im Straßenverkehr unter „normalen“ Bedingungen nicht manövrieren. Dies ist der Hauptgrund, warum kamerabasierten Assistenzsystemen eine solche Aufmerksamkeit entgegengebracht wird.

Kamerasysteme sind aufgrund der großen Variationsbreite an möglichen physikalischen Parametern und dem vielgestaltigen geometrischen Aufbau dem menschlichen Gesichtssinn mit seinen natürlichen Grenzen und Einschränkungen grundsätzlich gleichrangig, in manchen Teilbereichen gar überlegen. Dies gilt jedenfalls dann, wenn wir nur die objektiven optisch-physikalischen Daten zugrunde legen.

Trotz dieser hohen Leistungsfähigkeit ist die verlässliche Überwachung der Fahrzeugumgebung im Sinne einer universellen Assistenz oder gar die Übernahme von sicherheitsrelevanten Teilaufgaben mit aktuellen Systemen nicht möglich. Hier ist der Mensch, ungeachtet seiner biologischen Limitierungen, deutlich überlegen. Das technische System ist zwar gut in manchen Spezialfunktionen, z.B. Nachtsichtfähigkeit, versagt aber insbesondere bei den komplexeren übergreifenden Aufgaben, z.B. bei der Klassifizierung von Objekten, beim Herstellen von Objektrelationen oder bei der Zuordnung von Kontextinformationen. Eine echte Szeneninterpretation, wie sie der Fahrer ohne größere Mühen jederzeit bereit hält und permanent aktualisiert, liegt weit jenseits der Möglichkeiten aktueller Systeme. Dazu sind heutige Assistenzsysteme noch nicht in der Lage. Dieses ganzheitliche Szenenverständnis ist aber erforderlich für die Realisierung von komplexeren Komfort- und Sicherheitsfunktionen und ist eine unbedingte Notwendigkeit für die Umsetzung des Fernziels *Autonomes Fahren*.

Seeing and understanding: Camera-based supervision of the vehicle surroundings

Summary

In today's motor vehicles, more and more sensor systems for detection and interpretation of the vehicle environment are being used: they provide the relevant data to complex driver assistance systems and are the essential input to assist the driver in various driving scenarios by means of different functions.

The most important driver’s sensory organ necessary for the operation of a vehicle is undoubtedly the sense of sight (the eye). Manoeuvring a vehicle in road traffic is not possible otherwise. This is the main reason why camera-based assistance systems attract such a great deal of attention.

Due to the wide variation of possible physical parameters and the varied geometric structure, camera systems principally are coequal compared with the human sense of sight with its natural limits and restrictions. More over, such systems are even superior in some areas. This is certainly the case, if we focus on the bare optical-physical data.

Despite of this high efficiency, the reliable supervision of the vehicle surroundings in terms of a universal assistant or even the adoption of safety-related tasks is not possible with current systems. Here, the human is clearly superior, despite of its narrow biological limits. The technical system is excellent in some special functions, e.g. night vision. However, in a lot of principal issues the technical system fails, especially in the more complex cross-functional tasks, e.g. in the classification of objects, in the determination of object relations, or in the assignment of context information. A real scene interpretation is far beyond the capability of current systems. On the other side, the driver has no difficulties to create such a clear scene understanding and to keep it constantly updated. Those modern assistance systems are still not able to replace the human excellence. Nevertheless, this holistic scene understanding is required for the realization of more complex comfort, safety and security features and is a crucial need to implement the ultimate objective *Autonomous Driving*.

Einleitung und Status

Kameras erfassen das Systemumfeld, Objekte werden detektiert und klassifiziert: Das ist „Sehen“. Es wird die Frage nach dem Was und Wo beantwortet. An welchem Ort in der Szene befinden sich welche Objekte? In Ansätzen sind das zugleich die ersten, noch lange nicht abschließend ausgearbeiteten Lösungen in Richtung „Verstehen“. Der darauf folgende, noch sehr große Schritt: Die ganzheitliche logische Interpretation von Szenen, das kognitive Verarbeiten der physikalischen Außenweltsicht ist sehr komplex und wird heute in den vielfältigen Querbeziehungen und aufwendigen Verarbeitungsprozessen noch nicht im Sinne einer praxistauglichen und echtzeitfähigern Anwendung beherrscht. In der Abfolge Sehen – Verstehen sind wir erst auf dem Weg vom Ersteren zum Zweiten.

Gerade im Bereich der Automotive-Bildverarbeitung wurden bisher viele Lösungen für sehr spezielle Probleme wie die Fußgängererkennung oder die Fahrspurerkennung entwickelt. Die realisierten Systeme erreichen zwar eine akzeptable Erkennungsrate, agieren aber noch weit weg von einem Szenenverständnis. Der Hauptgrund ist: Ihr Auswertealgorithmus beschränkt sich auf simple Mustervergleiche und Plausibilitätschecks zur Erkennung ganz bestimmter vordefinierter Objekte.

Ein Beispiel zur Darstellung der daraus resultierenden Problematik: Aktuelle Verkehrszeichenerkennungssysteme erkennen auch solche Verkehrszeichen als relevant, die beispielsweise an den Rückwänden von LKWs angebracht sind. Im Effekt zeigt dann ein solches System die Höchstgeschwindigkeit „80 km/h“ mitunter auch dann an, wenn tatsächlich gar keine Beschränkung vorliegt. Ein anderes Beispiel: Spurerkennungssysteme sehen fälschlicherweise eine Leitplanke als Spurmarkierung an.

Solche Unzulänglichkeiten führen dazu, dass diese Systeme derzeit nicht für sicherheitskritische Anwendungen eingesetzt werden können. Die für die Umsetzung des Autonomen Fahrens erforderliche sehr hohe Verlässlichkeit ist damit lange noch nicht erreicht. Die Schwächen sind maßgeblich dadurch zu beheben, dass die Bildauswertung sukzessiv durch weitere Kontextinformation angereichert wird. Dabei können die Zusatzinformationen größtenteils aus dem Bild selber gewonnen werden, bisweilen macht es aber auch Sinn, sie von externen Quellen zu beziehen (Digitale Karte, Car2X, etc.).

In einem verwandten Gebiet, dem Luftfahrtsektor, geht der Trend mit hoher Geschwindigkeit zu (teil-) autonomen Fluggeräten (Unmanned Aerial Vehicles = UAVs), die sich weitgehend autark in ihrer Umgebung bewegen. Umfangreiche Sensorik soll dabei für eine schnelle Lagebeurteilung sorgen. Für die Verarbeitung der hierbei anfallenden Fülle an Informationen ist eine extrem leistungsfähige Echtzeit-Bilddatenverarbeitung für Objekterkennung und Tracking etc. erforderlich. Und auch hier zeigt sich die Notwendigkeit eines umfassenden Verarbeitungsansatzes.

Auch jenseits dieser Beispiele gibt es vielfältige Aufgabenstellungen in der Echtzeit-Bilddatenverarbeitung, z.B. kann eine Kamera auch als metrisches Instrument, etwa zur exakten Größenbestimmung von Bauteilen in der industriellen Bildverarbeitung verwendet werden.

Immer dann, wenn die Aufgabenstellungen eng umrissen und die Lösungen auf diese ganz bestimmten Anwendungen zugeschnitten werden, lassen sich auf der Basis von optischen Systemen sehr effektive Assistenzsysteme entwickeln. Wird das Umfeld komplexer, stoßen heutige kamerabasierte Assistenzsysteme indessen schnell an ihre Grenzen. So ist es derzeit z.B. noch nicht möglich, für ein Kreuzungsszenario im Straßenverkehr ein adäquates Szenenverständnis aufzubauen und dauerhaft aufrecht zu erhalten.

Für den Typus der letztgenannten Aufgabenstellung ist eine Interpretation der Bilddaten notwendig, wie sie mit vertretbarem Aufwand an Ressourcen nur von biologischen Systemen in ausreichender Qualität beherrscht wird. Hierbei geht es um Klassifikationsaufgaben, also die Zuordnung von Objekten im Bildraum zu Symbolen real existierender Entitäten (Fußgänger, Fahrzeuge, Verkehrsanlagen, Flugobjekten bis hin zu Menschenansammlungen). Eine weitere zentrale Aufgabe ist die Bestimmung von Entfernungen, Objektgrößen und Bewegungen in der realen Umgebung. Diese Informationen werden benötigt, um den räumlichen und zeitlichen Kontext herzustellen. Dazu mehr im nächsten Abschnitt.

Analyse und Lösungsansatz

Rein physikalisch gesehen können Kameras einen ähnlichen Ausschnitt der relevanten Fahrzeugumgebung wahrnehmen, wie er typischerweise auch vom Fahrer erfasst wird. Die technischen Parameter

von Kameras, Öffnungswinkel, Brennweite, Auflösungsvermögen, erfasstes Frequenzspektrum können, je nach genauer Aufgabe, fast immer so gewählt werden, dass die Leistungsfähigkeit des menschlichen Auges in Summe erreicht oder gar übertroffen wird. In manchen Bereichen haben Kameras gar deutliche Vorteile, z.B. bei der Nachtsichtfähigkeit. Durch den Zusammenschluss mehrerer Kameras kann sogar eine permanente Rundumsicht erreicht und damit eine vollständige Erfassung der Fahrzeugumgebung jederzeit gewährleistet werden (Surround-View-System), was dem Fahrer schon wegen seines begrenzten Blickwinkels und aufgrund von Abschattungen durch die Karosserie nicht möglich ist. Freilich erfordert die Verarbeitung der Datenströme von solchen Kameraverbünden mit unterschiedlichen optischen Parametern und Frequenzbereichen einen höheren technischen Aufwand für die Herstellung einer geometrisch und physikalisch konsistenten Gesamtsicht. Grundsätzlich ist das aber machbar und wird auch schon technisch umgesetzt.

Optische Systeme zur Erfassung der Fahrzeugumgebung sollen das Auge des Fahrers zunächst einmal nicht ersetzen, in erster Linie zielen sie daraufhin ab, den Fahrer zu unterstützen, z.B. durch Überwachung von Bereichen, die vom Fahrer nicht einsehbar sind. Beispiele hierfür sind Rückfahrkamera und Blind Spot Detection (Überwachung des „toten Winkels“). Durch solche Assistenzsysteme mit ganz eng abgesteckten Aufgaben wird in bestimmten Fahrszenarien der Komfort für den Fahrer gesteigert. In Ansätzen wird damit bereits eine im weitesten Sinne sicherheitsrelevante Funktion zur Vermeidung einer Fehlleistung des Fahrers realisiert. Für diese vergleichsweise einfachen Aufgaben reicht es aus, das betreffende Kamerabild im Sichtfeld des Fahrers zur Anzeige zu bringen oder, im Falle der Blind Spot Detection, auf Basis eines relativ einfachen Objekterkennungsalgorithmus ein Signal auszulösen.

Sehr viel anspruchsvoller ist es, die Fahrzeugumgebung insgesamt zu überwachen, also Objekte zu detektieren und klassifizieren sowie die Situation ganzheitlich zu interpretieren. Dazu sind heutige Assistenzsysteme noch nicht in der Lage. Dieses ganzheitliche Szenenverständnis ist aber erforderlich für die Umsetzung des Fernziels *Autonomes Fahren*.

In der nachfolgenden Tabelle werden einige wichtige Merkmale von optischen Assistenzsystemen zu den grundsätzlichen menschlichen Fähigkeiten in Bezug gesetzt.

Leistungsmerkmal	**Mensch und Auge**	**Assistenzsystem und Kamera(s)**
Frequenzspektrum	O	++
Erfassung von Farbinformationen	O	++
Fähigkeit zur Objekterkennung (Kategorisierungsleistung)	++	-
Nachtsichtfähigkeit	-	++
Fusionsfähigkeit	+	-
Schnelligkeit der Objekterkennung	++	-
Bestimmung der Objektposition	+	O
Bestimmung der Objektgeschwindigkeit	+	O
Objektklassifizierung	++	--
Erkennung der Eigenbewegung	++	O
Multi Objekt-Verfolgung	+	+
Herstellen von Objektrelationen	++	--

Leistungsmerkmal	Mensch und Auge	Assistenzsystem und Kamera(s)
Ableitung der räumlichen Kontextinformation	++	-
Ableitung der zeitlichen Kontextinformation	++	-
Herstellen eines Szenenverständnisses	++	--

Tabelle 1: Relative Leistungen von heutigen kamerabasierten Assistenzsystemen im Vergleich zum Menschen

Wie bereits oben angedeutet, sind Kamerasysteme aufgrund der großen Variationsbreite an möglichen physikalischen Parametern und dem vielgestaltigen geometrischen Aufbau dem menschlichen Gesichtssinn mit seinen natürlichen Limitierungen grundsätzlich gleichrangig, in manchen Teilbereichen gar überlegen. Dies gilt jedenfalls dann, wenn wir nur die nackten optisch-physikalischen Daten zugrunde legen. Warum aber ist die verlässliche Überwachung der Fahrzeugumgebung im Sinne einer universellen Assistenz oder gar der Übernahme von sicherheitsrelevanten Teilaufgaben trotzdem so schwierig? – Die Antwort liegt in der Betrachtung der jeweiligen Gesamtsysteme: Zwar sind die optischen Sensoren von Mensch und Maschine grundsätzlich vergleichbar, die nachgeschaltete Verarbeitung der Signale ist indessen vollkommen verschiedenartig. In den meisten Fällen ist hier ist das biologische System sehr viel leistungsfähiger. Der Verarbeitungsprozess macht also den wesentlichen Unterschied.

Worin aber genau liegen die Unterschiede? Wenn man sich Tabelle 1 betrachtet, dann erkennt man, dass das technische System auf der physikalischen Ebene durchaus Vorteile hat, es ist aber insbesondere dann deutlich schwächer, wenn es darum geht, komplexe logische Bezüge herzustellen: Um was für eine Art von Objekt handelt es sich (Objektklassifizierung)? Wie steht dieses Objekt in Beziehung zu anderen Objekten der Szene (Objektrelation)? Wie ist die Beziehung zur Objektumgebung (räumlicher Kontext)? Welche zeitlichen Abhängigkeiten gibt es (zeitlicher Kontext)? Und wie wirken alle Objekte zusammen (Szenenverständnis)?

Über allem schwebt die Frage: Was überhaupt ist relevant? Für das technische System ist das die ganz große, die zentrale Herausforderung. Es geht darum, aus dem komplexen Echtzeitdatenstrom, ggf. von mehreren Kameras, die für die Szeneninterpretation wichtigen Bildinformationen zu identifizieren. Für den Menschen ist das in den meisten Fällen vergleichsweise einfach.

Ein Ansatz für die Herstellung der nötigen Szeneninterpretation liegt nahe: Vollständige Erfassung und Verarbeitung der physikalisch-geometrischen Umwelt- und Objektparameter (Orts- und Bewegungsvektor, Zeitstempel) im Sinne einer Objektliste und die permanente Umweltbeschreibung. Das ist, technisch gesehen, sehr anspruchsvoll, weil sehr große Datenmengen von unterschiedlichen Quellen in Echtzeit verarbeitet werden müssen, erscheint aber, algorithmisch gesehen, fast schon trivial und als die einzig richtige Lösung.

Diese Vorgehensweise läuft im Effekt auf eine Brute-Force-Methode (Exhaustionsmethode) hinaus: Es wird einfach alles betrachtet, was möglich ist. Der Lösungsraum möglicher Trajektorien, Entscheidungen, Aktionen im Szenario wird dazu erschöpfend durchlaufen und nach der jeweiligen Wahrscheinlichkeit des Eintretens bewertet. Aufgrund der hohen Dimensionalität des zugrunde liegenden Merkmalsraums und der daraus resultierenden Kombinatorik erweist sich dieser Ansatz indessen als höchst problematisch. Die Anforderungen an die HW-Ressourcen sind so enorm hoch, dass auf absehbare Zeit auf diesem Wege und zu vertretbaren Kosten ein Erfolg nicht zu erwarten ist. – Selbstverständlich muss diese Erfassung der objektiven Umweltdaten stattfinden, dies allein führt aber noch nicht weiter. Denn, auch wenn jederzeit die Koordinaten und Bewegungsvektoren aller Objekte in der Fahrzeugumgebung bekannt sind, wissen wir kaum mehr über die Relevanz für das eigene Fahrzeug und die eigenen Handlungsalternativen. Allenfalls können wir eine simple Betrachtung zu eventuell möglichen Kollisionen vornehmen und darauf reagieren

Der Anspruch auf vollständige, Erfassung und Verarbeitung der objektiven, physikalisch-geometrischen Umwelt- und Objektparameter erinnert an das philosophische Konzept des französischen Mathematikers und Physikers Pierre Simon de Laplace (1749 – 1827). Er konstruierte einen die

Welt rational vollständig erfassenden universellen Geist, den sogenannten *Laplaceschen Dämon*, der aufgrund seiner erschöpfenden Kenntnis der Gegenwart in allen physikalischen Fakten die Zukunft vorherzusehen in der Lage sei. Dahinter steckt letztlich die Annahme eines totalen Determinismus. Angewandt auf unser Problem sollten wir also nach der kompletten Erfassung der Umwelt- und Objektparameter und unserer Vertrautheit mit den Naturgesetzen in der Lage sein, die Fahrzeugumgebung detailliert zu beschreiben und die Entwicklung der Szene in allen Einzelheiten genau und zuverlässig zu prognostizieren.

Aufgrund des heute in aller Breite akzeptierten quantentheoretischen Erklärungsrahmens und der damit einhergehenden stochastischen Interpretation von quantenmechanischen Experimenten wird solcher Determinismus allgemein als nicht schlüssig angesehen. Mit anderen Worten: Der *Laplacesche Dämon* ist tot. Ungeachtet dessen könnten wir mit der objektiven Erfassung der physikalisch-geometrischen Umwelt- und Objektparameter dessen hypothetische Fähigkeiten ohnehin nur unvollkommen nachvollziehen. Wir erfahren auf diesem Wege wenig über den Charakter der Objekte als systemische Entitäten und absolut nichts über deren Absichten.

Die nackte und beziehungslose Datenaufnahme bringt uns also der Problemlösung nicht näher. Darüber hinaus brauchen wir insbesondere die

- Ableitung der räumlichen Kontextinformation
- Ableitung der zeitlichen Kontextinformation
- Die Herstellung von Objektrelationen

Diese Informationen bekommen wir aber nicht als objektive Größen von der Kameraoptik. Wir brauchen also einen anderen Ansatz.

Worum eigentlich geht es im Kern bei dem Anspruch, die Situation, in welcher sich das Fahrzeug befindet, die Szene, ganzheitlich zu interpretieren, zu verstehen. Ziel ist es doch, Erkenntnis über die Welt zu gewinnen, zumindest Erkenntnis über die unmittelbare Fahrzeugumgebung, und daraus Schlüsse zu ziehen. Es geht also um die Wahrnehmung der Fahrzeugumgebung in Raum und Zeit und die Spiegelung dieser Eindrücke an unserem eigenen Verstand, unserer Erfahrung, unserem Vorwissen. Imanuel Kant (1724 – 1804) hat die diesem Erkenntnisprozess zugrundeliegenden Voraussetzungen in aller Breite und Tiefe beleuchtet und diskutiert.

Nach Kant erkennen wir die Dinge in Bezug auf die in uns angelegten raum-zeitlichen Vorstellungen. In gewissem Sinne sind also unsere Anschauung und unser Begriff von der Welt nicht objektiv, sondern subjektiv. Und trotzdem ist dieser Ansatz offenbar überaus erfolgreich. Warum? Aus zwei Gründen: 1. Weil in der Subjektivität Vorwissen über die Welt steckt. 2. Weil dieses Vorwissen dazu benutzt werden kann, den Prozess der Welterkenntnis extrem schnell und effektiv zu machen. Interessanterweise werden z.B. nur etwa 10% der Nervenfasern aus dem Auge ins Sehzentrum geleitet. Nichtsdestotrotz bekommen wir auf dieser Basis bekanntermaßen ein stimmiges Bild der Welt. Es gelingt, weil das gewissermaßen fragmentarisch Gesehene im Zuge der Verarbeitung in Bezug gesetzt wird zum Erfahrungsschatz des Individuums und der Spezies und auf diesem Wege die fehlenden Informationen, ohne dass dies bewusst wird, ergänzt werden. Dieses von der Evolution gefundene Konzept ist sehr leistungsfähig, es ist aber nicht perfekt, wie die Beispiele von optischen Täuschungen zeigen. Dennoch ist der Ansatz in fast allen praktisch relevanten Fällen der oben skizzierten Exhaustionsmethode weit überlegen.

Zusammenfassend kann man sagen: Wahrnehmen und Verstehen sind keine absoluten Kategorien, sie beziehen sich stets auf das eigene Denken und die eigenen Ziele. Das ist die Sicht, wie sie im Übrigen auch von der modernen Wahrnehmungspsychologie einhellig vertreten wird. Dieser kleine Exkurs soll hinführen auf den vorgeschlagenen Lösungsansatz.

Übertragen auf das technische System steht zunächst einmal die objektive Erfassung von Daten im Vordergrund (das ist die Entsprechung zur sinnlichen Wahrnehmung), dann aber kommt etwas Wesentliches hinzu was in Analogie steht zu den Verarbeitungsabläufen im menschlichen Erkenntnisprozess: Die Bezugnahme auf geeignet repräsentiertes semantisches Wissen, die Referenzierung auf ein episodisches Gedächtnis und die Fähigkeit zur Inferenz in diesem Gesamtgefüge.

Im Ergebnis folgt daraus ein hierarchisches Modell der Fahrzeugumgebung, in das sich die Modelle der Objekte mit ihren hypothetischen (wahrscheinlichen) Verhaltensweisen nahtlos einbetten.

Als grundsätzliche Zielrichtung wird also vorgeschlagen, die Verarbeitungsprozesse an dem in Jahrmillionen von der Natur im Rahmen der evolutionären Auslese erarbeiteten Vorbild auszurichten. Dazu gehören:

- Konsequente Informationsfusion in einer frühen Verarbeitungsphase
- Objektklassifizierung unter Einbeziehung von Erfahrungswissen
- Bestimmung von räumlichen und zeitlichen Kontextinformationen
- Dauerndes Herstellen der Objektrelationen
- Permanente Antizipation des Verhaltens der Objekte

Im Kern geht es darum, durch permanentes Herstellen, Überprüfen, Verwerfen und Bestätigen von Hypothesen zur Objektklassifizierung, zu den räumlichen und zeitlichen Kontextinformationen, zu den Objektrelationen und zum Objektverhalten die Szene in einer dem menschlichen Verständnis nahe kommenden Weise kognitiv zu verarbeiten und so gesamthaft darzustellen und fortgesetzt zu interpretieren.

Die zugrunde liegenden Umweltinformationen werden durch eine Fülle von unterschiedlichen Sensoren (auch nicht-optischen) dezentral erfasst. Rein technisch gesehen ist dabei ein wichtiger Aspekt, dass die daraus abgeleitete Umweltrepräsentation letzten Endes zentral erfolgt und zentral verfügbar gemacht wird. Dies ist die Basis für den darauf fußenden, permanent laufenden Prozess der ganzheitlichen Szeneninterpretation.