Künstliche Wahrnehmung auf Basis ortsfester Sensoren

Kurzfassung der Dissertation

André Ibisch

Ein künstliches Wahrnehmungssystem auf Basis ortsfester Sensoren ist ein in sich geschlossenes datenverarbeitendes Hardware- und Softwaresystem in einer vorhandenen Infrastruktur. Im Vergleich zu mobilen Systemen (z. B. Roboter) weisen ortsfeste Wahrnehmungssysteme aufgrund der nicht vorhandenen Eigenbewegung der Sensoren eine geringere Unsicherheit auf und sind somit in der Lage, eine Lokalisierung mit einer höheren Genauigkeit durchführen zu können. Für die Bereiche des Autonomen Fahrens sowie für die Detektion von Gruppenemotionen wird die Entwicklung und Umsetzung verschiedener Systeme dargestellt.

Für die Unterstützung des autonomen Fahrens in einem Parkhaus werden zwei Lokalisierungssysteme, ein LIDAR-basiertes und ein kamerabasiertes, entwickelt, die die Position eines autonomen Fahrzeugs sowie aller anderen bewegten Objekte in Echtzeit mit einer hohen Präzision schätzen und vor Gefahren (z. B. kreuzende Fußgänger oder Fahrzeuge) auf dem geplanten Pfad warnen.

Für diese Lokalisierung wird ein LIDAR-basiertes System vorgestellt: Die Laserstrahlen eines einzelnen LIDAR-Sensors werden durch eine adaptive *Grid*-Repräsentation des Hintergrunds als bewegt bzw. statisch klassifiziert. Anschließend werden die bewegten Strahlen aller Sensoren aggregiert und durch eine Transformation in ein gemeinsames Koordinatensystem des Parkhauses überführt. Auf diesen Daten werden durch einen adaptierten RANSAC-Algorithmus Fahrzeug- und Fußgänger-Hypothesen generiert. Diese werden durch einen erweiterten Kalman-Filter zeitlich geglättet und bilden somit die finale Systemausgabe als Position auf einer Karte des Parkhauses.

Da Überwachungskameras bereits in einem Großteil der Parkhäuser installiert sind, wird als kostengünstigere Alternative ein kamerabasiertes Wahrnehmungssystem für die Lokalisierung allgemeiner Objekte vorgeschlagen. Für jede an das System angeschlossene Kamera wird eine Vordergrundsegmentierung durchgeführt. Die resultierenden Bildbereiche aller Kameras werden aggregiert, indem diese in eine gemeinsame 3D-Repräsentation des Parkhauses transformiert werden. In dieser Repräsentation werden Sichtstrahlen durch die bewegten Bildbereiche einer jeden Kamera projiziert. Durch eine Kombination der Sichtstrahlen werden plausible Objekthypothesen generiert. Diese werden zeitlich integriert und an angeschlossene Systeme weitergeleitet. Eine experimentelle Klassifikation der allgemeinen Objekte wird ebenfalls beschrieben.

Die durchschnittliche Abweichung des LIDAR-basierten Systems zu einem DGPS-Referenzsystem mit 0,19 m sowie die des kamerabasierten Systems zum eigenen LIDAR-Referenzsystem mit 0,24 m bieten eine in der Literatur noch nicht erreichte Genauigkeit.

In einer von der DFG geförderten Zusammenarbeit mit Soziologen wird ein System für die videobasierte Detektion von Gruppenemotionen bei Großveranstaltungen (wie z. B. in Fußballstadien, Musikfestivals etc.) beschrieben. Dieses soll die Ordnungskräfte (z. B. Polizei, Feuerwehr etc.) mit Informationen unterstützen: Zum einen soll der Fokus auf relevante Ereignisse gelenkt werden, zum anderen soll prädiziert werden, welche abnormalen Situationen (z. B. eine Massenpanik durch Gedränge oder Ausschreitungen durch kleine Gruppen) in unmittelbarer Zukunft auftreten können. Da das soziologische Konzept von Gruppenemotionen abstrakt definiert ist, müssen konkrete Hinweise aus Videos mit Methoden der digitalen Bildverarbeitung extrahiert werden, die in verschiedenen eskalierenden Situationen beobachtet wurden und ein Indiz dafür sind, dass in einer vergleichbaren Situation das Risiko einer abnormalen Situation gleich hoch ist.

Die vorgeschlagene *Dichteschätzung* wird durch eine neue Methode für die Detektion von dichten Menschenmengen realisiert: Durch einen erweiterten Personendetektor werden 91 % aller Personen einer Menschenmenge erkannt. Anschließend werden diese Detektionen auf eine Repräsentation der Umgebung projiziert, um dort die Dichte bestimmen zu können. Ziel ist es, in einem späteren Anwendungsszenario zu hohe Dichten (wie z. B. bei der Loveparade 2010 in Duisburg) rechtzeitig erkennen zu können, um entsprechende Gegenmaßnahmen einzuleiten.

Das Auftreten von aggressiven *Aktivitäten* von Einzelpersonen (wie z. B. treten, schlagen etc.) sowie kollektive Aktivitäten in und von Gruppen (wie z. B. schnelleres geschlossenes Laufen) deutet auf eine drohende oder bereits stattfindende Eskalation hin. Die Detektion dieser Aktivitäten wird durch das Trainieren von annotierten Beispielhandlungen und deren Klassifikation als raum-zeitliche Merkmale anhand einer bildbasierten Verarbeitungskette ermöglicht: Für jede detektierte Person werden Raum-Zeit-Subvolumen abhängig von den Personen im Umfeld generiert und durch ein *Random-Forest*-Verfahren mit einer durchschnittlichen Genauigkeit von 65 % klassifiziert.

Die Detektion von *Umschlagsverhalten* sucht in Aufnahmen von Menschenmengen nach einer qualitativen Änderung des optischen Flusses und kombiniert diese mit einem Modell, das soziale Kräfte in Menschenmengen abbildet. Das vorgestellte System erkennt diese mit einer durchschnittlichen Genauigkeit von ca. 90 %. Ziel in einer möglichen Anwendung ist es, den zeitlichen Wendepunkt, an dem eine normale Situation in eine abnormale (z. B. eine Panik) umschlägt, direkt zu erkennen, um die Eingriffszeiten von Rettungs- und Ordnungskräften zu verkürzen.