Visual urban scene analysis by moving platforms

Author(s):
Ess, Andreas

Publication Date:
2009

Permanent Link:

https://doi.org/10.3929/ethz-a-005957024 →

Diss. ETH No. 18663

Visual Urban Scene Analysis by Moving Platforms

A dissertation submitted to the

ETH ZURICH

for the degree of

Doctor of Science ETH

presented by

Andreas Ess
M.Sc. ETH Zurich
born 15. December 1980
citizen of Austria

accepted on the recommendation of

Prof. Dr. Luc Van Gool, ETH Zurich and KU Leuven, examiner
Prof. Dr. Ram Nevatia, University of Southern California, co-examiner
Prof. Dr. Bastian Leibe, RWTH Aachen University, co-examiner

2009

Abstract

In recent years, interest in mobile robots and intelligent vehicles that are able to act autonomously in scenarios of daily human living has been increasing constantly. This trend has been further fostered by advances in input sensor technology such as cameras or lasers, as well as in the corresponding algorithms for data interpretation. Due to their relatively low price and similarity to human vision, cameras are an especially intriguing sensor for such tasks.

This dissertation presents a purely vision-based system for the tasks of self-localization, scene analysis, and object tracking in semi-crowded urban environments, with a mobile platform as observer. Starting from a set of off-the-shelf components for object detection and stereo depth estimation, we first propose a probabilistic combination of these independent cues to simultaneously validate object hypotheses and estimate a common ground plane where these objects reside. Based on the the improved detections and self-localization from visual odometry, we then introduce a multi-object tracking system that recovers the objects' trajectories while explicitly reasoning about physical space requirements between each other. The different algorithms are closely coupled by a set of feedback channels, with failure detection mechanisms in each component to ensure robust operation in crowded scenarios.

The resulting, integrated mobile vision system is then applied to the task of pedestrian and car tracking in urban environments, where we demonstrate robust and computationally efficient tracking over prolonged time spans. Taking this system as a basic component, we then explore its use for creating dynamic occupancy maps that could serve as input to a path planning module. For pedestrians, we moreover show that the system can be used as a building block for articulated multi-body tracking; and that the motion model for pedestrians can be improved when

incorporating the interaction of agents along with the knowledge about the scene in an approach inspired by social simulations. Furthermore, we propose a method to analyze the scene by segmenting it into a set of texture labels. The resulting intermediate representation can be used to infer the road type ahead of the observer as well as the presence of various object classes, even without employing a dedicated detector.

All algorithms are evaluated on several challenging, realistic video sequences recorded in busy inner-city locations with a set of representative platforms. Our experiments corroborate our claim that sensor-related robotic tasks in daytime navigation and object tracking can be performed using vision only instead of using often considerably more expensive sensor arrays.

Zusammenfassung

In den letzten Jahren nahm das Interesse an mobilen Robotern und intelligenten Autos, die sich selbständig in natürlichen Umgebungen zurechtfinden können, ständig zu. Dieser Trend wurde weiter gefördert durch Fortschritte im Bereichen von Sensortechnologien, wie z.B. Kameras oder Laser, als auch den entsprechenden Algorithmen zur Dateninterpretation. Aufgrund ihres relativ niedrigen Preises und ihrer Ähnlichkeit zum menschlichen Sehen sind vor allem Kameras interessante Sensoren für Aufgaben im Bereich der Robotik.

Diese Dissertation beschreibt ein System zur Selbstlokalisierung, Szenenanalyse und Objektverfolgung in bevölkterten Stadtgebieten. Das System basiert ausschliesslich auf Bilderkennung, wobei eine mobile Plattform als Beobachter eingesetzt wird. Ausgehend von einer Reihe verfügbarer Komponenten zur Objekterkennung und Stereotiefenschätzung schlagen wir als erstes eine probabilistische Kombination dieser unabhängigen Hinweise zur gleichzeitigen Validierung von Objekthypothesen und Schätzung einer gemeinsamen Strassenebene vor. Aufbauend auf den dadurch verbesserten Detektionen sowie Selbstlokalisierung mittels kamerabasierter Odometrie führen wir als nächstes ein Multi-Objekt-Verfolgungssystem ein. Dieses findet Objekttrajektorien indem es explizit den physikalisch benötigten Platz zwischen den einzelnen Objekten in Betracht zieht. Die verschiedenen Algorithmen sind durch diverse Informationskanäle miteinander verbunden, wobei Mechanismen zur Fehlerdetektion in jeder einzelnen Komponente vorhanden sind, um so eine robuste Anwendung in verkehrsreichen Szenarien zu garantieren.

Das daraus entstehende, integrierte mobile Bilderkennungssystem wird dann auf die Verfolgung von Fussgängern und Autos in städtischen Umgebungen angewandt, wobei eine robuste und rechnerisch effiziente Objektverfolgung über längere Zeiträume demonstriert wird. In einem wei-

teren Schritt untersuchen wir das System als Eingabe für weitere Aufgaben in der Erstellung von Belegungskarten. Für die Objektklasse Fussgänger demonstrieren wir zudem, dass das zugrundeliegende System als Baustein für die Verfolgung von mehreren Objekten mit artikulierten Modellen verwendet werden kann und dass das Bewegungsmodell für Fussgänger verbessert werden kann, wenn die Interaktionen zwischen den einzelnen Agenten in einem von Simulationen inspirierten Ansatz beachtet wird. Des weiteren schlagen wir eine Methode zur Szenenanalyse vor, die ein Bild in eine vorgegebene Anzahl von Texturarten segmentiert. Die entstehende Zwischenrepräsentation kann benutzt werden, um Aussagen über die Strassengeometrie und das Vorhandensein diverser Objektklassen zu machen, ohne einen dedizierten Objektdetektor zu verwenden.

Sämtliche Algorithmen werden auf mehreren anspruchsvollen und realistischen Videosequenzen getestet, die in belebten innerstädtischen Umgebungen mit einer Reihe repräsentativer Plattformen aufgenommen wurden. Die vorgestellten Experimente stützen unsere Behauptung, dass sensorbezogene Robotikaufgaben in Navigation und Objektverfolgung mit Bilderkennung alleine bewerkstelligt werden können, ohne dass dabei auf weitaus teurere Sensoranordnungen zurückgegriffen werden muss.