Selected Topics in Computer Vision

Dennis Mitzel

Taking Mobile Multi-Object Tracking to the Next Level

Band 1

Herausgeber: Prof. Dr. Bastian Leibe

Lehr- und Forschungsgebiet Informatik 8
Computer Vision Group

RWTH AACHEN UNIVERSITY

Selected Topics in Computer Vision

herausgegeben von
Prof. Dr. Bastian Leibe
Lehr- und Forschungsgebiet Informatik 8
(Computer Vision)
RWTH Aachen University

Band 1

Dennis Mitzel

Taking Mobile Multi-Object Tracking to the Next Level

Shaker Verlag
Aachen 2014

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Zugl.: D 82 (Diss. RWTH Aachen University, 2013)

Printed in Germany.

ISBN 978-3-8440-2524-8
ISSN 2198-3372

Shaker Verlag GmbH • P.O. BOX 101818 • D-52018 Aachen
Phone: 0049/2407/9596-0 • Telefax: 0049/2407/9596-9
Internet: www.shaker.de • e-mail: info@shaker.de

Taking Mobile Multi-Object Tracking to the Next Level

Der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University vorgelegte Dissertation zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften

von Diplom-Informatiker
Dennis Mitzel
aus Katschar, Kasachstan

Berichter: Prof. Dr. Bastian Leibe
Prof. Dr. Luc Van Gool

Tag der mündlichen Prüfung: 30.09.2013

Abstract

Recent years have seen considerable progress in automotive safety and autonomous navigation applications, fueled by the remarkable advance of individual Computer Vision components, such as object detection, tracking, stereo and visual odometry. The goal in such applications is to automatically infer semantic understanding from the environment, observed from a moving vehicle equipped with a camera system. The pedestrian detection and tracking components constitute an actively researched part in scene understanding, important for safe navigation, path planning, and collision avoidance.

Classical *tracking-by-detection* approaches require a robust object detector that needs to be executed in every frame. However, the detector is typically the most computationally expensive component, especially if more than one object class needs to be detected. A first goal of this thesis was to develop a vision system based on stereo camera input that is able to detect and track multiple pedestrians in real-time. To this end, we propose a hybrid tracking system that combines a computationally cheap low-level tracker with a more complex high-level tracker. The low-level trackers are either based on level-set segmentation or stereo range data together with a point registration algorithm and are employed in order to follow individual pedestrians over time, starting from an initial object detection. In order to cope with drift and to bridge occlusions that cannot be resolved by low-level trackers, the resulting tracklet outputs are fed to a high-level multi-hypothesis tracker, which performs longer-term data association. With this integration we obtain a real-time tracking framework by reducing object detector applications to fewer frames or even to few small image regions when stereo data is available. Reduction of expensive detector evaluations is especially relevant for the deployment on mobile platforms, where real-time performance is crucial and computational resources are notoriously limited.

To overcome another limitation of a classical *tracking-by-detection* pipeline, employment only for tracking of objects for which a pre-trained object classifier is available, we propose a *tracking-before-detection* system that is able to track known and unknown

objects robustly, based purely on stereo information. With this approach we track all visible objects in the scene by first segmenting the point cloud into individual objects and associating them to trajectories based on a simple registration algorithm. The core of our approach is a compact 3D representation that allows us to robustly track a large variety of objects, while building up models of their 3D shape online. In addition to improving tracking performance, this representation allows us to detect anomalous shapes, such as carried items on a person's body. Moreover, classical pedestrian tracking approaches ignore important aspects of human behavior, that should be considered for better scene understanding. Humans are not moving independently, but they closely interact with their surroundings, which includes not only other persons, but also further scene objects. Being able to track not only humans but also their objects, such as child strollers, suitcases, walking aids and bicycles, we propose a probabilistic approach for classifying person-object interactions, which associates objects simultaneously to persons and predicts their interaction type.

In order to demonstrate the capabilities of proposed tracking algorithms, we evaluated them on several challenging video sequences, captured in busy and crowded shopping street environments. As our experiments prove we come closer to the goal of better scene understanding, being able to detect and track multiple objects in the scene in real time and to predict their possible interactions.

Zusammenfassung

In den letzten Jahren hat die Entwicklung von Fahrerassistenzsystemen und mobilen Robotern erhebliche Fortschritte gemacht. Dies wurde möglich durch bemerkenswerte Fortschritte von einzelnen Methoden des maschinellen Sehens wie Objekterkennung, Objektverfolgung, Stereotiefenschätzung und Stereokamera-basierte Odometrie. Das Ziel dieser Methoden beim Einsatz in mobilen Robotern ist es, dem Roboter ein Szenenverständnis zu vermitteln. Möglich wird dies durch das automatische Auswerten von Bildern einer auf dem Roboter montierten Kamera. Objekterkennung und Objektverfolgung sind die für das Szenenverständnis wichtigsten Komponenten, da diese sichere Navigation, Pfadplanung und Kollisionsvermeidung ermöglichen und deshalb zu stark erforschten Gebieten des maschinellen Sehens gehören.

Ein klassisches Verfahren zur Objektverfolgung wird durch den sogenannten *Tracking-by-Detection* Ansatz realisiert. Hierbei wird für jedes Videobild ein Objektdetektor ausgewertet und die resultierenden Objektdetektionen dann mit Hilfe der Odometrie frameübergreifend zu Trajektorien verbunden. Der Nachteil dieses klassischen Ansatzes ist der zwingend notwendige Einsatz eines Objecktdetektors auf jedem Frame. Da dieser Detektor typischerweise die rechenintensivste Komponente der Tracking-Pipeline ist, wird dadurch der Einsatz vom *Tracking-by-Detection* für echtzeitkritische Anwendungen unmöglich. Aus diesem Grund war das erste Ziel der Arbeit die Entwicklung eines Objektverfolgungsverfahrens, welches ausgehend von Bildern einer Stereokamera Fußgänger in Echtzeit finden und verfolgen kann. Dazu haben wir einen hybriden Objektverfolgungsansatz entwickelt, welcher einen recheneffizienten *Low-Level Tracker* und einen *High-Level Tracker* kombiniert. Der *Low-Level Tracker* basiert entweder auf einer *Level-Set* Segmentierung oder Stereotiefe kombiniert mit dem *ICP* Algorithmus. Diese Tracker sind verantwortlich für die Verfolgung von Fußgängern über die Zeit basierend auf einer initialen Objektdetektion. Da die *Low-Level Tracker* nicht mit Abweichungen von der echten Position des Objektes, oft verursacht durch Verdeckungen, umgehen können wird das Verfolgungssystem durch einen *High-Level Tracker* erweitert. Der *High-*

Level Tracker erzeugt lange Trajektorien und erkennt durch entsprechende Konsistenztests die Divergenz der *Low-Level Tracker*. Durch diese Kombination wird die Auswertung eines Detektors auf wenige Frames oder sogar wenige kleine Bildregionen pro Frame reduziert. Diese drastische Reduktion schafft die Voraussetzung für ein echzeitfähiges System, das den Einsatz auf mobilen Robotern erst möglich macht.

Im zweiten Teil der Arbeit stellen wir einen neuen *Tracking-before-Detection* Ansatz vor. Dieser erlaubt es uns, nicht nur bekannte Objektkategorien, wie Fußgänger, sondern auch unbekannte, vorher ungesehene Kategorien zu verfolgen. Mit diesem Ansatz überwinden wir auch die starke Einschränkung von typischen *Tracking-by-Detection* Verfahren, dass ein vortrainierter Objektdetektor erforderlich ist und können somit alle sichtbaren Objekte der Szene verfolgen. Dazu verwenden wir die Punktwolken, die mit Hilfe der Stereoschätzung extrahiert werden. Die Punktwolken werden dabei in individuelle Objekte segmentiert und zu Objekttrajektorien verbunden. Dies geschieht mit Hilfe eines Registrierungsverfahrens, welches zwei Punktwolken auf einander registriert. Den Kern des Verfahrens bildet eine neue, kompakte 3D Objektrepräsentation, die uns auf der einen Seite robuste Verfolgung von beliebigen Objekten erlaubt und auf der anderen Seite das Lernen von 3D-Objektformen online ermöglicht. Die gelernten 3D-Objektformen für Fußgänger erlauben uns die Detektion von getragenen Objekten wie Taschen. Basierend auf der Fähigkeit der Verfolgung von allen Objekten einer Szene wurde in Rahmen dieser Arbeit ein weiterer wichtiger Aspekt der Bewegung von Menschen untersucht. Menschen bewegen sich nicht unabhängig, sondern interagieren sehr stark mit ihrer Umgebung. Diese besteht nicht nur aus anderen Menschen, sondern auch aus weiteren unbekannten Objekten wie Kinderwägen, Koffern, Gehhilfen und Fahrrädern. Um diese Interaktionen modellieren zu können, stellen wir einen neuen probabilistischen Ansatz vor, der uns erlaubt Objekte mit Personen zu assoziieren. Gleichzeitig lässt sich die Art der Interaktion vorhersagen, was wiederum für die Verbesserung der Objektverfolgung verwendet werden kann.

Um die Leistungsfähigkeit der vorgestellten Verfahren zu demonstrieren, haben wir die Algorithmen auf mehreren anspruchsvollen Sequenzen aus sehr belebten Einkaufstraßen evaluiert. Unsere Experimente zeigen, dass wir dem Ziel von einem besseren Szenenverständnis deutlich näher gekommen sind. Wir sind in der Lage Objekte in Echtzeit zu finden, zu verfolgen und ihre Interaktionen vorherzusagen.

Acknowledgments

This dissertation is a product of the invaluable support I received in the last few years from a number of great people.

First and foremost I thank my supervisor Prof. Dr. Bastian Leibe for his continuing intensive support in all stages during my PhD time, for great and inspirational ideas, countless fruitful discussions and teaching me what research is all about. My thanks goes also to Prof. Dr. Luc Van Gool for his interest in my work and agreeing to co-examine my thesis.

I would like to express my gratitude to my diploma thesis supervisor Prof. Dr. Daniel Cremers who intrigued my interest in Computer Vision from the first lecture.

I am deeply grateful to all my colleagues Esther Horbert, Tobias Weyand, Patrick Sudowe, Georgios Floros and Wolfgang Mehner making the life in UMIC really fun. Furthermore, Tobias thank you for offering me a place to sleep during the nights close to the deadlines, that become longer and longer the closer the deadline was approaching and especially thank you for the excellent espresso in the morning. Georgios thank you for all the great Greek supplies as the amazing feta and spinach pies or the excellent olive oil from Crete. Esther thank you very much for being supportive during the difficult time in the first year of my PhD. Patrick thank you for all the discussions about cars in context of car detection in video sequences, but also convincing me why Volkswagen Golf GT is the best car ever. Moreover, I would like to thank the team of KHAO-LAK-BEACH being always punctual in delivering the best Vietnamese food during rough deadline time.

Thanks also to my diploma/master theses students who have contributed to this theses with their work: Esther Horbert, Tobias Baumgartner, Philipp Fischer, Stefan Breuers, Wolfgang Mehner, Emmanouil Tzouridis, Seyed Hamidreza Odabai-Fard, Jonathan Meyer.

Furthermore, I would like to thank Tanja, my two years old daughter, for coaching me to get along with only 3-4 hours sleep per night, which helped me to stay awake and concentrated during long nights before the deadlines.

Last but not least, I would like to thank my family and my friends, especially my parents always believing in me and being always there.

Contents

1 Introduction

1.1. Motivation

Computer Vision is a broad and varied research field concerned with the problem of extracting semantic information from the images of a scene. Having its beginning in the early 1970s, it has been a very vivid research area. Scientists around the world put enormous effort into the development of algorithms and methods trying to tackle the problem of extracting relevant information from existing images. As a result, the techniques of this field have applications in a wide range of scenarios, including manufacturing, security, robotics, car industry, communication and many more. For many applications the behavior of humans in urban scenarios is of particular interest. For example, a traffic safety application could analyze a video stream from a camera system, mounted inside a car or on a mobile robot, in order to issue warnings in case of future path intersections or possible collisions. In order to achieve this goal, methods are required that can process video streams automatically and in real-time. Furthermore, in order to understand the behavior of people, it is also important to recognize and track other objects in their surroundings. In practical scenarios, this includes a large variety of objects such as bicycles, child strollers, shopping carts, trolleys, or wheelchairs. In recent years a number of tracking-by-detection approaches have been proposed to address these goals, reaching remarkable performance for robust people detection and tracking in dynamic and complex real-world scenes. However, those approaches have two major limitations. On the one side they are not yet satisfactory for use on autonomous platforms with respect to their requirements for computational power and energy consumption. On the other side they are naturally restricted to tracking objects for which pre-trained detector models (e.g., pedestrians) are available.

In this thesis, we investigate the problem of multi-object tracking in busy inner-city scenarios. Starting with classical tracking-by-detection approaches, we focus on algorithmic means for improving run-time efficiency in order to make them applicable for use on a mobile robot. Based on the lessons learned from this endeavor, we investigate different

means for reducing the dependency on an expensive object detector by introducing a hybrid tracking framework. Such a framework is a combination of a computationally cheap low-level tracker with a high-level tracker. The low-level tracker follows pedestrians over time after an initial detection and thus takes over the role of the computationally expensive object detector. We investigate different choices for the low-level tracker, exploring both appearance-based and depth-based approaches. In both cases the low-level tracker is augmented by a high-level tracker that, using the tracklets output by the low-level tracker, performs longer-term data association bridging drift and occlusions that cannot be resolved by the low-level trackers. In the second part of the thesis, we then spin this idea further. Assuming that a mechanism exists for extracting regions-of-interest from the input video data (in our case from stereo data), we explore how those ROIs can be used for both simplifying and improving the object detection and tracking stages. Finally, we address the problem of tracking both known and unknown scene objects, which is a prerequisite for robust performance in many real-world settings such as mobile robotics and intelligent vehicles. For this, we extend the ROI-based scheme to a true tracking-before-detection approach, which can automatically track a large number of object candidates even before knowing their categories. This paradigm shift has important consequences for the design of the tracking pipeline. In particular, before tracking we first need to decide how an object candidate that we want to track is defined. To this end, we make use of regions-of-interest which are robustly segmented into candidate objects. Each such region is then tracked independently in 3D using a model-based point cloud registration tracker. In order to learn a representation of the objects, we develop an approach that reconstructs 3D shape models of each tracked object, which allow us besides robustly tracking a large variety of objects, an analysis of their shape. In addition, relying on the tracking results of known and unknown objects, we analyze person-object interactions and use this knowledge to make improved predictions for the continuation of observed trajectories. In this sense, we believe that the contributions of this thesis have brought tracking a significant step towards the next level.

1.2. Contributions

In detail we have made the following contributions:

- We show how the classical *tracking-by-detection* framework can be complemented by a cheap and fast low-level tracker, based either on appearance or depth, resulting in a real-time tracking system. We systematically present the required consistency checks and interactions between the components in order to solve the difficulties in street-level mobile tracking tasks with a number of non-trivial challenges.

- We present an integrated system for upper body detection which is purely based on depth information. The system overcomes the drawback of classical full-body detectors, which often fail to detect pedestrians close to the camera, due to strong