Doctoral Thesis

Detection of Structured Objects with a Range Camera

Author(s):
Gächter Toya, Stephan M.B.

Publication Date:
2008

Permanent Link:

https://doi.org/10.3929/ethz-a-005679968 →

Diss. ETH No. 17825

Detection of Structured Objects with a Range Camera

A dissertation submitted to the
Eidgenössische Technische Hochschule Zürich (ETHZ)

for the degree of

Doctor of Sciences

presented by

Stefan Martin Benjamin Gächter Toya

Electrical Engineer,
École Polytechnique Fédérale de Lausanne (EPFL)
born February 12, 1973
citizen of Switzerland and Canada

accepted on the recommendation of

Prof. Dr. Roland Siegwart, ETH Zürich, Principal Advisor
Prof. Dr. Markus Vincze, TU Wien, Member of the Jury
Dr. Nicola Tomatis, BlueBotics SA, Lausanne, Member of the Jury

2008

Abstract

In recent years, a novel type of range camera emerged on the market to capture 3D scenes for mobile robotic applications. A current issue in mobile robotics is to endow a robot with a human-compatible representation of the environment, which demands object classification. However, mobile robots have to act in complex surroundings. Therefore, a classification algorithm has to deal with uncertainties and incompleteness originating from the fact that the number and location of the query objects are unknown beforehand, which is aggravated in addition by sensory limitations as well as variations of the object aspects due to robot motion and human intervention. Thus, the main goal and contribution of this work is to bring well grounded approaches from different domains together, extend and adapt them to enable a novel object detection with a range camera in 3D mobile robotic applications.

In order to address this problem, three steps are proposed. Firstly, objects are modeled by primitive parts, secondly, primitive parts are tracked while detected in a sequence of images and, thirdly, objects are detected by votes from the primitive parts; whereas primitive part models and object structural model are learned by a reasonable large training set. In more detail, in the *first* step, the range data is sequentially acquired, quantized, and merged, over a limited time-frame using odometry information of a robot for the spatial range data alignment, into a discrete volume representation of voxels. The resulting sequence of voxel images is further processed to emphasize the local structure information in form of a set of shape factors, which characterizes the linear, planar, or spherical likeliness of the voxel distribution in a local neighborhood. In the *second* step, a multiple-hypothesis tracking algorithm known as particle filter is proposed to detect primitive parts in the sequence of voxel images. The multiple hypotheses are used to represent hypothetical primitive parts of unknown number and locations in the field-of-view of the robot. In order to enable the hypothesis verification, and therewith the primitive part detection, a part classifier is designed and trained with support vector machine (SVM) on a large set of labeled primitive parts, where a primitive part is modeled by a part descriptor. The part descriptor consists of a shape-factor histogram to describe the shape and bounding-box to circumscribe the dimension of a primitive part. In the *third* step, the detected primitive parts undergo a localization and structure verification process to assign parts to their potentially originating objects and to disambiguate parts belonging to objects from parts being clutter. Therefore, a structural object model in form of voting vectors, known as implicit shape model

(ISM), is learned to enable object localization by casting votes from each detected primitive part to potential object reference points and searching for maxima in the voting space. Finally, a simple structure verification identifies the actual object parts.

The appropriateness of the algorithm is successfully demonstrated by detection of chairs in real world experiments, where the robot is navigating through an indoor scenario composed of a dining table, two chairs, and a coffee table. The detection approach is demonstrated based on two different primitive parts classifiers, one for stick- and one for plate-like structures, and of a structural chair model learned with a training set of twelve four-legged chairs. The outcome of the experiments show that chairs are correctly detected in sequences of up to 550 range images, despite occlusion and clutter.

Kurzfassung

Eine neuartige Distanzbildkamera zur dreidimensionalen Erfassung der Umgebung kam in den letzten Jahren auf den Markt. Diese Kamera eignet sich ebenfalls für Anwendungen in der mobilen Robotik. Ein aktuelles Thema in der mobilen Robotik ist die Erforschung einer menschengerechten Darstellungsweise der Umgebung. Eine solche Darstellungsweise ist auf Objektklassifizierung angewiesen. Da mobile Roboter jedoch in einer komplexen Umgebung agieren, muss ein Algorithmus zur Objektklassifizierung fähig sein, mit unsicheren und unvollständigen Daten umgehen zu können. Die Ursachen von Unsicherheit und Unvollständigkeit sind die nicht im vornherein bekannte Anzahl und Positionen der zu klassifizierenden Objekte, was zusätzlich durch Beschränkungen des Sensors sowie Änderungen in der Objektansicht, die auf Bewegungen des Roboters oder durch den Menschen zurückzuführen sind, erschwert wird. Das Ziel und der Beitrag dieser Arbeit ist deshalb das Zusammenführen von fundierten Methoden aus verschiedenen Gebieten, deren Erweiterung und Anpassung, so dass eine neuartige Objektdetektion mit der Distanzbildkamera für Robotikanwendungen in 3D entsteht.

Um das Problem der Objektdetektierung anzugehen, werden drei Schritte vorgeschlagen. Erstens, Objekte werden mittels einfacher Einzelteile modelliert. Zweitens, die Einzelteile werden in einer Bildfolge detektiert und verfolgt und, drittens, die Objekte werden durch die kumulative Stimmabgabe der Einzelteile detektiert; wobei die Modelle der Einzelteile sowie der Objektstruktur anhand einer angemessen grossen Menge an Beispielen gelernt werden. Ausführlicher, in dem *ersten* Schritt werden die Distanzdaten sequentiell aufgenommen, quantisiert und in einer diskreten Raumdarstellung bestehend aus Voxels vereinigt, wobei dies während eines begrenzten Zeitabschnitts geschieht und dabei die Odometrieinformation des Roboters zur Bildausrichtung verwendet wird. In dem *zweiten* Schritt wird ein Mehrfach-Hypothesen-Verfolgungsalgorithmus, bekannt als Partikelfilter, vorgeschlagen, um die Einzelteile in der Voxelbildfolge zu detektieren. Die Hypothesen stellen mögliche Einzelteile im Sehbereich des Roboters dar, deren Anzahl und Positionen nicht bekannt sind. Um die Hypothesenverifikation zu ermöglichen, und damit die Detektion der Einzelteile, wird ein Klassifikator entworfen und mittels einer Supportvektor-Maschine (SVM) anhand einer grossen Menge an Beispieleinzelteilen trainiert, wobei die Einzelteile durch einen Deskriptor beschrieben werden. Der Deskriptor besteht aus einem Formfaktorhistogramm, das die Form des Einzelteils beschreibt, und einem rechteckigen Volumen, das die Dimensionen umschreibt. In dem *dritten*

Schritt werden die detektierten Einzelteile einem Lokalisierungs- und Strukturverifikationsprozess unterzogen, um die Einzelteile möglichen Objekten, deren sie entspringen, zuzuweisen und zwischen Einzelteilen, die zu tatsächlichen Objekten gehören, von anderen, störenden Einzelteilen zu unterscheiden. Dafür wird ein Strukturmodell eines Objekts, bekannt als implizites Strukturmodell (ISM), in Form von Referenzvektoren gelernt, das die Objektlokalisierung mittels kumulativer Stimmabgabe ermöglicht, indem die Referenzvektoren von jedem detektierten Einzelteil ausgesandt werden um auf mögliche Objekte hinzuweisen. Zuletzt wird eine einfache Strukturüberprüfung vorgenommen, die die Einzelteile, die zu einem tatsächlichen Objekt gehören, identifiziert.

Die Eignung des Algorithmus wird erfolgreich aufgezeigt, indem Stühle in einem realen Experiment detektiert werden. Dabei fährt der Roboter durch ein Innenraumszenario das aus einem Esstisch, zwei Stühlen, und einem Kaffeetisch besteht. Die Detektierungsmethode wird anhand zweier verschiedener Klassifikatoren, einer für stabähnliche und einer für plattenähnliche Strukturen, und einem Strukturmodell eines Stuhls aufgezeigt, wobei eine Gruppe von zwölf Beispielstühlen zum Lernen verwendet wird. Das Resultat der Experimente zeigt, dass Stühle in einer Tiefenbildfolge von bis zu 550 Bildern richtig detektiert werden, trotz störenden Elementen und Verdeckungen.