High-Resolution 3D Layout from a Single View

Author(s):
Zia, Muhammad Z.

Publication date:
2014

Permanent link:
https://doi.org/10.3929/ethz-a-010150291

Originally published in:
Mitteilungen / Institut für Geodäsie und Photogrammetrie an der Eidgenössischen Technischen Hochschule ETH Zürich 114

DISS. ETH NO. 21923

High-Resolution 3D Layout from a Single View

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

MUHAMMAD ZEESHAN ZIA

Master of Science (M.Sc.), Technische Universität München

born on 26.07.1984

citizen of Pakistan

accepted on the recommendation of

Prof. Dr. Konrad Schindler, ETH Zurich
Prof. Dr. Tinne Tuytelaars, KU Leuven

2014

Abstract

Scene understanding based on photographic images has been the holy grail of computer vision ever since the field came into existence some 50 years ago. Since computer vision comes from an Artificial Intelligence background, it is no surprise that most early efforts were directed at fine-grained interpretation of the underlying scene from image data. Unfortunately, the attempts proved far ahead of their time and were unsuccessful in tackling real-world noise and clutter, due to unavailability of vital building blocks that came into existence only decades later as well as severely limited computational resources.

In this thesis, we consider the problem of detailed 3D scene level reasoning from a single view image in the light of modern developments in vision and adjoining fields. Bottom-up scene understanding relies on object detections, but unfortunately the hypotheses provided by most current object models are in the form of coarse 2D or 3D bounding boxes, which provide very little geometric information - not enough to model fine-grained interactions between object instances. On the other hand, a number of detailed 3D representations of object geometry were proposed in the early days of computer vision, which provided rich description of the modeled objects. At the time, they proved difficult to match robustly to real world images. However over the past decade or so, developments in local image descriptors, discriminative classification, and numerical optimization methods have made it possible to revive such approaches for 3D reasoning and apply them to challenging real-world images. Thus we revisit detailed 3D representations for object classes, and apply them to the task of scene-level reasoning. The motivation also comes from recent revival of coarse grained 3D modeling for scene understanding, and demonstrations of its effectiveness for 3D interpretation as well as 2D recognition. These successes raise the question of whether finer-grained 3D modeling could further aid scene-level understanding, which we try to answer in our work.

We start with 3D CAD training data to learn detailed 3D object class representations, which can estimate 3D object geometry from a single image. We demonstrate applying this representation for accurate estimation of object shape, as well as for novel applications namely, ultra-wide baseline matching and fine-grained object categorization. Next, we add an occluder representation comprising of a set of occluder masks, which enables the detailed 3D object model to be applied to occluded object instances, demonstrated over a dataset with severely occluded objects. This object representation is lifted to metric 3D space, and we jointly model multiple object instances in a common frame. Object interactions are modeled at the high-resolution of 3D wireframe vertices: deterministically modeling object-object occlusions and long-range dependencies enforcing all objects to lie on a common ground plane, both of which stabilize 3D estimation. Here, we demonstrate precise metric 3D reconstruction of scene layout on a challenging street scenes dataset. We evaluate parts of our approach on five different datasets in total, and demonstrate superior performance to state-of-the-art over different measures of detection quality. Overall, the results support that detailed 3D reasoning benefits both at the level of individual objects, and at the level of entire scenes.

Zusammenfassung

Seit sich die *computer vision* vor ca. 50 Jahren als eigenständiges Feld etabliert hat ist das Szenenverstehen, also die semantische Interpretation der abgebildeten Szene, eines ihrer fundamentalen Probleme. Da der Ursprung der *computer vision* in der künstlichen Intelligenz liegt überrascht es nicht, dass zu ihren Zielen von Beginn an das automatische Verstehen der beobachteten Szene gehörte. Aus heutiger Sicht ist es verwundert es auch nicht, dass die anfänglichen Versuche scheiterten, einerseits weil wesentliche Grundlagen erst Jahrzehnte später entwickelt wurden, andererseits weil die damaligen Computer nicht die notwendige Rechenleistung hatten.

Die vorliegende Arbeit untersucht das Problem des detaillierten, 3-dimensionalen Szenenverstehens auf Basis eines Einzelbildes, ausgehend von den heutigen Möglichkeiten der *computer vision* und verwandter Disziplinen. Ein grundlegender Baustein des Szenenverstehens ist die Erkennung von Objekten im Bild. Die gebräuchlichen Detektoren liefern jedoch als Objektmodell nur 2D oder 3D *bounding boxes*, und diese grobe Repräsentation ist nicht geeignet, die Objektgeometrie und die Interaktionen zwischen verschiedenen Objekten im Detail zu modellieren. In Gegensatz dazu wurden in der Frühzeit der *computer vision* Repräsentationen der Objektgeometrie entwickelt, die eine wesentlich höheren Detailgrad aufweisen. Es gelang damals aber nicht zuverlässig, das Modell mit dem Bildinhalt in Korrespondenz zu bringen. Die Entwicklungen der letzten Jahre im Bereich der lokalen Bild-Deskriptoren, der diskriminativen Klassifikation und der numerischen Optimierung ermöglichen es, diese Ansätze wiederzubeleben und auf das Verstehen komplexer 3-dimensionaler Szene anzuwenden. In der vorliegenden Arbeit wird daher eine solches klassisches, detailreiches 3D Objektmodell für das bildbasierte Szenenverstehen benutzt. Der vorgestellte Ansatz ist unter anderem dadurch motiviert, dass in den letzten Jahren das 3-dimensionale Szenenverstehen – mit eher groben Modellen – wieder vermehrt untersucht wurde. Dabei zeigte sich ,dass es sowohl für die 3D Modellierung als auch für die Objekterkennung im Bild Vorteile bringt. Diese Erfolge werfen die Frage auf, ob detailliertere Modelle das Szenenverstehen weiter verbessern können. Die vorliegende Arbeit ist ein Versuch, die Frage zu beantworten.

Den Ausgangspunkt der Arbeit bilden 3D CAD-Modelle. Auf deren Basis werden detaillierte, deformierbare Objektrepräsentationen gelernt, mit deren Hilfe die 3D Geometrie des Objekts auf Basis eines Einzelbildes geschätzt werden kann. Neben der Rekonstruktion der genauen geometrischen Objektform ermöglichen solche Modelle auch neue Anwendungen wie das *matching* über extrem grosse Basislinien und die Klassifizierung in nur durch geometrische Details unterscheidbare Unterkategorien. Um Verdeckungen in den Bildern verarbeiten zu können wird das Modell um eine Verdeckungsmaske erweitert. Die Maske ermöglicht es, die Verdeckung einzelner Objektteile darzustellen, und es wird gezeigt, dass sich damit auch stark verdeckte Objektinstanzen detektieren lassen. Schliesslich wird das Modell noch so modifiziert, dass Objekte im metrischen 3D Koordinaten repräsentiert werden. Somit können mehrere Objekte in einem gemeinsamen Koordinatensystem modelliert werden. Weiters werden Interaktionen zwischen den verschiedenen Objekten auf dem Niveau einzelner Objektpunkte und -flächen berücksichtigt,

im speziellen gegenseitige Verdeckungen und eine gemeinsame Geländeebene, auf der alle Objekte stehen. Es wird gezeigt, dass mit einem derart stabilisierten Modell komplexe Strassenszenen metrisch korrekt rekonstruiert werden können. Die einzelnen Teile der vorgeschlagenen Methode wurden auf mehreren verschiedenen Datensätzen evaluiert, dabei wurden signifikante Verbesserungen hinsichtlicher verschiedener Qualitätsmasse beobachtet. Insgesamt stützen die Ergebnisse die Hypothese, dass die detaillierte 3-dimensionale Modellierung vorteilhaft für das Bildverstehen ist, sowohl auf der Stufe einzelner Objekte als auch auf der Stufe kompletter Szenen.