Interleaved object categorization and segmentation

Author(s):
Leibe, Bastian

Publication Date:
2004

Permanent Link:

https://doi.org/10.3929/ethz-a-004949680 →

DISS. ETH NO. 15752

Interleaved Object Categorization and Segmentation

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY ZURICH

for the degree of
Doctor of Technical Sciences

presented by

BASTIAN LEIBE

Dipl. Inform., M. Sc.

born 23rd of April, 1975
citizen of
Germany

accepted on the recommendation of

Prof. Dr. Bernt Schiele, examiner
Prof. Dr. Andrew Zisserman, co-examiner

2004

Abstract

This thesis is concerned with the problem of visual object categorization, that is of recognizing unseen-before objects, localizing them in cluttered real-world images, and assigning the correct category label. This capability is one of the core competencies of the human visual system. Yet, computer vision systems are still far from reaching a comparable level of performance. Moreover, computer vision research has in the past mainly focused on the simpler and more specific problem of identifying known objects under novel viewing conditions.

The visual categorization problem is closely linked to the task of figure-ground segmentation, that is of dividing the image into an object and a non-object part. Historically, figure-ground segmentation has often been seen as an important and even necessary preprocessing step for object recognition. However, purely bottom-up approaches have so far been unable to yield segmentations of sufficient quality, so that most current recognition approaches have been designed to work independently from segmentation.

In contrast, this thesis considers object categorization and figure-ground segmentation as two interleaved processes that closely collaborate towards a common goal. The core part of our work is a probabilistic formulation which integrates both capabilities into a common framework. As shown in our experiments, the tight coupling between those two processes allows them to profit from each other and improve their individual performances. The resulting approach can detect categorical objects in novel images and automatically compute a segmentation for them. This segmentation is then used to again improve recognition by allowing the system to focus its effort on object pixels and discard misleading influences from the background.

In addition to improving the recognition performance for individual hypotheses, the top-down segmentation also allows to determine exactly from where a hypothesis draws its support. We use this information to design a hypothesis verification stage based on the MDL principle that resolves ambiguities between overlapping hypotheses on a per-pixel level and factors out the effects of partial occlusion. Altogether, this procedure constitutes a novel mechanism in object detection that allows to analyze scenes containing multiple objects in a principled manner. Our results show that it presents an improvement over conventional criteria based on bounding box overlap and permits more accurate acceptance decisions.

Our approach is based on a highly flexible implicit representation for object shape that can combine the information of local parts observed on different training examples and interpolate between the corresponding objects. As a result, the proposed method can learn object models already from few training examples and achieve competitive object detection performance with training sets that are between one and two orders of magnitude smaller than those used in comparable systems. An extensive evaluation on several large data sets shows that the system is applicable to many different object categories, including both rigid and articulated objects.

Zusammenfassung

Diese Arbeit beschäftigt sich mit der visuellen Objektkategorisierung, d.h. dem Problem, zuvor noch nie gesehene Objekte zu erkennen, in realen Szenen zu lokalisieren, und die Objekte der korrekten Kategorie zuzuweisen. Diese Fähigkeit ist eine der Kernkompetenzen des menschlichen Sehsystems. Die maschinelle Bildverarbeitung is jedoch noch weit davon entfernt, eine vergleichbare Leistung erbringen zu können. Darüberhinaus hat sich die Forschung in der Vergangenheit hauptsächlich auf das einfachere und speziellere Problem konzentriert, bekannte Objekte unter geänderten Bedingungen wiederzuerkennen.

Die visuelle Kategorisierung ist eng mit dem Figure-Ground Segmentierungsproblem verbunden, d.h. mit der Aufgabe, das Bild in eine Objekt- und eine Hintergrund-Region zu trennen. In der Vergangenheit wurde dieses Problem oft als ein wichtiger und sogar notwendiger Vorverarbeitungsschritt für die Objekterkennung betrachtet. Reine Bottom-up Verfahren haben sich aber bis heute als ungeeignet erwiesen, Segmentierungen von genügender Qualität hervorzubringen, so dass die meisten aktuellen Erkennungsansätze dahingehend entworfen wurden, ohne eine vorausgehende Segmentierung auszukommen.

Im Gegensatz dazu betrachtet diese Arbeit die Objektkategorisierung und Figure-Ground Segmentierung als zwei miteinander verwobene Prozesse, die eng zusammenarbeiten, um ein gemeinsames Ziel zu erreichen. Das Herzstück unseres Ansatzes ist eine probabilistische Formulierung, die beide Fähigkeiten in einem gemeinsamen Rahmen verbindet. Wie wir in unseren Experimenten zeigen, erlaubt die enge Zusammenarbeit dieser beiden Prozesse ihnen, voneinander zu profitieren und ihre Einzelleistungen zu verbessern. Der daraus entstehende Ansatz ermöglicht es, Kategorie-Objekte in neuen Bildern zu detektieren und automatisch eine Segmentierung für sie zu berechnen. Diese Segmentierung trägt dann dazu bei, die Erkennungsergebnisse wiederum zu verbessern, indem sie es dem System ermöglicht, sich auf Objektpixel zu konzentrieren und irreführende Einflüsse von Hintergrundstrukturen zu ignorieren.

Zusätzlich zu der besseren Erkennungsleistung für Einzelhypothesen erlaubt die so berechnete Top-Down Segmentierung es unserem Ansatz ebenfalls, zu ermitteln welche Bildstrukturen eine Hypothese stützen und somit für ihr Zustandekommen verantwortlich sind. Wir verwenden diese Information, um eine auf dem MDL-Prinzip beruhende Verifikationsstufe zu entwerfen, die Konflikte zwischen überlappenden Hypothesen Pixel für Pixel auflöst und somit die Folgen partieller Verdeckungen ausklammert. Insgesamt stellt dieses Verfahren einen neuartigen Mechanismus dar, der es erlaubt, Szenen mit mehreren Objekten auf eine fundierte Weise zu untersuchen. Unsere Ergebnisse zeigen, dass dieser Mechanismus eine Verbesserung gegenüber herkömmlichen Kriterien basierend etwa auf dem Bounding-Box Überlappungsgrad darstellt und dass er genauere Akzeptanzentscheidungen ermöglicht.

Unser Ansatz basiert auf einer sehr flexiblen impliziten Darstellung der möglichen Objektformen, die es gestattet, die Informationen von Objektteilen aus unterschied-

lichen Trainingsbeispielen zu kombinieren und zwischen den entsprechenden Objekten zu interpolieren. Als Folge davon kann das vorgeschlagene Verfahren Objektmodelle schon aus sehr wenigen Trainingsbeispielen lernen und gute Detektionsleistungen schon mit Trainingsdatenmengen erzielen die ein bis zwei Grössenordnungen unter denen vergleichbarer Systeme liegen. Eine ausführliche Auswertung auf mehreren grossen Bildersammlungen zeigt, dass das vorgestellte System auf viele verschiedene Objektkategorien, mit sowohl starren als auch artikulierten Objekten, anwendbar ist.