Scalable Visual Recognition Using a Shared Vocabulary

Author(s):
Razavi, Nima

Publication Date:
2012

Permanent Link:

https://doi.org/10.3929/ethz-a-007593593 →

DISS. ETH NO. 20721

Scalable Visual Recognition using a Shared Vocabulary

A dissertation submitted to

ETH ZURICH

for the degree of

Doctor of Sciences

presented by

Nima Razavi
M.Sc. ETH Zürich
born 2. September 1984
citizen of Iran

accepted on the recommendation of

Prof. Dr. Luc Van Gool, ETH Zurich and KU Leuven, examiner
Prof. Dr. Bernt Schiele, MPI and Saarland University, co-examiner
Dr. Jürgen Gall, MPI for Intelligent Systems, Tübingen, co-examiner

2012

Abstract

Building autonomous systems that enable machines to understand and interpret images is a long-standing goal in the field of computer vision. Solving this problem will have a huge impact on the society as it leads to many life-changing applications such as autonomous driving, image search, and surveillance just to name a few. Visual object detection is a key component of such systems that makes sense of an input image by establishing a correspondence of its content to previously observed data and compactly represents this information in forms of labels.

The primary goal of this dissertation is to present a scalable non-parametric visual object detector with a shared vocabulary. Scalability of object detectors with respect to the number of classes and training images is a critically important issue for applications with a large number of classes. Detecting objects with a shared vocabulary is also very attractive as it facilitates better generalization and incremental learning of the detector.

This thesis shows how to learn and utilize a shared vocabulary to detect objects accurately and in a scalable manner. To this end, it argues that the vocabulary entries need to be generalizable across instances and yet remain as discriminative as possible. This way, by matching a patch to the vocabulary, a sparse and reliable distribution of object labels and configurations can be predicted. It is proposed to learn the vocabulary by simultaneously maximizing the sparsity of the entries and controlling their complexity to ensure generalization. Also, a data-driven semantic hierarchy acting on the patch level is introduced and used to speed up the detection.

In addition to the category label and location, it is desired in many applications to retrieve auxiliary information for a detected instance such as type and pose. For this purpose, this thesis proposes to describe a detection by the configuration and appearance of its supporting patches.

Using this description, a novel occlusion-insensitive similarity measure between two detections is introduced. The proposed similarity measure is used to retrieve similar previously seen objects; transferring their information to the new detection. The support of a hypothesis is also used for tracking where the detector is adapted to the appearance of an instance.

When detecting objects in an image, it is important to only combine consistent patches into a hypothesis. Semantic properties such as pose or color can be used for detecting consistent hypotheses. However, these properties are often either unobserved in the training data or not clearly known to be optimal for enforcing consistency. This thesis proposes to treat these additional properties as global and local latent variables and discriminatively learn their optimal assignments for the training data. This way, only patches consistent in their latent assignments are combined into an object hypothesis, substantially increasing the detection accuracy.

All methods presented in this thesis are evaluated on challenging and realistic benchmark databases. The experiments confirm the benefits of using a discriminative shared vocabulary of middle complexity patches as a building block for visual recognition tasks such as scalable multi-class object detection, tracking, and object property estimation.

Zusammenfassung

Autonome Systeme zu bauen, die Bilder verstehen und interpretieren, ist seit Langem Ziel des Forschungsgebiets Computer Vision. Die Lösung dieses Problems verspricht viele Anwendungen zu ermöglichen, die unser Leben beeinflussen: von selbstfahrenden Autos zu Bildersuche und Überwachung, um nur ein paar Beispiele zu nennen. Visuelle Objekterkennung ist eine Schlüsselkomponente solcher Systeme und interpretiert ein Eingabebild mittels vorher annotierter Bilddaten. Bildelemente, die in einem Eingabebild wiedererkannt wurden, können so auf kompakte Weise durch Annotationen dargestellt werden.

Das Hauptziel dieser Dissertation ist eine skalierbare und nichtparametrische, visuelle Objekterkennung basiert auf ein gemeinsame Vokabular zu präsentieren. Skalierbarkeit in der Anzahl der erkennbaren Klassen und der Grösse des Trainingsdatensatzes ist insbesondere für Anwendungen mit vielen zu erkennenden Klassen wichtig. Das gemeinsame Vokabular vereinfacht es, die Objekterkennung zu generalisieren und das System inkrementell zu erweitern.

Diese Dissertation beschreibt, wie ein Vokabular erlernt und benutzt werden kann, um Objekte zielsicher und auf skalierbare Art und Weise zu erkennen. Ein Kernpunkt ist, dass die einzelne Vokabel generalisierbar sein müssen, um auf weitere Objektinstanzen angewendet werden zu können, aber dennoch so spezifisch wie möglich sein müssen. Eine dünn besetzte und zuverlässige Verteilung von Objektannotationen und Konfigurationen kann durch eine Paarung von Patches zu Vokabeln vorhergesagt werden. Erlernung des Vokabular kann durch Maximieren der Dünnbesetztheit, unter gleichzeitiger Kontrolle der Komplexität von Einträgen zum Zwecke der Generalisierbarkeit, stattfinden. Weiterhin wird eine datengetriebene semantische Hierarchie, die auf Patchebene arbeitet, eingesetzt, um die Erkennung zu beschleunigen.

Fuer viele Anwendunge ist es wünschenswert neben Annotation und Lage eines Objektes weitere unterstützende Informationen wie Typ und Pose zu extrahieren. Hierzu beschreibt diese Dissertation eine Erkennung mittels Anordnung und Aussehen der unterstützenden Patches. Mit dieser Beschreibung ist es möglich ein gegenüber Verdeckung robustes Ähnlichkeitsmass zwischen zwei Erkennungen zu definieren. Dieses Ähnlicheitsmass kann dann genutzt werden, um ähnliche Objekte, die bereits vorher erkannt wurden, zu finden, und deren Informationen auf die neue Erkennung zu transferieren. Die Unterstützung einer Hypothese wird auch für Tracking benutzt, wenn die Erkennung auf ein bestimmtes Aussehen einer Instanz angepasst wurde.

Für Objekterkennung in Bildern ist es wichtig nur konsistente Patches zu einer Hypothese zu kombinieren. Semantische Eigenschaften wie Pose und Farbe können benutzt werden, um konsistente Hypothesen zu erkennen. Diese Eigenschaften sind jedoch oftmals nicht in den Trainingsdaten enthalten oder tragen nicht zur Formung von konsistenten Hypothesen bei. In dieser Dissertation werden diese weiterführenden Eigenschaften als, globale und lokale, latente Variablen behandelt. Die optimale Zuordnung für den Trainingsdatensatz kann auf diese Weise individuell erlernt werden. Dadurch ist es möglich, nur konsistente Patches in ihren latenten Zuordnungen zu einer Hypothese zu kombinieren und damit die Erkennungsgenauigkeit wesentlich zu erhöhen.

Alle Methoden, die in dieser Dissertation präsentiert werden, wurden auf anspruchsvollen und realistischen Benchmark-Datenbanken getestet. Diese Experimente unterstuetzen die Benutzung eines unterscheidenden gemeinsame Vokabular von mittelkomplizierten Patches als Baustein für Anwendungen visueller Objekterkennung wie zum Beispiel skalierbare Mehrklassen-Objekterkennung, Tracking und Schätzung von Objekteigenschaften.