Objects in Relation for Scene Understanding

Diss. ETH No. 23548

Objects in Relation for Scene Understanding

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

Marian Nasr Amin George

BSc., MSc. Computer and Systems Engineering

born on 12 June 1986

citizen of Egypt

accepted on the recommendation of

Prof. Dr. Friedemann Mattern, examiner

Prof. Dr. Marc Pollefeys, co-examiner

Prof. Dr. Richard Green, co-examiner

Dr. Christian Floerkemeier, co-examiner

2016

Abstract

The goal of visual image understanding is to have machines which can perceive the world similar to a human. Achieving this goal will provide numerous opportunities for machines to seamlessly interact with users, improving the quality of life of individuals. As the need for understanding a large number of scene classes with thousands of objects grows, there is a gradual shift towards a more fine-grained understanding of scenes. Thus, image understanding algorithms are now faced with the need to be able to scale to an increasing number of scenes and objects, and to better discriminate between fine-grained scenes. Furthermore, we would like to achieve these goals in a robust and generalizable manner, in such a way that the developed algorithms are effective in understanding scene images taken with smartphones, digital cameras, security cameras, or any other means without any further modification or adjustment.

Traditional methods rely on visual appearance information to understand scenes. In contrast, to achieve the desired detailed, yet generalizable, level of scene understanding, we need to explore high-level semantic information of scenes. Such information will enable machines to perceive relationships, co-occurrences, and informativeness of the different components of a scene image in a similar manner to a person. Among the different components of scenes, objects provide the richest semantic entity of a scene image; they provide hints about the type of the scene, its location, and how closely it relates to other scenes. Furthermore, objects enable algorithms to reason about the semantic relationships among the different components of a real world scene. In this thesis, we propose techniques for a fine-grained level of scene understanding through exploiting high-level contextual scene knowledge. We show how to jointly exploit the visual appearance and context of objects in scenes; how to explore the underlying semantic space of related fine-grained scenes; and how to recognize a wide range of objects in scenes and exploit

global scene context.

In many real-world applications, like assistive vision or robotics, a visual recognition system is faced with the challenge that there is a significant mismatch between the distribution of the training data and the test data where the system will be applied. An even more challenging scenario happens when no data is available from the test domain during the training process. As for scenes, we argue that describing a scene image in terms of its constituent objects provides an effective approach to tackle this challenge, where objects provide a high level of abstraction which enhances the generalization ability of the representation. This is especially true if there are no available scene images during the training process, but only images of the fine-grained objects that may occur in them. We propose to describe a scene image by retrieving all its constituent fine-grained objects in a multi-label image classification scheme. We jointly reason about the visual appearance of objects, their co-occurrence statistics, and the amount of expected overlap among the retrieved objects in a given scene image. This is achieved by optimizing an energy function which incorporates the three criteria to reach a final labeling of a given scene image. Results show the effectiveness and efficiency of our approach in simultaneously retrieving all the specific objects in a given scene image in a single optimization step.

While objects provide a powerful notion for describing scenes, some fine-grained scenes may share common objects which imposes challenges on the ability to differentiate between them. In several fine-grained scene domains, e.g. the domain of store scenes, there exists subgroups of scene images that are more related to each other than to other scene images, for example by sharing more common objects with each other. Automatically discovering these more confusing groupings allows the system to learn more discriminant models for each subgroup that yield a better consensus decision when combined. We propose an approach to describe scene images using conditional scene probabilities, where each image is represented by how likely it belongs to each scene class conditioned on its constituent objects. We then cluster scene images in this semantic space to enable the system to exploit the underlying semantic structure of scene images and learn a more discriminant model for each subgroup. We show that our proposed approach outperforms traditional scene recognition methods when faced with challenging fine-grained scenes.

Motivated by the significant importance of objects in achieving a better scene understanding, we finally propose an approach to recognize a wide range of objects in scene parsing methods. Scene parsing aims at labeling regions of a scene image with their semantic classes, as a way of holistic scene understanding. Retrieval-based parsing systems rely on retrieving similar images to a given scene image and then computing label likelihoods

for each region in the given image. These likelihoods are obtained through matching the regions with those of the set of retrieved images in a nonparametric scheme. These systems have the advantage of scaling to a large number of scenes and objects, however they are heavily biased towards the recognition of background regions which harms the recognition of more salient foreground objects. We propose an approach that boosts the recognition of foreground objects in scene images by combining the label likelihoods from several nonparametric classifiers. We show how to design the different classifiers with the goal of maximizing the gain when combining their decisions. We also propose a method that reasons about which region labels often co-occur in one scene to discover outlier labels and recover missing labels in parsing results. We demonstrate that combining likelihoods and exploiting the scene context in terms of label statistics yields better parsing results than traditional retrieval-based systems.

Zusammenfassung

Das Ziel der automatischen Analyse und Erkennung visueller Szenen besteht darin, maschinelle Systeme zu befähigen, die Welt ähnlich wie Menschen wahrnehmen zu können. Dies würde Maschinen zahlreiche Möglichkeiten zur nahtlosen Interaktion mit Menschen eröffnen, was in geeigneten Szenarien die Lebensqualität Betroffener erhöhen kann.

Durch das Streben nach Erkennung von immer mehr unterschiedlichen Szenen mit tausenden von Objekten vollzieht sich allmählich ein Übergang hin zu einer feingranularen Perzeption visueller Szenen. Dabei steht man vor der Herausforderung, Algorithmen zur Szenenanalyse für die wachsende Zahl von Szenenklassen und Objekten skalierbar zu machen sowie genauer zwischen einzelnen Szenen hoher Auflösung diskriminieren zu können. Gleichzeitig soll dies in einer robusten und generalisierbaren Weise erreicht werden, sodass die entwickelten Algorithmen ohne Modifikation oder Anpassung Szenen erkennen können, seien sie von Smartphones, Digitalkameras, Überwachungskameras oder anderen Geräten aufgezeichnet worden.

Herkömmliche Methoden zur Szenenanalyse beruhen lediglich auf Informationen zum visuellen Erscheinungsbild. Um den angestrebten Detaillierungsgrad bei gleichzeitiger Generalisierbarkeit zu erreichen, muss auf abstrakterer Ebene zusätzlich die latente Semantik der Szenen genutzt werden. Diese semantische Information sollte es maschinellen Systemen in einer ähnlichen Weise wie einem Menschen ermöglichen, Aussagen, Zusammengehörigkeit und wechselseitige Bezüge der verschiedenen Bildkomponenten zu erkennen. Dabei stellen unter den verschiedenen Komponenten einer Szene die abgebildeten Realweltobjekte die semantisch reichhaltigsten Entitäten dar; sie geben Hinweise auf die Art der Szene, den Ort sowie die Bezugsstärke zu anderen Szenen. Darüber hinaus ermöglichen sie geeigneten Algorithmen, Schlüsse über die semantischen Beziehungen zwischen den verschiedenen Komponenten der Szene zu ziehen. Dementsprechend

präsentieren wir in der vorliegenden Arbeit Techniken zur feingranularen Szenenerkennung unter Nutzung von abstraktem kontextuellem Szenenwissen; wir zeigen dabei auf, wie in Szenen das visuelle Erscheinungsbild und der Kontext von Objekten gemeinsam genutzt werden kann, wie der zugrundeliegende semantische Raum exploriert werden kann, wie eine grosse Anzahl verschiedener Objekte erkannt werden kann und wie hierfür der globale Szenenkontext verwendet werden kann.

Bei vielen Anwendungen, wie zum Beispiel bei Assistenzsystemen oder in der Robotik, steht ein optisches Erkennungssystem vor der Herausforderung, dass eine signifikante Diskrepanz zwischen den Trainingsdaten und den Falldaten, auf denen das System ausgeführt wird, besteht. Noch grösser ist die Schwierigkeit, wenn während des Trainingsprozesses keine Daten der Anwendungsdomäne verfügbar sind. Wir zeigen, dass die Beschreibung eines Szenenbildes durch die abgebildeten Szenenobjekte eine effektive Herangehensweise zur Bewältigung dieser Herausforderung darstellt. Dies ist insbesondere dann der Fall, wenn keine Szenenbilder der Anwendungsdomäne während des Trainingsprozesses verfügbar sind, sondern höchstens Einzelbilder der Szenenobjekte. Wir schlagen vor, ein Szenenbild durch Erfassen aller feingranularen Objekte, die Teil der Szene sind, zu beschreiben. Das Erkennen dieser Objekte geschieht durch ein Multi-Label-Bildklassifikationsschema. Wir steuern den Klassifikationsvorgang unter Einbezug der visuellen Information, der Statistiken bezüglich des gemeinsamen Auftretens von Objekten und der Grösse des erwarteten Überschneidungsbereichs der erkannten Objekte im Szenenbild. Dies wird durch die Optimierung einer Energiefunktion erreicht, die diese drei Kriterien miteinbezieht, um eine abschliessende Kennzeichnung einer gegebenen Szene zu erzielen. Versuchsergebnisse belegen die Wirksamkeit und Effizienz unseres Ansatzes beim gleichzeitigen Erkennen aller Objektinstanzen eines gegebenen Szenenbildes in einem einzigen Optimierungsschritt.

Abgebildete Objekte stellen ein ausdrucksstarkes Konzept zur Beschreibung von Szenen dar. Jedoch kann es sein, dass unterschiedliche Szenen auf feingranularer Ebene aus teilweise gleichen Objekten bestehen, was Schwierigkeiten bei der Unterscheidung dieser Szenen mit sich bringt. In vielen Domänen, beispielsweise bei Ladenszenen, gibt es Teilgruppen von Szenenbildern, die stärker miteinander in Beziehung stehen als mit anderen Szenenbildern, etwa aufgrund des Vorhandenseins gemeinsamer Objekte. Das automatische Erkennen solcher Konfusionsgruppen erlaubt es dem System, gut diskriminierende Modelle für die einzelnen Teilgruppen zu erlernen, die dann in Kombination eine bessere Entscheidung ermöglichen. Dementsprechend schlagen wir vor, Szenenbilder durch bedingte Wahrscheinlichkeiten zu beschreiben, die jedes Bild dadurch charakte-

risieren, wie wahrscheinlich es – bedingt durch dessen konstituierende Objekte – zu einer bestimmten Szenenklasse gehört. Anschliessend werden die Szenenbilder in diesem semantischen Raum zu Clustern gruppiert, um dem System zu ermöglichen, die zugrundeliegenden semantischen Strukturen der Bilder zu nutzen und für jede Teilgruppe stärker diskriminierende Modelle zu erlernen. Wir zeigen, dass unser vorgeschlagener Ansatz die herkömmlichen Szenenerkennungsmethoden übertrifft, sobald er auf den schwierigeren feingranularen Szenen zur Anwendung kommt.

Motiviert durch die hohe Bedeutung der konstituierenden Objekte für die Optimierung der Szenenerkennung schlagen wir schliesslich einen Ansatz vor, mit dem bei der Szenenanalyse eine grosse Bandbreite an Objekten erkannt werden kann. Die Szenenanalyse zielt darauf ab, im Sinne eines ganzheitlichen Szenenverständnisses Teilbereiche eines Szenenbildes mit der zugehörigen semantischen Klasse zu kennzeichnen. Retrieval-basierte Analysesysteme beruhen darauf, zu einem gegebenen Szenenbild ähnliche Bilder abzurufen und mit diesen die Wahrscheinlichkeiten ("Likelihoods") zutreffender Bildlabels für jeden Teilbereich des gegebenen Szenenbildes zu berechnen. Die Werte erhält man durch einen nichtparametrischen Abgleich der Teilbereiche des Szenenbildes mit entsprechenden Bereichen in den abgerufenen Bildern. Diese Verfahren haben den Vorteil, dass sie gut mit der Zahl von Szenen und Objekten skalieren. Allerdings fokussieren sie stark auf die Erkennung von Hintergrundregionen, was die Erkennung von relevanten Objekten im Vordergrund verschlechtert. Wir schlagen daher für die Objekterkennung im Vordergrund von Szenenbildern einen Ansatz vor, bei dem die Bildlabel-Wahrscheinlichkeiten mehrerer nichtparametrischer Klassifikatoren kombiniert werden. Wir zeigen, wie diese Klassifikatoren konzipiert werden können, um den Gewinn ihrer kombinierten Entscheidung zu maximieren. Zudem schlagen wir eine Methode vor, die ermittelt, welche Labels von Teilbereichen in einer Szene oft gemeinsam auftreten, um Ausreisser zu erkennen und fehlende Labels zu ergänzen. Wir zeigen, dass man durch die Kombination von Bildlabel-Wahrscheinlichkeiten und die Nutzung des Szenenkontexts im Sinne der Label-Statistiken bessere Analyseergebnisse erzielt als mit herkömmlichen retrieval-basierten Systemen.