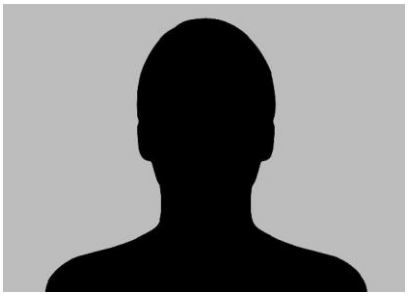
Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

InIT Institut für angewandte Informationstechnologie

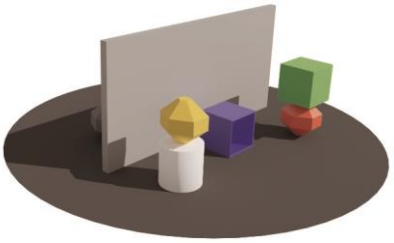
# Active Scene Understanding from Image Sequences for Next-Generation Computer Vision



Diplomierende
Ralph Meier
Dano Roost

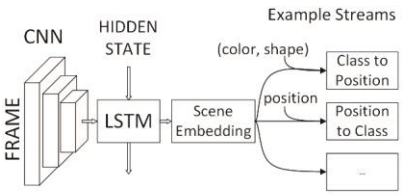
Dozierende
Thilo Stadelmann
Giovanni Toffetti Carughi

In den vergangenen Jahren konnten enorme Fortschritte im Bereich der digitalen Objekterkennung erzielt werden. Ansätze basierend auf Convolutional Neural Networks sind in der Lage, Katzen und Hunde voneinander zu unterscheiden oder zu erkennen, ob ein Bild an einem Strand oder in einem Wald aufgenommen wurde. Die meisten dieser Techniken arbeiten jedoch mit einzelnen, unabhängigen Bildern und können Informationen aus verschiedenen Blickwinkeln nicht kombinieren. Deshalb haben sie auch kein Verständnis für Objektpermanenz, also dass ein Objekt immer noch existiert, auch wenn es gerade nicht sichtbar ist. Ebenfalls sind die meisten dieser Ansätze nicht in der Lage zu entscheiden, von welchem Winkel die Szene als nächstes betrachtet werden sollte, um ein möglichst komplettes Bild zu erhalten.



Beispielsszene mit mehreren zu identifizierenden Objekten und einem Hindernis in der Mitte.

Diese Arbeit untersucht Ansätze, um diese Mängel zu beseitigen und Computern ein tieferes Verständnis einer Szene zu ermöglichen. Dazu benutzen wir synthetische 3D-Szenen und rendern Bilder aus 36 Blickwinkeln. Unter Verwendung von Sequenzen dieser Bilder muss das System Fragen zur Szene beantworten, wie zum Beispiel die Auswertung von Form und Farbe für ein Objekt an einer bestimmten Position im 3D-Raum oder die Auflistung aller gefundenen Objekte. Wir evaluieren drei verschiedene Netzwerkarchitekturen (CNN, CNN + RNN, ConvLSTM), von denen zwei in der Lage sind, Informationen über mehrere Zeitschritte zu aggregieren. Die Kamerapose steht dabei nicht als Input zur Verfügung, sodass das System selbst lernen muss, Informationen von verschiedenen Blickwinkeln zu kombinieren.




Die am besten performende Netzwerkarchitektur, bestehend aus Convolutional, Recurrent und Fully Connected neuralen Netzwerken.

Als besonders erfolgreich zeigt sich dabei eine Kombination von Convolutional und Recurrent Neural Networks. Eine noch bessere Leistung wird erzielt, wenn das System zudem lernt, die Kamera selbst zu steuern. Mittels dieser Methode werden im Durchschnitt nur noch fünf verschiedene Blickwinkel benötigt, um bessere Resultate zu erreichen als mit jeder anderen getesteten Art der Kamerasteuerung. In Szenen mit einem Hindernis-Objekt übertrifft die Architektur, die Convolutional und Recurrent Neural Networks kombiniert, den reinen Convolutional-Ansatz um bis zu 97 %. Wir können zeigen, dass unser Ansatz in der Lage ist, viele gängige Unzulänglichkeiten traditioneller Objekterkennungsansätze zu lösen, darunter das Verständnis von 3D-Okklusion, die Fähigkeit, Informationen über viele Einzelbilder hinweg zu aggregieren oder die aktive Steuerung der Kamera, um die Szene aus einem optimalen Blickwinkeln zu erfassen.