Top-Down Multimodal Object Detection for Interactive Tables

Thema:
Top-Down Multimodal Object Detection for Interactive Tables
Art:
BA
BetreuerIn:
Andreas Schmid
BearbeiterIn:
Maximilian Hogger
ErstgutachterIn:
Bernd Ludwig
ZweitgutachterIn:
Raphael Wimmer
Status:
abgeschlossen
Stichworte:
vigitia, object detection, interactive tables, computer vision, machine learning
angelegt:
2023-06-29
Antrittsvortrag:
2023-07-24
Abgabe:
2023-10-16
Textlizenz:
CC-BY
Codelizenz:
MIT

Hintergrund

Im Rahmen des Projekts VIGITIA werden herkömmliche Tische mit Projected Augmented Reality um virtuelle Elemente erweitert. Da Tracking von Gegenständen und Interaktion durch über der Tischoberfläche montierte Kamereas umgesetzt werden, kommt es zu einigen technischen Herausforderungen. Ein Beispiel dafür ist die Erkennung von Objekten auf der Tischoberfläche.

Wenngleich Machine-Learning-Ansätze für Objekterkennung mittlerweile sehr robust funktionieren, basieren solche Modelle auf Trainingsbilder, die meist frontal aufgenommen wurden. Einige Objektklassen, wie zum Beispiel Flaschen und Geschirr, werden deshalb aus der Vogelperspektive nicht erkannt.

In einem Projekt aus dem Master-Forschungsseminar [1] konnte gezeigt werden, dass es durch dedizierte Trainingsdatensätze möglich ist, diese Objekte auch aus der Vogelperspektive zu klassifizieren. Jedoch hatte dieser Ansatz immer noch das Problem, dass sehr ähnlich aussehende Objekte, wie zum Beispiel Teller und Schüsseln, noch nicht wirklich robust klassifiziert werden

konnten.

Dieses Problem könnte gelöst werden, indem zusätzlich zu einem RGB-Bild ein Tiefenbild für die Klassifizierung herangezogen wird.

Zielsetzung der Arbeit

In dieser Arbeit soll evaluiert werden, inwiefern sich Tiefenbilder eignen, um Objektklassifikation im genannten Kontext zu verbessern. Dazu sollen verschiedene Architekturen und Vorverarbeitungsmethoden verglichen werden. Zum Schluss soll eine Referenzimplementierung erstellt und mit einem rein auf RGB-Bildern basierenden Modell verglichen werden.

Konkrete Aufgaben

- **1 Integration von Farb- und Tiefeninformationen zur gegenseitigen Vorverarbeitung**

- 1.1 Vorverarbeitung des Tiefenbilds anhand des Farbbildes[2] Auffüllen von Löchern und Rauschen Reduzieren im Tiefenbild anhand der zugehörigen Farbinformation
- 1.2 Integration der Tiefeninformation in das Farbbild Unter anderem Multiplikation der Farbkanäle mit normalisiertem Tiefenbild.

- **2 Objekterkennung anhand eines Farb- und eines Tiefenmodells.**

  Es werden zwei YOLOv8 [3] Objekterkennungsmodelle trainiert, eines auf den Farbbildern und eines auf den Tiefenbildern. Mit Hilfe beider Modelle soll dann die Objekterkennung durchgeführt werden. (Aufgrund dessen, dass Yolov8 nur die Verarbeitung von 3 Farbkanälen unterstützt müssen dafür zwei Modelle verwendet werden.)

- 2.2 Weighted Boxes Fusion [4] Mit Hilfe dieses Algorithmus sollen auf basis der Predictions des Farb- und Tiefenmodells eine finale Vorhersage berechnet werden.
- 2.3 Ensemble learning in CNN – EnsNet[5] Die in dieser Arbeit vorgeschlagene Architektur muss an die Objekterkennung mit zwei Modalitäten angepasst werden. Mit Hilfe der Implementierung soll die Objekterkennung mit dem Farb- und dem Tiefenmodell ermöglicht werden.

- **3 Vergleichende Analyse**

- 3.1 Erstellen eines neuen Repräsentativen Datensatzes Auf Basis der Erkenntnisse aus dem Ergebnisbericht zur Umfrage „Tischnutzung im Alltag“, durchgeführt im Rahmen des VIGITIA Projekts, soll ein kleiner repräsentativer Datensatz zur Evaluation erstellt werden. Dabei sollen systematisch Parameter wie die Entfernung des Tischs zur Kamera, und die Beschaffenheit der Tischoberfläche systematisch variiert werden.
- 3.2 Vergleichende Analyse der Stärken und Schwächen des verschiedenen Ansätze

Die verschiedenen Bildvorverarbeitungsschritte und die Objekterkennungsmodelle sollen anhand von Metriken wie mAP, IoU, Precision, Recall, F1-Score und der Prediction Zeit verglichen werden. Die Evaluation erfolgt dabei auf dem eigens erstellten Datensatz.

Erwartete Vorkenntnisse

Grundlagen Python, Vorkenntnisse in PyTorch vorteilhaft. Interesse an

Weiterführende Quellen

[1] Markus Bink & Julian Höpfinger (2023). A Matter of Perspektive: Top-Down Object Detection for Interactive Tables. Forschungsseminar Master Medieninformatik, WS22/23.

[2] W. Li, W. Hu, T. Dong and J. Qu, „Depth Image Enhancement Algorithm Based on RGB Image Fusion,“ 2018 11th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 2018, pp. 111-114, doi: 10.1109/ISCID.2018.10126.

[3] https://github.com/ultralytics/ultralytics

[4] Weighted boxes fusion: Ensembling boxes from different object detection models, Image and Vision Computing, Volume 107, 2021

[5] Hirata, D., & Takahashi, N. (2023). Ensemble learning in CNN augmented with fully connected subnetworks. IEICE TRANSACTIONS on Information and Systems, 106(7), 1258-1261.[4] Roman Solovyev, Weimin Wang, Tatiana Gabruseva,