HOCHSCHULE MAINZ
UNIVERSITY OF
APPLIED SCIENCES

SEMANTISCHE SEGMENTATION FÜR MIXED REALITY

Projektleitung:
Prof. Dr.-Ing. Anita Sellent

Laufzeit:
1.4. – 30.9.2023

Kooperationspartner:
Sonali Patil, Georgia Albuquerque, Andreas Gerndt, Deutsches Zentrum für Luft- und Raumfahrt

Kontakt, weitere Informationen:
anita.sellent@hs-mainz.de

Motivation

Mixed Reality-Anwendungen erlauben es, Bilder eines virtuellen Systems mit der echten, physikalischen Umgebung des Benutzers zu überblenden. Um eine exakte Überlagerung, z.B. virtuell eingeblendeter Leitungen mit physikalisch vorhandenen Steckdosen zu ermöglichen, sind aktuell lästige Vorbereitungen der Umgebung notwendig, z.B. das Anbringen von Marken.

In der Computer-Vision wurden in dem letzten Jahrzehnt durch Convolutional Neural Networks (CNNs) große Fortschritte in der Bildklassifizierung, der Objekt-Detektion und der semantischen Segmentierung ermöglicht. Ziel des Gesamtprojekts ist es, diese Fortschritte für das Kamera-Tracking von Mixed Reality-Anwendungen zu verwenden. Ein erster Meilenstein ist dabei die Auswertung der Qualität der semantischen Segmentierung in typischen Mixed Reality-Umgebungen.

Realisierung

Die Anwendung von semantischer Segmentierung in der Mixed Reality stellt hohe Anforderungen, da Daten in Echtzeit verarbeitet werden müssen. Andererseits sind typische Mixed Reality-Brillen mit mehreren kalibrierten Kameras ausgestattet, die eine Entfernungsberechnung zulassen. Im Stand der Technik zeichnet sich EMSANet [1] als eine Umsetzung der semantischen Segmentierung auf RGB-D Daten durch seine Echtzeitfähigkeit und seine hohe Genauigkeit aus.

In unserer Anwendung soll EMSANet auf Innenraumdaten angewendet werden, da dies eine typische Arbeitsumgebung für Mixed Reality-Anwendungen, z.B. im Renovierungs- oder Remote-Support-Bereich, ist. Für die Anwendung wurde eine RGB-D Kamera mit einem aktiven Stereo-Modul im Infrarotbereich gewählt. Der Algorithmus benötigt zum Training viele Daten, um die semantischen Klassen der Pixel zu erlernen. Da die anwendungsspezifische Aufbereitung der Daten gerade für kleine Unternehmen oft viel zu aufwändig ist, wurde eine öffentliche Trainingsdatenbank verwendet, die möglichst ähnliche Daten aufweist, hier [2].

Ergebnis

Die verwendete Stereo-Kamera erlaubt es, Tiefenkarten in Echtzeit direkt auf der Kamera zu berechnen. Aufgrund der Beschränkung der Rechenkapazität sind diese Tiefenkarten stark verrauscht, siehe Abb. 2 links. Damit unterschieden sie sich stark von den Trainingsdaten. Die damit erzielten semantischen Segmentierungen zeigen ein hohes Rauschverhalten, Abb. 3 links, und die erfolgreiche Klassifizierungsrate in einem solchen Aufbau wurde stark reduziert, siehe Tab. 1.

Trainingsdaten	Testdaten	mIoU
NYU refined	NYU refined	49.42%
NYU refined	NYU raw	37.75%
NYU raw	NYU refined	44.30%
NYU raw	NYU raw	48.12%

Tabelle 1: Numerische Evaluation der semantischen Segmetierung zeigt eine bessere Übereinstimmung zwischen vorhergesagten und bekannten Pixellabeln (mean intersection over union, mIoU), wenn Tiefenwerte mit gleicher Qualität verwendet werden.

Um bessere Trainingsdaten zu erzeugen, wurden die Infrarotbilder von der Kamera heruntergeladen und ein CNN-basierter Stereoalgorithmus verwendet [3]. Die damit berechneten Tiefenkarten gleichen den Trainingsdaten mehr und die Ergebnisse sind stabiler und genauer, siehe Abb. 3 rechts.

Abbildung 1: Die verwendete AR Kamera nimmt gleichzeitig ein Farbbild (links) und zwei Infrarotbilder (Mitte und rechts) auf.

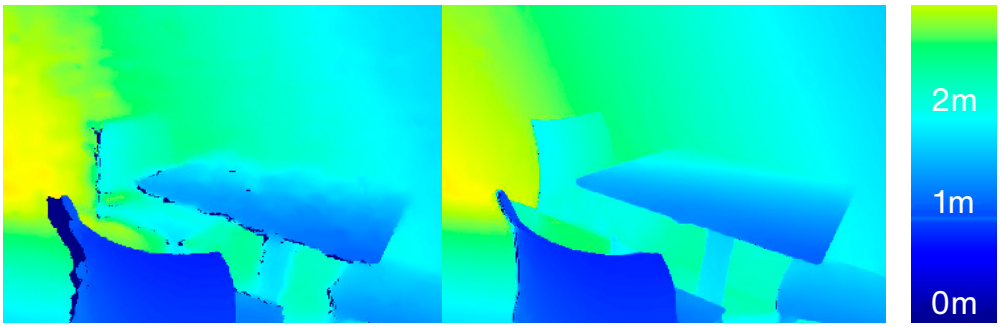

Abbildung 2: Das Infrarot-Stereopaar aus Abb. 1 wurden zur Berechnung einer Tiefenkarte verwendet: die kamerainterne Software benötigt keine weitere Rechenzeit, liefert aber stark verrauschte Tiefenkarten (links). Ein externer Echtzeit-Algorithmus [3] liefert genauere Tiefenkarten (rechts).

Abbildung 3: Die semantische Segmentierung auf den verrauschten Eingabedaten aus Abb. 2 (links) zeigt häufigere Fehlklassifizierungen als auf den genaueren Eingabedaten aus Abb. 2 (rechts).

Um von der Effizienz der Kamera zu profitieren, wurde der Algorithmus auch auf verrauschten Tiefendaten der Trainingsdatenbank trainiert. Obwohl das quantitative Korrektheitsmaß mIoU damit wieder etwas verbessert werden konnte, Tab. 1, zeigen sich in der Anwendung weiterhin stark verrauschte Segmentierungsergebnisse. Für das weitere Vorgehen im Gesamtprojekt kann also die klare Empfehlung ausgesprochen werden, Rechenzeit für die Bestimmung exakter Tiefenkarten zu verwenden, da die semantische Segmentierung davon stark profitiert.

Literatur

[1] SEICHTER, Daniel, et al. Efficient multi-task rgb-d scene analysis for indoor environments. In: 2022 International Joint Conference on Neural Networks (IJCNN). IEEE, 2022. S. 1-10.

[2] SILBERMAN, Nathan, et al. Indoor segmentation and support inference from rgbd images. In: Computer Vision–ECCV 2012. S. 746-760.

[3] LIPSON, Lahav; TEED, Zachary; DENG, Jia. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In: 2021 International Conference on 3D Vision (3DV). IEEE, 2021. S. 218-227.