



Hochschule  
Bonn-Rhein-Sieg  
*University of Applied Sciences*

**Fachbereich Informatik**  
*Department of Computer Science*

## Bachelorarbeit

# **Automatisches Labeln von Objekten in einer Augmented Reality Umgebung**

von  
**Janelle Pfeifer**

Erstprüfer: Prof. Dr. Ernst Kruijff  
Zweitprüfer: Prof. Dr. André Hinkenjann  
Eingereicht am: 8. Oktober 2020

## **Erklärung**

Janelle Pfeifer  
Delpstraße 28  
53359 Rheinbach

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbst angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Die Arbeit wurde bisher keiner Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Ort, Datum Unterschrift

**Inhaltsverzeichnis**

<b>Abkürzungsverzeichnis</b>	v
<b>1 Einleitung</b>	1
1.1 Zielsetzung . . . . .	1
<b>2 Related Work</b>	3
2.1 Huynh et al. (2019): In-Situ Labeling for Augmented Reality Language Learning . . . . .	3
2.2 View Management for Virtual and Augmented Reality . . . . .	3
2.3 3. Related Work . . . . .	4
<b>3 Grundlagen</b>	5
3.1 Grundlagen zu Augmented Reality . . . . .	5
3.1.1 Augmented Reality . . . . .	5
3.2 AR System . . . . .	5
3.3 Spatial Mapping . . . . .	5
3.4 View Management . . . . .	6
3.5 Magic Leap AR Brille . . . . .	6
3.5.1 Hardware . . . . .	6
3.5.2 Betriebssystem . . . . .	7
3.5.3 Unity Applikationen für Magic Leap One . . . . .	8
3.6 Grundlagen zu 3D Szenen für AR . . . . .	8
3.6.1 Lokale und globale Koordinatensysteme in 3D Szenen . . . . .	8
3.6.2 Kamera in 3D-Computergrafik . . . . .	8
3.7 Computer Vision . . . . .	8
3.7.1 Object Detection . . . . .	9
3.7.2 Artificial Neural Networks . . . . .	9
3.7.3 Convolutional Neural Networks . . . . .	9
3.7.4 Azure Computer Vision . . . . .	10
3.7.5 Azure Custom Vision . . . . .	10
<b>4 Umsetzung</b>	11
4.1 Design des Vorgang der Objekt Erkennung . . . . .	11
4.2 Architektur . . . . .	11
4.3 Interaktion . . . . .	13
4.4 Implementierung der Objekt Erkennung . . . . .	14
4.5 Ein Foto aufnehmen . . . . .	15
4.5.1 Object Detection . . . . .	15
4.5.2 Von dem Foto zum 3D Raum . . . . .	17
4.5.3 Raycast . . . . .	19
4.5.4 LabelCreator . . . . .	20
4.5.5 Azure Custom Vision . . . . .	22
4.6 Entwicklung der Foto-Repräsentation . . . . .	23
<b>5 Auswertung</b>	27
5.1 Laufzeitanalyse . . . . .	27
5.2 Analyse durch Azure Objekt Detection . . . . .	28
5.3 Azure Custom Vision . . . . .	28
5.4 Objekte in 3D Szene Lokalisieren . . . . .	29

<b>6 Zusammenfassung</b>	<b>31</b>
6.1 Konzepte und Implementierte Funktionen . . . . .	31
6.2 Ausblick . . . . .	31
<b>7 Literaturverzeichnis</b>	<b>32</b>
<b>8 Anhang</b>	<b>35</b>

**Abkürzungsverzeichnis**

## 1 Einleitung

Augmented Reality (AR) ist eine Vermischung der realen Welt mit virtuellen Elementen. Es wird durch Anzeigegeräte, wie Handys, Tablets oder Augmented Reality Brillen präsentiert und bietet ein intuitives Benutzerinterface um Informationen über Objekten der realen Welt anzuzeigen. Eine AR Umgebung bietet dem Nutzer eine erweiterte Wahrnehmung der realen Welt, indem sie diese anzeigt und gleichzeitig 3D Objekte, 2D Overlays oder Audioelemente hinzufügt.

Die Interaktion der virtuellen Elemente miteinander, mit dem Nutzer und der realen Umgebung ist ein Grundbestandteil von AR Anwendungen. Um die Interaktion mit der Umgebung zu ermöglichen müssen Informationen über die Geometrie der Umgebung vorliegen. Es gibt somit ein grobes Verständnis davon, wie ein Mesh der Umgebung aussieht.

Dieses Geometrische Verständnis kann durch ein semantisches Verständnis der Umgebung erweitert werden. Dieses ermöglicht komplexere, Interaktionen zwischen digitalen und virtuellen Elementen. Sematische Informationen über die Umgebung können durch die Gegenstände erschlossen werden, die sich darin befinden.

Es gibt mehrere Möglichkeiten Gegenstände zu erkennen. Zum einen können Markierungen in der realen Welt verwendet werden. Dabei handelt es sich um statische Bilder, beispielsweise ein Foto, oder ein QR Code, die von einer Kamera eingescannt werden. Der Marker ist einzigartig für jedes Objekt, damit sie voneinander unterschieden werden können. Der Nachteil bei diesem Vorgehen ist der Arbeitsaufwand, der damit verbunden ist, jeden Gegenstand einzeln zu markieren.

Wenn man Markierungen in der realen Welt umgehen möchte, kann man den Nutzer der Applikation bitten, beispielsweise per Geste auf Objekte der realen Welt zu weisen, die erkannt werden sollen. Für jedes der Objekte muss der Nutzer angeben, um welche Art von Gegenstand es sich handeln, damit die Applikation unterschiedliche Objekte auseinander halten kann und die korrekten Informationen mit ihnen assoziiert. Auch hier ist der Arbeitsaufwand hoch.

Beide der Verfahren lassen sich schlecht skalieren um große AR Umgebungen abzudecken. Nur eine vollautomatische Objekterkennung ist skalierbar. Damit könnte man mit deutlich weniger aufwand Semantische Informationen über eine reale Umgebung erfahren und somit Komplexere Anwendungsgebiete für AR erschließen.

Um diese Automatisierung zu erreichen, kann Image based Object Detection aus dem Bereich der Computer Vision verwendet werden. Dieses Verfahren ist darauf ausgelegt Objekte in Bildern zu erkennen. Die Objekterkennung mithilfe von 2D Abbildungen ist am performantesten. Und Besser als versuche 3D Wolken zu interpretieren.(O'Shea und Nash 2015)

### 1.1 Zielsetzung

In dieser Thesis wird das Erkennen und Labeln von Objekten in einer AR Umgebung, mithilfe von Image based Objekt Detection, automatisiert. Das AR Gerät Magic Leap One Lightwear wird als Benutzerinterface und Plattform verwendet.

Das Minimalziel besteht darin Fotos der Umgebung zu analysieren und gefundene Objekte in einer 3D Szene mit Labels zu versehen. Mithilfe der Kamera des AR Gerätes werden Fotos von der Umgebung aufgenommen. Diese Fotos werden, durch ein trainiertes Neuronales Netzwerk nach Objekten durchsucht. Dies Positionen der werden in einer digitalen Abbildung der Umgebung lokalisiert und mit Labels markiert.

Das erweiterte Ziel besteht darin ein zweites Neuronales Netzwerk in die Objekt Erkennung einzubinden. Dieses kann trainiert werden, spezifischen Objekten zu erkennen und damit die Semantische Information zu erweitern die Erkannt werden kann.

Als Maximalziel werden die Labels der erkannten Objekte als virtuelle Elemente mit der Magic Leap One dargestellt.

## 2 Related Work

### 2.1 Huynh et al. (2019): In-Situ Labeling for Augmented Reality Language Learning

Huynh et al. (2019) schafft ein Framework, mit dem die Lernmethode "lociün Augmented Reality umgesetzt und erweitert werden kann. Die Lernmethode beruht darauf, Gegenstände der Welt mit Notizen zu beschriften.

Dafür wurde folgende automatische real-time Objekt Erkennung entwickelt:

Mithilfe von Image Based Object Detection werden Objekte auf Fotos der AR Umgebung erkannt. Diese Objekte werden dann in die 3D Szene der Umgebung übernommen.

Die AR Brille hat zu wenig Rechenleistung, um Image Based Object Detection durchzuführen. Daher wurde eine Server Client Architektur aufgesetzt.

Die Videokamera der AR Brille wird verwendet um Bilder von der Umgebung aufzunehmen. Die einzelnen Frames werden an den Server geschickt. Dieser nutzt ein Object Recognition Learning Modell um alle erkennbaren Objekte in dem Bild zu finden und mit Bounding Boxen zu lokalisieren.

Die ObjectDetection API von TesnorFlow wird verwendet. Es findet mehrere Objekte in einem Foto in einer Analyse und gibt Bounding Boxen an. Damit die Objekt Erkennung in real-time durchgeführt werden kann, wird die niedrigste Kamera Auflösung mit 896x504 verwendet. Zusätzlich werden die Fotos als JPEG mit 50 Prozent Qualität komprimiert. Damit braucht die Analyse 30 ms pro Foto, was eine Real-Time Erkennung mit 30 frames per second erlaubt.

Trotzdem ist die Erkennung in der Applikation verspätet, durch einen Netzwerk Delay von 150ms zwischen der Hololens und dem Server, der die Foto-Analyse durchführt.

Die Hololens nummeriert die Frames, die an den Server versicht werden. Zusätzlich wird für jedes der Frames, die Kameraposition gespeichert, mit der es aufgenommen wurde. So können Frames asynchron analysiert werden.

Ist die Analyse durchgelaufen, wird die Bounding Boxen der Objekte und die Kameraposition genutzt, um die Objekte in der 3D Umgebung zu lokalisieren. Dafür wird der Mittelpunkt jeder Bounding Box mithilfe eines Raycastes in die 3D Szene projiziert.

Um Fehlern bei der Object Erkennung entgegenzuwirken, wird ein Objekt erst als endgültig erkannt angesehen, wenn es auf mehreren Fotos erkannt erkannt wurde. Mehrere Frames werden verwendet um die Position des Objekte abzuschätzen. Dann werden bereits existierende Label untersucht, die in den letzten 60 Frames aufgenommen wurden. Wenn die Labels nah beieinander liegen wird davon ausgegangen, das es sich um das-selbe Objekt handelt. Der Mittelpunkt der Labels wird zu der Position des Objektes und wird mit einem endgültigen Label versehen.

### 2.2 View Management for Vitual and Augmented Reality

Bell et al. (2001) beschreibt View Management für interaktive 3D Benutzeroberflächen. Als View Management wird das positionieren von Labels bezeichnet. Die Labels können sich auf eine 2D Ebene beschränken, oder im 3D Raum liegen. Das Ziel des View Management für AR ist es die Labels so zu positionieren, das sie einander und relevante reale Objekte nicht verdecken. Gleichzeitig sollen die Label Gegenständen der Realen Welt auf eine verständliche weise annotieren. Sie sollen beispielsweise nahe bei den Objekten liegen, zu denen sie gehören.

Die Applikation die in dieser Arbeit erstellt wurde, verfügt über Label im 3D Raum. Die Lesbarkeit der Label kann durch View Management verbessert werden, indem die Positionen der Labels über zeit verändert wird, wenn mehr Labels hinzukommen. Durch das hinzukommen von Labels werden auch Gegenstände der realen Welt markiert. Im Zuge

des View Managements kann sichergestellt werden, das die Gegenstände nicht verdeckt werden.

View Management geht jedoch über den Rahmen dieser Arbeit hinaus.(Bell et al. 2001)

### 2.3 3. Related Work

Chen et al. (2018) stellen ein Framework vor, in dem einer AR Umgebung semantische Eigenschaften zugewiesen wird um realistische Interaktionen zwischen virtuellen und realen Objekten zu erreichen. Insbesondere sollen physikalische Interaktionen realistischer werden.

Das Framework reichert die Umgebung mit Informationen über die Materialien an, aus denen reale Oberflächen und Gegenstände bestehen. Die Materialien werden mit Labels versehen und die physikalischen Interaktionen berücksichtigen die Materialien der Umgebung.

Als beispielhafte Applikation wurde ein First-Person-Shooter vorgestellt, bei dem das aussehen von Einschusslöchern davon anhängt auf welches Material geschossen wurde.

Für die Erkennung der Materialien werden Frames der Hololens nach Materialien segmentiert. Diese Analyse ist nicht Echtzeit fähig, die Interaktionen können jedoch in Echtzeit ablaufen. Die semantischen Informationen werden abgespeichert um bei späteren Interaktionen abgerufen zu werden. Die Erkennung der Materialien ist nicht Echtzeit fähig, aber durch das speichern der Materialien im Raum können die Interaktionen in echt Echtzeit ablaufen. Die Semantischen Informationen müssen nicht zu jedem Frame bestimmt werden, sondern nur in Abständen erhoben werden.

Um die Semantik der Umgebung zu erheben, werden RGB-Bilder von ihr aufgenommen und mit einem neuronales Netzwerk analysiert. Das neuronale Netz wurde von Chen et al. (2018) für den First-Person-Shooter trainiert. Es kann 23 unterschiedliche Materialien erkennen und segmentiert Bilder danach. Dabei wird für jedes Pixel ein Material angegeben.

Mithilfe der Camera Position des Frames werden die Material-Informationen auf das 3D Modell der Umgebung projiziert.

Ein Kinect Sensor wird als Tiefenkamera verwendet.

Porblem damit das es pixelweise ist. und sich überlappen kann.

### 3 Grundlagen

#### 3.1 Grundlagen zu Augmented Reality

##### 3.1.1 Augmented Reality

Augmented Reality vermischt die reale Welt mit digitalen (virtuellen) Elementen um dem Nutzer eine erweiterte Wahrnehmung zu ermöglichen. Es können 3D Objekte, 2D Overlays oder Audioelemente verwendet werden um eine reale Umgebung mit Informationen zu bereichern.

Die Umgebung bezeichnet den Teil der realen Welt, der in Augmented Reality abgebildet und erweitert werden soll. Beispielweise ein Zimmer, in dem eine AR Anwendung ausgeführt wird. Die AR Umgebung umfasst die reale Umgebung und die virtuellen Elemente. Augmented Reality weist drei grundlegende Merkmale auf.

- Die Kombination der Realität mit dem Virtuellen. Besteht darin, dass die Realität mit virtuellen Elementen überlagert wird.
- Die Interaktion mit virtuellen Elementen erfolgt in Echtzeit.
- Virtuelle Elemente haben einen festen räumlichen Platz in der AR Umgebung.

Die Merkmale unterstützen ein möglichst nahtloses verschmelzen der realen Welt mit den virtuellen Elementen.

Die Navigation in einer AR Umgebung funktioniert, indem der Nutzer sich durch physisch durch die reale Umgebung bewegt. Die reale Umgebung und die virtuellen Elemente stehen immer in dem gleichen räumlichen Verhältnis zueinander.

Da die reale Welt immer zu sehen ist, gibt sie eine Referenz und einen Kontext für die virtuellen Objekte an. Beispielsweise steht die Größe von virtuellen Objekten immer in Relation zu der realen Umgebung.(Dörner et al. 2019)

#### 3.2 AR System

Ein AR System besteht aus Hardware und Software, die benötigt wird um die Wahrnehmung der realen Welt mit virtuellen Elementen zu erweitern.

Die Vermischung der realen Welt mit virtuellen Elementen muss angezeigt werden.

Die Interaktion des Nutzers mit virtuellen Elementen, und die Interaktion von virtuellen Elementen mit der realen Welt muss simuliert werden.

Ein AR System ist in der Regel nicht an einen bestimmten Ort gebunden. Das System kann in unterschiedlichen Umgebungen eingesetzt werden, die unterschiedliche reale Gegenstände aufweisen. AR Applikationen müssen unterschiedliche Umgebungen unterstützen.(Dörner et al. 2019)

#### 3.3 Spatial Mapping

Um virtuelle Objekte an eine Umgebung anzupassen und die Interaktion zwischen virtuellen Objekten und der realen Umgebung zu ermöglichen, benötigt ein AR System Informationen über die Geometrie der Umgebung.

Mit den Sensoren der AR Hardware werden Informationen gesammelt, die Aussagen über die Geometrie der Umgebung geben. Beispielsweise haben AR Geräte eine Tiefenkamera, die die Entfernung messen kann. Die Daten der Sensoren werden gesammelt und in Relation zu der Bewegung des Gerätes gesetzt um die Umgebung zu Rekonstruieren. Dieser Vorgang nennt sich Spatial Mapping.

Mit der Entstehenden Spatial Map können digitale Elemente mit der Umgebung interagieren, diese verdecken oder von ihr verdeckt werden.(Microsoft 2018b)

### 3.4 View Management

Die virtuelle Information, die einen teil der realen Welt bereichern werden meistens als 3D Elemente angezeigt. Die Information kann jedoch auch in 2D Elementen angezeigt werden, die sich auf eine 2D Ebene beschränkt. Insbesondere Labels, die reale Objekte erklären, können auf diese Art angezeigt werden.

Das Layout der 2D Elemente auf der Ebene wird durch View Management optimiert. Idealerweise werden die Elemente so positioniert, das sie sich Gegenseitig nicht verdecken, relevante Bereiche der realen Welt nicht verdecken. Zusätzlich sollen sie nah an den Gegenständen der Realen Welt bleiben, die sie annotieren. Siehe Abbildung 2.



Abbildung 1: Labels durch View Management positioniert.(Azuma und Furmanski 2003)

Das Layout muss angepasst werden, wenn sich der View verändert. Gleichzeitig soll das Layout stabil bleiben und sich nicht verändern, wenn ein Anderes Layout ein wenig besser wäre, um ein hin und her springen zwischen zwei möglichen Layouts zu vermeiden.(Azuma und Furmanski 2003)

### 3.5 Magic Leap AR Brille

Die Magic Leap One Lightwear ist eine Augmented Reality Brille, die von dem Unternehmen Magic Leap entwickelt wurde. Sie verfügt über ein Head-Mounted Display und einer Recheneinheit die über ein Kabel mit dem Display verbunden ist. Die Recheneinheit kann an der Hüfte getragen werden, was die AR Brille komplett Mobil macht.

#### 3.5.1 Hardware

Die Recheneinheit besitzt zwei Denver 2.0 64 Bit Prozessor-Kerne und vier ARM Cortex A57 46 bit Kerne. Davon ist einer der Denver Kerne und zwei der ARM Cortex Kerne für Applikationen nutzbar.

Sie besitzt neun Sensoren und mehrere Kameras. Dazu gehören:

- ein Infrarot Tiefen-Sensor,
- ein Eye Tracker,
- eine Foto und Video Kamera, die im Format 16:9 mit einer Auflösung von 1920 x 1080 Pixeln aufnehmen,
- Umgebungskameras die in unterschiedliche Richtungen ausgerichtet sind. (Magic Leap 2018, 2020b)

Der Output geschieht über ein Display mit einem 50 Grad Field of View und einem Seitenverhältnis von 4:3. Das Display ist transparent. Daher kann die reale Welt immer betrachtet werden. Selbst wenn ein weißes Objekt angezeigt wird, schimmert die reale Welt noch durch. Das Display kann keine Schwarzen Objekte anzeigen.

Eingaben erfolgen über einen 6 Degree of Freedom Controller. Er verfügt über 3 Knöpfe (Trigger, Bumper, Home Button) und ein Touchpad. (MagicLeap 2018, 2020b)

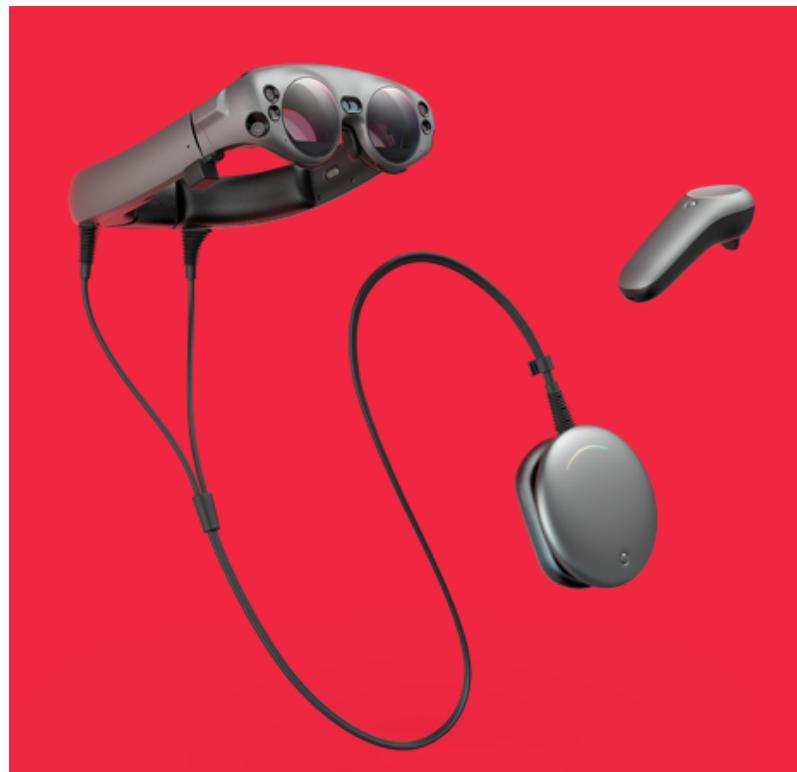


Abbildung 2: Magic Leap One AR Brille.(MagicLeap)

### 3.5.2 Betriebssystem

Die Magic Leap Brille läuft auf dem Betriebssystem Lumin OS. Dieses wurde für Augmented Reality entwickelt und bietet Applikationen entsprechende Funktionalitäten an. Beispielsweise führt das Betriebssystem Spatial Mapping durch.(MagicLeap 2019a, 2020c)

Dabei werden mit den Sensoren und Kameras der Brille Daten aufgenommen und in einen zeitlichen Zusammenhang mit der Bewegung der Brille gesetzt, um eine Rekonstruktion des Raumes zu erhalten.(MagicLeap 2019a, b, 2020a, c)

Lumin OS bietet es Applikationen an,

- Raycast auf die Umgebung durchzuführen und
- ein Mesh der Rekonstruktion zu erhalten.

Neben dem Spatial Mapping unterstützt Lumin OS das Verarbeiten vom Input des Controllers und verwaltet die Zugriffsrechte der Applikationen. Dazu gehört beispielsweise der Zugriff auf die Fotokamera und das Netzwerk.(Leap 2018; MagicLeap 2020c)

Es können niemals mehrere Applikationen zugriff auf die physikalische Kamera der Magic Leap One haben. Kamera Ressource Permission stufen

### 3.5.3 Unity Applikationen für Magic Leap One

## 3.6 Grundlagen zu 3D Szenen für AR

Die Virtuellen Inhalte der AR Anwendung werden in einer virtuellen Szene gespeichert. AR Anwendungen müssen in Echtzeit laufen. Daher muss die virtuelle Szene echtzeitfähig sein. Im besten Fall ist für den Nutzer kein Unterschied zwischen der virtuellen Welt und der realen Welt zu bemerken bezüglich auf zeitliche Verzögerungen.

Für eine AR Anwendung werden relevante Teile der realen Welt in der 3D Szene repräsentiert, um die Interaktion mit digitalen Elementen zu ermöglichen. So sind beispielsweise die Wände und der Boden eines Raumes, sowie die Position und die Blickrichtung des Nutzers. Auch die Position eines Eingabekontrollers und die Blickrichtung des Nutzers kann mit entsprechenden Sensoren verfolgt und in die Szene miteinbezogen werden.

### 3.6.1 Lokale und globale Koordinatensysteme in 3D Szenen

In einer 3D Szene werden die Positionen von Objekten als Matrizen in dreidimensionalen Koordinatensystemen verwaltet. Es gibt ein globales Koordinatensystem (auch Weltkoordinatensystem oder World Space), in dem alle Objekte relativ zu einem gewählten Ursprung liegen.

Jedes Objekt hat zusätzlich ein eigenes, lokales Koordinatensystem (Objektkoordinatensystem). Dessen Ursprung liegt in dem jeweiligen Objekt. Die Position und Rotation des Objektes in dem globalen Koordinatensystem bestimmt die Relation zwischen dem globalen und dem lokalen Koordinatensystem.

Das lokale Koordinatensystem einer Kamera wird auch Camera Space genannt. Die Relation zwischen dem Camera Space und dem globalen Koordinatensystem wird in Unity durch die cameraToWorld Matrix beschrieben. Mithilfe dieser Matrix kann eine Koordinate aus dem Camera Space in die entsprechende Koordinate des globalen Koordinatensystems transformiert werden.(Unity 2020a)

Dazu wird die Koordinate als Vektor angegeben und mit der cameraToWorld Matrix multipliziert. Das Resultat ist ein Vektor, der eine Koordinate im globalen Koordinatensystem angibt.(Unity 2020a, c)

### 3.6.2 Kamera in 3D-Computergrafik

**View Frustum** ist das Teilvolumen einer 3D Szene, die auf den zweidimensionalen Bildschirm abgebildet wird. Alle Objekte die von der Kamera gesehen werden, befinden sich in dem View Frustum.

**Clipping Plane** bezeichnet eine Ebene, die den View Frustum quer zur Blickrichtung begrenzt. Es gibt eine vordere und eine hintere Clipping Plane. Die vordere Clipping Plane liegt nah an der Kamera. Alle Objekte die zwischen der Kamera und der vorderen Clipping Plane liegen, werden nicht angezeigt.

Die hintere Clipping Plane limitiert wie weit Objekte entfernt sein können, bevor sie nicht mehr zu sehen sind.

## 3.7 Computer Vision

In dem Bereich der Bildverarbeitung gibt es viele Algorithmen, die es ermöglichen Objekte in einem Bild erkennbar zu machen und zu verarbeiten.

### 3.7.1 Object Detection

Objekt Detection ist eine Aufgabe der Computer Vision. Dabei werden Objekte in einem Bild erkannt. Für die Objekte wird eine Klasse und eine Bounding Box bestimmt. Die Klasse gibt an, um welche Art von Objekt es sich handeln. Beispielsweise ob es eine Tastatur oder ein Computerbildschirm ist. Die Bounding Box gibt ein Viereck auf dem Bild an, in dem sich das Objekt befindet.

### 3.7.2 Artificial Neural Networks

Artificial Neural Networks sind Machine Learning Architekturen. Sie können beispielsweise Musik, Text oder Bilder nach Mustern durchsuchen. Sie sind für keine genaue Aufgabe programmiert, sondern lernen indem sie mit Beispieldaten trainiert werden.

Für jedes Beispiel gibt es ein Label, das angibt ob es das gesuchte Muster enthält oder nicht. Die Struktur des Networks verfügt über Gewichte, die Einfluss auf den Output haben. Mit jedem Trainingsbeispiel passt das Network die Gewichte an, sodass der Output dem Label des Beispiels entspricht.(Jiao et al. 2019; O’Shea und Nash 2015)

Artificial Neural Networks bestehen aus einer Menge an verbundenen Knoten, die jeweils eine Berechnung durchführen. Diese Knoten sind in Ebenen aufgeteilt, den Input Layer, den Output Layer, und mehrere Hidden Layer dazwischen. Die Knoten einer Ebene sind mit allen Knoten der Vorherigen Ebene verbunden.(Jiao et al. 2019; O’Shea und Nash 2015)

Das Neural Network bekommt eine Menge an Daten als Input. Die Knoten arbeiten zusammen um den Output zu erzeugen. Dabei wird über Gewichte entschieden, wie viel Einfluss das Ergebnis der einzelnen Knoten auf die nächste Ebene hat.(Jiao et al. 2019; O’Shea und Nash 2015)

Um ein Neural Network zu trainieren, wird der Output von einem Mensch bewertet. Das Neural Network nutzt diese Bewertung, um die Gewichte der einzelnen Knoten zu verändern. So passt sich das Neural Network an. (Jiao et al. 2019; O’Shea und Nash 2015)

### 3.7.3 Convolutional Neural Networks

Convolutional Neural Networks sind auf das Verarbeiten von Bildern spezialisiert. Sie nutzen aus, dass Bilder viele Redundanzen und informationsarme Bereiche haben. Daher können mit jedem Verarbeitungsschritt des Networks Informationen weggelassen werden. So können Rechenzeit und Volumen der Trainingsdaten verringert werden.(Jiao et al. 2019; Jmour et al. 2018; O’Shea und Nash 2015)

Convolutional Neural Networks sind Machine Learning Architekturen, die darauf ausgelegt sind, Muster in Bildern zu erkennen. Sie müssen auf das Muster trainiert werden. Dazu wird ihnen eine Menge an Bildern, die Teilweise das Muster erhalten, und der gewünschte Output, der erreicht werden soll, gegeben. Die Struktur des Network verfügt über Gewichte, die die Berechnung des Outputs beeinflussen. Mit jedem Trainingsbild passt das Network die Gewichte an, damit es die Mustern korrekt erkennen kann.(Jiao et al. 2019; O’Shea und Nash 2015)

Convolutional Neural Networks werden hauptsächlich eingesetzt um Muster in Bildern zu erkennen. Daher ist ihre Struktur und ihre Arbeitsweise auf Bilder spezialisiert. Sie brauchen weniger Rechenzeit und weniger Trainingsdaten als ein generelles Artificial Neural Network für dieselbe Aufgabe brauchen würde.(Jiao et al. 2019; Jmour et al. 2018; O’Shea und Nash 2015)

Die Knoten in einer Ebene eines Convolutional Neural Network sind nur mit wenigen Knoten der vorherigen Ebene verbunden. So sinkt die Menge an Informationen mit jeder Ebene. Das CNN wird gezwungen sich auf wesentliche Teile des Bildes zu konzentrieren, mit

denen beispielsweise ein Objekt oder Muster erkannt werden kann. (Jiao et al. 2019; O'Shea und Nash 2015)

### **3.7.4 Azure Computer Vision**

Microsoft Azure bietet einen Computer Vision Service an. Dabei handelt es sich um mehrere KIs, die für unterschiedliche Aufgaben trainiert wurden. Dazu gehört unter anderem ein Service für Object Detection.

Dabei sendet der Anwender ein Bild an Microsoft, dort wird es verarbeitet und ein Ergebnis zurückgeschickt.(Microsoft, 2018a, 2019a, 2020)

Die Object Detection basiert auf einem trainierten KI Modell. Dieses kann nur Objekte erkennen, für die es trainiert wurde. Zusätzlich können Objekte, die in dem Foto sehr klein sind oder nah bei anderen Objekten liegen, nicht erkannt werden.(Microsoft 2019b)

Der Service ist durch eine REST-API erreichbar. Mit einer Post Anforderung werden die Bilddaten übertragen und die Analyse angefragt. Die Response Nachricht beinhaltet eine Json-Datei, welche die gefundene Objekte und deren Positionen auf dem Foto beinhaltet.

### **3.7.5 Azure Custom Vision**

Azure bietet zusätzlich einen Computer Vision Service an, den der Nutzer trainieren kann. Das verwendete KI Modell ist für Objekt Detection entwickelt und ist nicht vor-trainiert.

Das Trainieren wird über eine Webseite

Custom Vision kann dann verwendet werden, wenn Azure Object Detection für Objekte nicht trainiert ist.(Micosoft 2018)

dass es sich bei einem erkannten Objekt tatsächlich um eine Nivea Dose handelt

todo: was ist die Prediction? Erkäre die Iterations Erkläre was die Genauigkeit der Prediction aussagt.

## 4 Umsetzung

Das Ziel ist das Erkennen und Labeln von Objekten in einer AR Umgebung, durch Image Based Objekt Detection.

### 4.1 Design des Vorgang der Objekt Erkennung

Im Folgenden werden die Arbeitsschritte einer Detection beschrieben.

Wenn der Nutzer das Signal gibt, beginnt die Detection. Als Erstes wird ein Foto mit der Kamera der AR Brille aufgenommen. Dieses Foto wird dann an Azure Object Detection und Azure Custom Vision geschickt. Die Services untersuchen das Foto nach Objekten, geben deren Klasse und Position auf dem Foto an.

Für jedes Objekt soll ein Label erstellt werden, die zeigt wo sich das Objekt in der realen Welt befindet. Dafür wird in der 3D Szene der AR Umgebung eine virtuelle Repräsentation des Fotos erschaffen. Die Fotorepräsentation muss die richtige Skalierung, Position und Rotation haben, um das räumliche Verhältnis zwischen der realen Foto-Kamera und der Umgebung nachzubilden.

Da die Foto-Kamera und das Display nahe beieinander liegen und den gleichen Blickwinkel haben, kann die Position des Displays als Repräsentation des Fotos genutzt werden. In der 3D Szene ist das Display mit der Hauptkamera gleichgesetzt. Die Clipping Plane der Kamera hat somit die gleiche Rotation und eine zumindest ähnliche Position und Skalierung wie das Foto.

Daher werden die Foto-Positionen auf Koordinaten der Clipping Plane abgebildet. Dabei werden verbleibende Positions- und Skalierungs-Unterschiede ausgeglichen. Für jedes Objekte wird so eine Koordinate auf der Clipping Plane bestimmt.

Als Nächstes wird ein Raycast, von der Kamera aus, durch die Clipping Plane Koordinate geschickt. Der Raycast schneidet sich mit einem Mesh, das die reale Welt abbildet. Die getroffene Position wird mit einem Schriftzug markiert. Dort befindet sich das Objekt, das auf dem Foto gefunden wurde.

Alle Objekte, die Azure Object Detection und Azure Custom Vision gefunden haben, werden so für den Nutzer in der AR Umgebung markiert.

### 4.2 Architektur

Magic Leap One übernimmt alle Berechnungen im 3D Raum und führt Spatial Mapping durch. Das Analysieren von 2D Fotos wird an eine REST-API delegiert, da es sehr Speicher und rechenintensiv ist. Die Magic Leap wird als Interaktionsmöglichkeiten für den Nutzer verwendet und zeigt die Ergebnisse der Objekt Detection an. Ergebnisse und Zwischenstände der Objekt Erkennung werden mit einem UI Element angezeigt in der Szene angezeigt. Siehe Abbildung ??

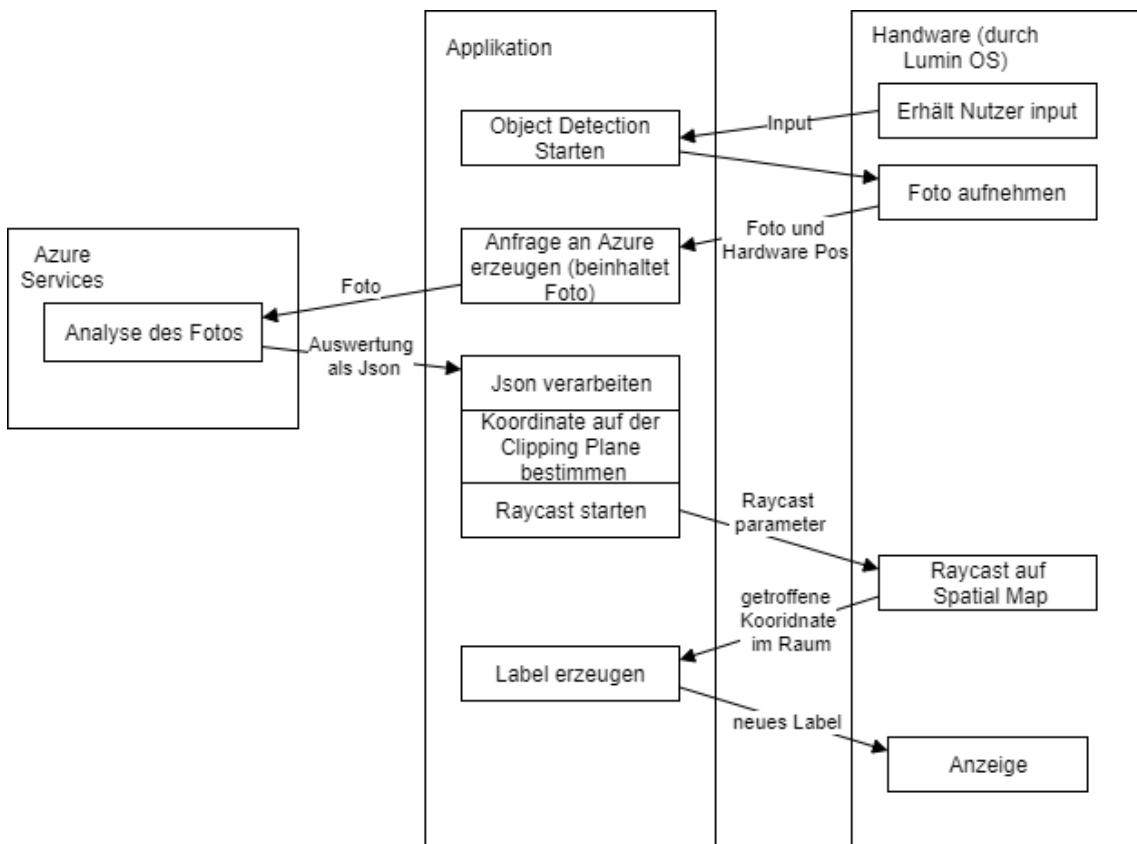


Abbildung 3: Diagramm der Architektur inklusive Bearbeitungsschritte und Informationsweitergabe.

Das Projekt wurde in Unity umgesetzt und für die Magic Leap AR Brille entwickelt. Es wurde ein Unity Projekt Template von Magic Leap verwendet.

Zusätzlich werden einige vorgefertigte Klassen von Magic Leap verwendet. Dazu gehören MLInput, ML Camera, ML Raycast, ML PrivilegeRequestBehavior und ML SpatialMapper. Diese Klassen greifen auf Funktionalitäten des Lumin OS zu.

Die Benötigten Funktionalitäten der Applikation wurden in mehreren Script Klassen umgesetzt. Der Großteil der Scripts verhält sich wie Singletons. Sie existieren nur einmalig in der Szene.

Das Klassendiagramm auf Abbildung 4 zeigt die Scripts und ihre Relationen zueinander.

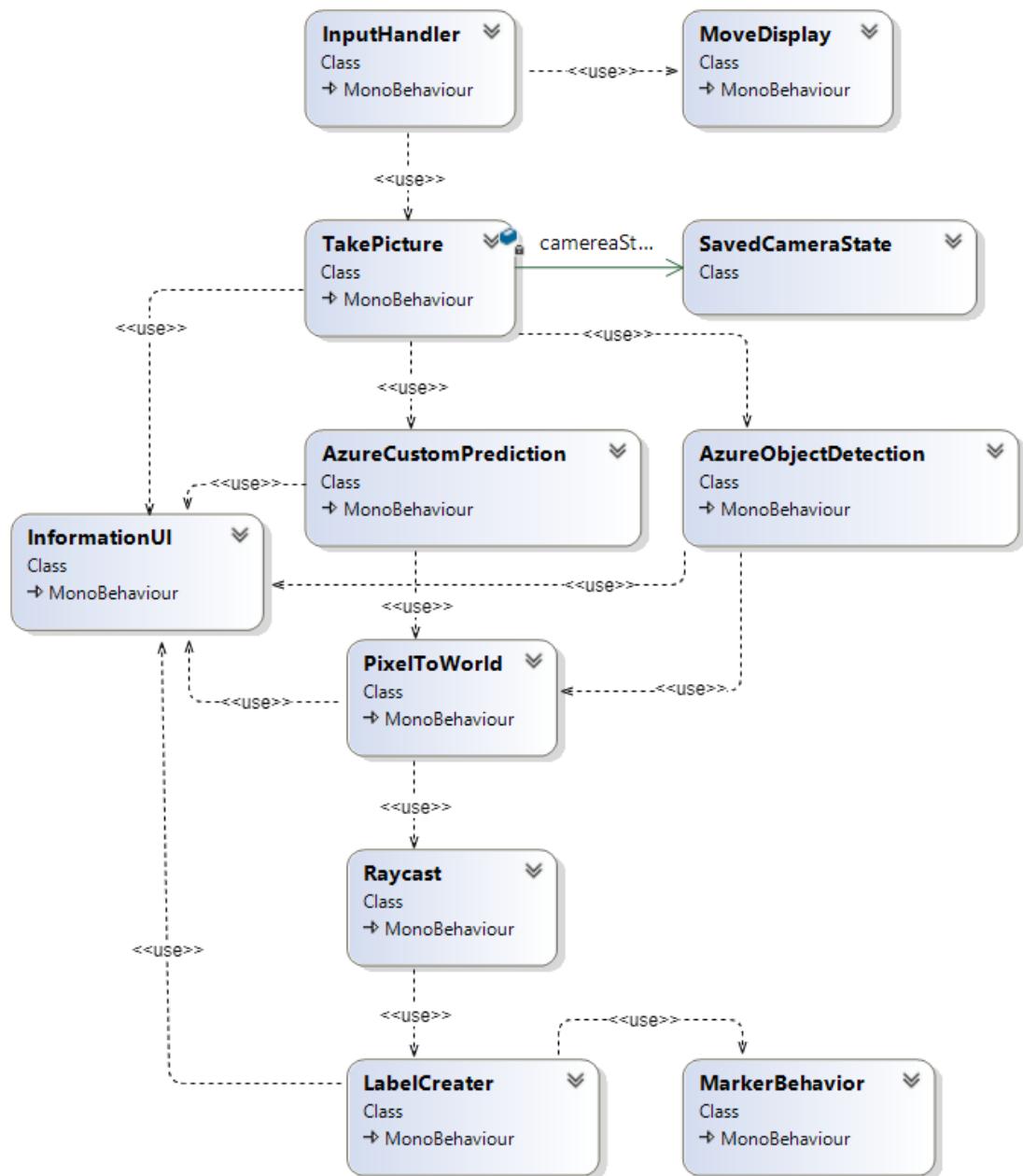


Abbildung 4: Klassendiagramme der Scripte

Die Klassen InputHandler, MoveDisplay, LabelCreator, InformationUI und MarkerBehavior sind für die Interaktion mit dem Nutzer zuständig.

TakePicture, SavedCameraState, AzureCustomPredicton, AzureObjectDetection, PixelToWorld und Raycast führen das Erkennen von Labels von Objekten anhand eines Aufgenommenen Fotos durch. Der Prozess wird durch den InputHander gestartet.

Wurde ein Objekt erkannt und eine Position auf dem Mesh der Umgebung bestimmt, wird der LabelCreator aufgerufen. Es erzeugt das Label mit dem entsprechenden Text.

MarkerBehavior ist eine Script das jedes Label GameObject hat, das erzeugt wird. Über das MarkerBehavior kann der Schriftzug des Labels angepasst werden.

## 4.3 Interaktion

InputHander verarbeitet den Input des Nutzers und startet entsprechende Aktionen durch die MoveDisplay und TakePicture Klassen. MLInput ist eine vorgefertigte Klasse von Ma-

gic Leap. Sie stellt Informationen über den Zustand des Controllers zur Verfügung. InputHander überwacht den Controller und startet die Aktionen, wenn die entsprechende Taste gedrückt wurde.

- Trigger: Objekt Erkennung starten mit TakePicture
- Home Button: UI Element Mittig vor das Display setzen.
- Bumper: Labels verstecken
- Bumper halten: zuletzt erzeugten Label entfernen

Das UI Element wird von InformationUI gesteuert. TakePicture nutzt InformationUI um das zuletzt aufgenommene Foto anzuzeigen. AzureCustomPrediction, Azure ObjectDetection und PixelToWorld dokumentieren ihre Arbeitsschritte mit dem UI Element und der LabelCreator lässt eine Liste aller Labels anzeigen, die in der Szene existieren. Siehe Abbildung ??.

Der LabelCreator ist für das erstellen der Labels verantwortlich und sorgt dafür, dass die Labels für den Nutzer lesbar sind. Dafür werden die Labels in Richtung der Kamera ausgerichtet und mitgeführt. Des Weiteren kann der LabelCreator Labels verstecken und entfernen.

Neben dem UI Element und den Labels wird auch ein Mesh angezeigt, das die Spatial Map der Umgebung wiedergibt. Das Spatial Mapping wird von Lumin OS durchgeführt und das Mesh wird durch die MLSpatialMapper Klasse von MagicLeap erzeugt.



Abbildung 5: Ausgabe

#### 4.4 Implementierung der Objekt Erkennung

Im folgenden werden die Scripts besprochen die für die Objekt Erkennung zuständig sind.

## 4.5 Ein Foto aufnehmen

Das Script TakePicture implementiert das Aufnehmen eines Fotos. Dabei wird ML Camera von MagicLeap genutzt, um die Kamera der Magic Leap Brille anzusteuern. Wenn die Applikation gestartet wird, stellt dieses Script sicher, dass die Applikation Permission hat die Kamera zu nutzen. Dann verbindet sich das Script über MLKamera mit der Kamera Ressource. Die Kamera Ressource wird wieder abgegeben, wenn die Applikation terminiert oder pausiert wird.

Wenn die Methode TakeImage aufgerufen wird, startet der Prozess der Objekt Erkennung. Das aufnehmen der Fotos geschieht asynchron. Für jedes Foto wird ein Thread erzeugt, in dem ML Camera ein Foto aufnimmt. In diesem Thread wird zusätzlich die Aktuelle Position der Unity Kamera als SavedCameraState gespeichert.

Die Methode OnCaptureRawImageComplete wird von ML Camera aufgerufen, wenn das Foto fertig ist. Die Daten des Bildes und der SavedCameraState werden von dort aus an die Scripts AzureObjectDetection und AzureCustomPrediction weitergegeben. Dort wird die Analyse der Bilder gestartet.

### 4.5.1 Object Detection

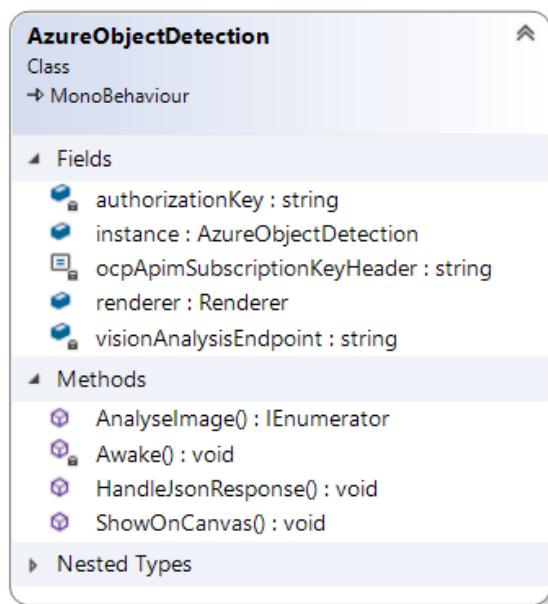


Abbildung 6: Klassendiagramm AzureObjectDetection

In der Methode AnalyselImage von AzureObjectDetection wird ein Web Request zusammengestellt, um die Azure REST API anzufragen. Der Request enthält einen Authentifizierung für die API und das zu analysierende Foto.

Der Webrequest wird verschickt und auf die Antwort gewartet. Wenn die Antwort eintrifft, wird anhand des ResponseCodes geprüft, ob es bei dem Request einen Fehler gab. Beispielsweise kann die Internetverbindung gestört sein oder die Authentifizierung abgelehnt werden. Wenn es keinen Fehler gab, wurde eine Json-Datei bei der Antwort mitgeschickt. Darin wird für jedes gefundene Objekt auf dem Foto eine Bezeichnung (Klasse) und eine Bounding Box angegeben.

Die Json-Datei wird in HandleJsonResponse verarbeitet. Für den erwarteten Aufbau der Datei wurden drei Klassen geschrieben. Der Json String wird mit JsonUtility in ein Detec-

tionResponse Object umgewandelt. Dabei werden alle gefundenen Foto-Objekte in einer Liste von DetectedObjects abgelegt. Siehe Abbildung 7. (Unity 2020b)

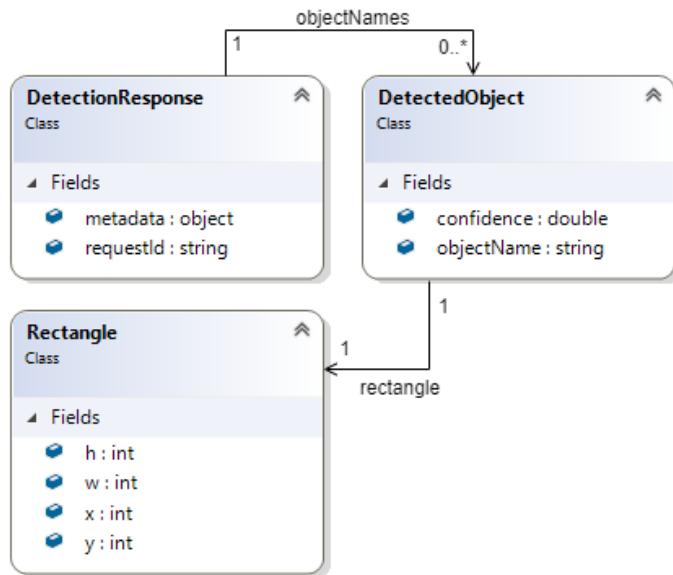


Abbildung 7: Umwandeln der Json Datei in Objekte.

Die gefundenen Objekte sollen im 3D Raum mit einem Label gekennzeichnet werden. Dafür wird für jedes DetectedObject die Methode Cast von der Klasse PixelToWorld aufgerufen. Der Methode wird der Mittelpunkt der BoundingBox als u,v Foto-Koordinate für das DetectedObject übergeben. Siehe Abbildung 8.

```

81     public void HandleJsonResponse(System.String jsonResponse, SavedCameraState cpos)
82     {
83         jsonResponse = jsonResponse.Replace("object", "objectName");
84         //c# doesn't like "public string object"
85         DetectionResponse det = new DetectionResponse();
86         det = JsonUtility.FromJson<DetectionResponse>(jsonResponse);
87         InformationUI.instance.Show(" Handle Json");
88         foreach (DetectedObject obj in det.objectNames)
89         {
90             Debug.Log(obj.objectName);
91             int x = obj.rectangle.x + (obj.rectangle.w / 2);
92             int y = obj.rectangle.y + (obj.rectangle.h / 2);
93             InformationUI.instance.Add(obj.objectName);
94             PixelToWorld.instance.Cast(x, y, cpos, obj.objectName);
95         }
96     }
  
```

Abbildung 8: Umwandeln der Json Datei in Objekte.

#### 4.5.2 Von dem Foto zum 3D Raum

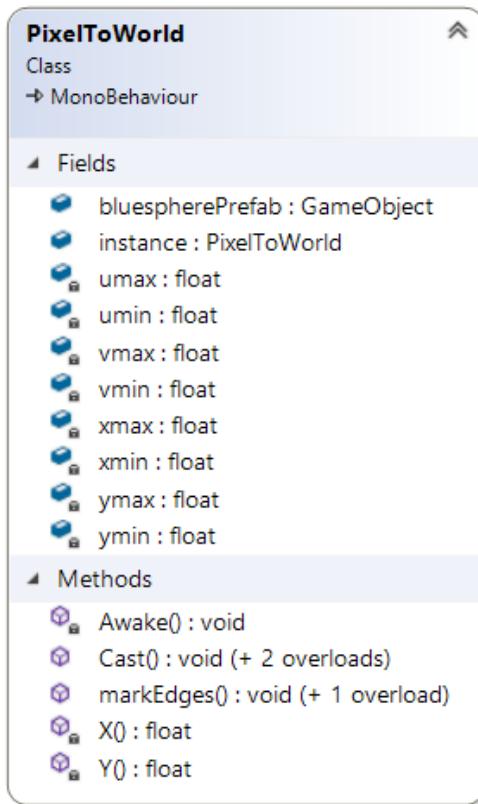


Abbildung 9: Klassendiagramm PixelToWorld

Ein gefundenes Foto-Objekt soll in der 3D Abbildung der realen Welt lokalisiert werden. Dafür nutzt die Methode Cast die u,v Foto-Koordinate des Objekts und einen SavedCameraState. Der SavedCameraState beschreibt die Positiong der Unity Kamera zu dem Zeitpunkt als das Foto aufgenommen wurde. SavedCameraState beinhaltete die cameraToWorldMatrix und den Ursprung der Kamera.

Das Foto kann mit dem Display und somit mit der Clipping Plane der Hauptkamera approximiert werden. Die u,v Foto-Koordinate wird zunächst in eine x,y,z Koordinate in dem Camera Space umgewandelt. Der z Anteil gibt die Entfernung von dem Ursprung der Kamera an in Blickrichtung an. Dabei befinden sich Punkte mit einer Entfernung von 0.4 Einheiten auf der Clipping Plane. In dem Camera Space mit  $z = -0.4$  angegeben.

Die x und y Dimensionen beschreiben die Achsen, die horizontal und vertikal zur Clipping Plane verlaufen. Mit dem festgelegten  $z = -0.4$ , kann jeder Punkt auf der Clipping Plane durch x und y angegeben werden. Dazu gehören auch Punkte die außerhalb des View Frustum liegen.

Es wurden Werte für x und y ausprobiert, mit denen die Ränder des Fotos auf der Clipping Plane angegeben werden können. Dabei wurde auf die unterschiedlichen Seitenverhältnisse des Fotos und des Displays geachtet. Darüber hinaus ist der Bildausschnitt des Displays kleiner. Daher liegen die Ränder des Fotos außerhalb des View Frustum.

Sind diese x und y Werte bekannt, ergibt sich für die Achsen jeweils ein Intervall, die kombiniert alle Foto-Koordinaten auf die Clipping Plane abbilden können. Die Intervall lauten:  $[-0.295, 0.2281]$  für x und  $[0.1546, -0.1507]$  für y. Mit den Intervallen wird die Position und Skalierung des Fotos in Relation zu dem Display - und der Hauptkamera - berücksichtigt.

Siehe Kapitel 4.6 für die Entwicklung der Cast Methode und die Ermittlung der Intervallwerte.

Es werden zwei lineare Funktionen aufgestellt:

- Die Funktion X bildet das Intervall für u [0,1920] auf das Intervall für x [-0.295,0.2281] ab.
- Die Funktion Y bildet das Intervall für v [0,1080] auf das Intervall für y [0.1546,-0.1507] ab.

Siehe Abbildung 10.

```

40 //Picture u and v ranges
41 private float umin = 0;//left
42 private float umax = 1920;//right
43 private float vmin = 0;//up
44 private float vmax = 1080;//down
45 //Offset Vektor x and y ranges
46 private float xmin = -0.295F;//left
47 private float xmax = 0.2281F;//right
48 private float ymin = 0.1546F;//up
49 private float ymax = -0.1507F;//down
50 private float X(float u)
51 {
52     float slope = ((xmax - xmin) / (umax - umin));
53     float b = xmin - slope * umin;
54     return slope * u + b;
55 }
56 private float Y(float v)
57 {
58     float slope = ((ymax - ymin) / (vmax - vmin));
59     float b = ymin - slope * vmin;
60     return slope * v + b;
61 }
```

Abbildung 10: Funktionen X und Y

Mit den Funktionen wird eine Koordinate im Camera Space für u,v berechnet. Diese Koordinate wird dann, mithilfe der cameraToWorldMatrix des SavedCameraState, in die Koordinate p des globale Koordinatensystem umgewandelt. Damit wird die Position und Rotation der Kamera - und somit des Fotos - in der 3D Szene berücksichtigt. Siehe Abbildung 11.

```

28 public void Cast(float u, float v, SavedCameraState cpos, GameObject clippingPlaneMarker,
29                     string objectName, bool showClippingPlane, int material)
30 {
31     //scale v,v to x,y range
32     Vector3 offset = new Vector3(X(u), Y(v), -0.4F);
33     Vector3 p = cpos.cameraToWorldMatrix.MultiplyPoint(offset);
34     Raycast.instance.StartCast(Raycast.instance.CreateRaycastParams(cpos.ctransform, p), objectName, material);
35     if (showClippingPlane)
36     {
37         GameObject sphere2 = Instantiate(clippingPlaneMarker, p, Quaternion.identity); // show point on clipping
38     }
39     ResultAsText.instance.Add(u + " " + v + " " + X(u) + " " + X(v) + " " + objectName + " object marked");
40 }
```

Abbildung 11: Cast Methode

#### 4.5.3 Raycast

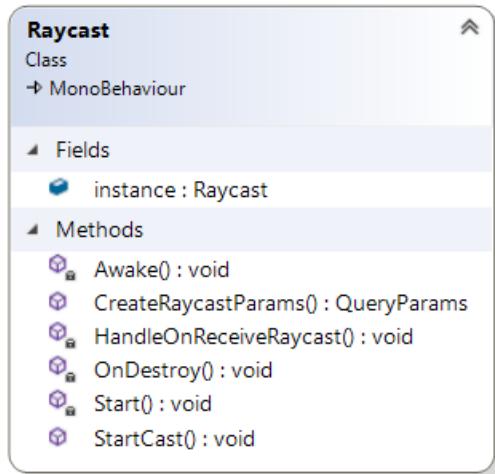


Abbildung 12: Klassendiagramm Raycast

Als Nächstes wird ein Raycast durch den Ursprung der Kamera und die Koordinate p gesendet. MLRaycast wird genutzt, um einen Schnittpunkt mit der Rekonstitution der Welt von Lumin OS zu bestimmen. Die Stelle, die der Raycast trifft beschreibt die Position des DetectedObject im 3D Raum.

Für den MLRaycast werden zwei Parameter benötigt:

- Ein QueryParams Objekt, das Ursprung und Richtung für den Raycast beinhaltet.
  - Ursprung: Kameraursprung aus SavedCameraState
  - Richtung: Richtungsvektor von dem Kameraursprung zu der Koordinate p
- Eine Methode die aufgerufen wird, wenn der Raycast fertig ist.
  - Callback Methode: HandleOnRecieveRaycast

Siehe Abbildung 13.

```

19     public MLRaycast.QueryParams CreateRaycastParams(Transform ctransform, Vector3 target)
20     {
21         MLRaycast.QueryParams _raycastParams = new MLRaycast.QueryParams
22         {
23             // Update the parameters with our Camera's transform
24             Position = ctransform.position,
25             Direction = target - ctransform.position,
26             UpVector = ctransform.up,
27             // Provide a size of our raycasting array (1x1)
28             Width = 1,
29             Height = 1
30         };
31         return _raycastParams;
32     }

```

Abbildung 13: Cast Methode

Wenn der Raycast fertig ist, wird die Methode HandleOnRecieveRaycast aufgerufen. Der Parameter point beinhaltet dabei die getroffene Koordinate. Diese wird an die Methode CreateMarker von der Klasse LabelCreater weitergegeben.

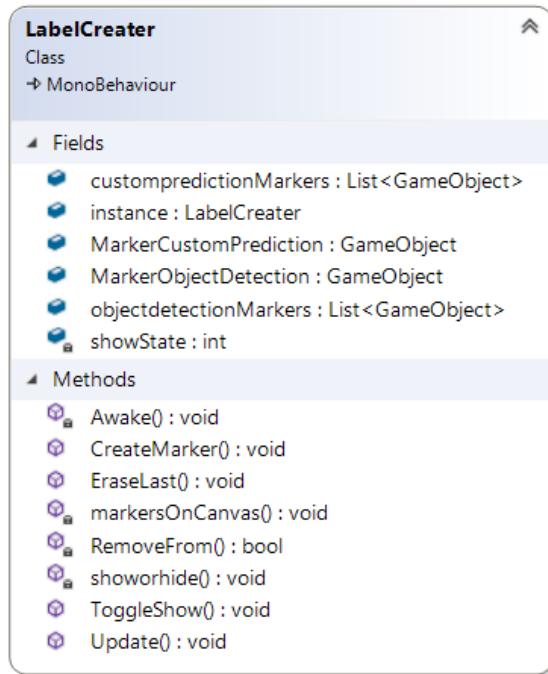


Abbildung 14: Markierungen in der Welt

#### 4.5.4 LabelCreator

CreateMarker erhält den Punkt point, der getroffen wurde und die Bezeichnung für das DetectedObject. An der Koordinate von point wird ein Prefab GameObject instanziert, das als Markierung für das DetectedObject in der 3D Umgebung dient.

Das Prefab besteht aus einer Kugel und einem Schriftzug, der den Namen des DetectedObject anzeigen soll. Dem neu instanzierten GameObject wird die Bezeichnung des DetectedObject als Schriftzug zugewiesen. Siehe Abbildung 15.

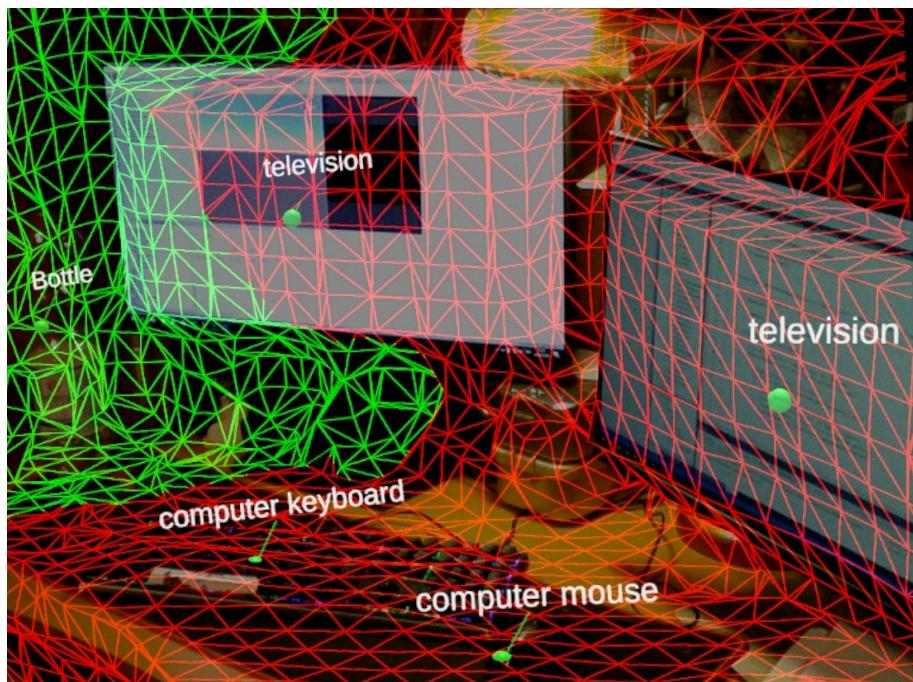


Abbildung 15: Labels in der Welt

Wenn ein Label nahe eines anderen Labels erzeugt werden soll, das denselben Schriftzug hat, wird davon ausgegangen, dass ein Objekt der Realen Welt zum zweiten mal erkannt wurde. Daher wird kein neues Label erstellt, sondern das alte Label modifiziert. Die Positionen an denen das Objekt in der Szene lokalisiert wurde werden mit jeweils einer grauen Sphäre markiert und das Label wird in den Mittelpunkt der grauen Sphären gesetzt. So wird die Position des Objektes genauer, wenn es häufiger Erkannt wurde. Siehe Abbildungen 16 und 17.

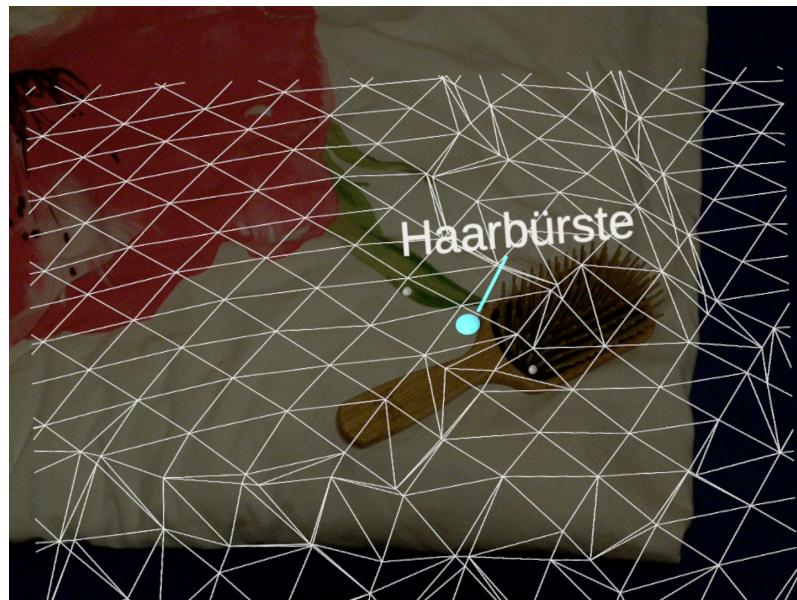


Abbildung 16: Haarbürste zwei mal Erkannt

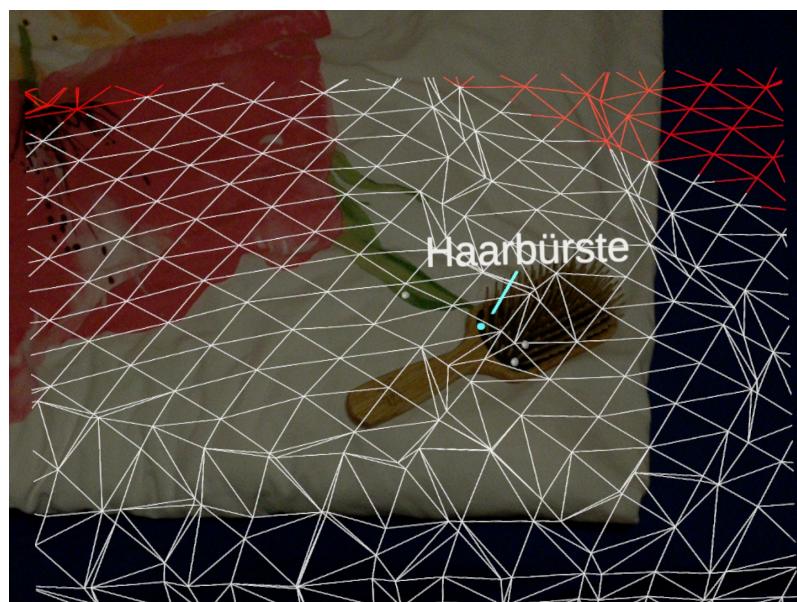


Abbildung 17: Haarbürste drei mal Erkannt

Die erzeugten Labels werden in den Listen `MarkerCustomPrediction` und `MarkerObjectDetection` gespeichert, je nachdem welcher Azure Service verwendet wurde. Die Listen werden genutzt, um die Labels zu der Unity Kamera auszurichten damit sie lesbar sind.

#### 4.5.5 Azure Custom Vision

Neben der Bildanalyse mit Azure Object Detection wird auch Azure Custom Vision verwendet. Die AI wurde über die Webseite trainiert.

Die Anfrage an den Service passiert in der Klasse AzureCustomPrediction. Ähnlich wie bei AzureObjectDetection wird ein Webrequest erstellt mit einem Authorization Key für den Service und einem Foto als Payload.

In der Antwort wird eine Json Datei zurückgeschickt, die die gefundenen Objekte angibt. Da die Json Datei ein etwas anderes Format hat, wurde eine eigene HandleJsonResponse Methode dafür geschrieben.

Für jedes erkannte Objekt wird die Methode Cast von PixelToWorld aufgerufen, um das Objekt in der realen Welt zu lokalisieren und zu markieren.

**Das Trainieren** Es wurde probiert das Custom Vision Modell auf drei unterschiedliche Objekte zu trainieren. Dabei wurden vier Iterationen erstellt.

Iteration 1:

Zunächst wurde probiert Tuben von Acrylfarbe zu erkennen. Die Genauigkeit davon war nicht sehr hoch. Es wurden in Fotos Acrylfarben an Stellen erkannt, an denen es keine gab. Siehe Abbildung 18.

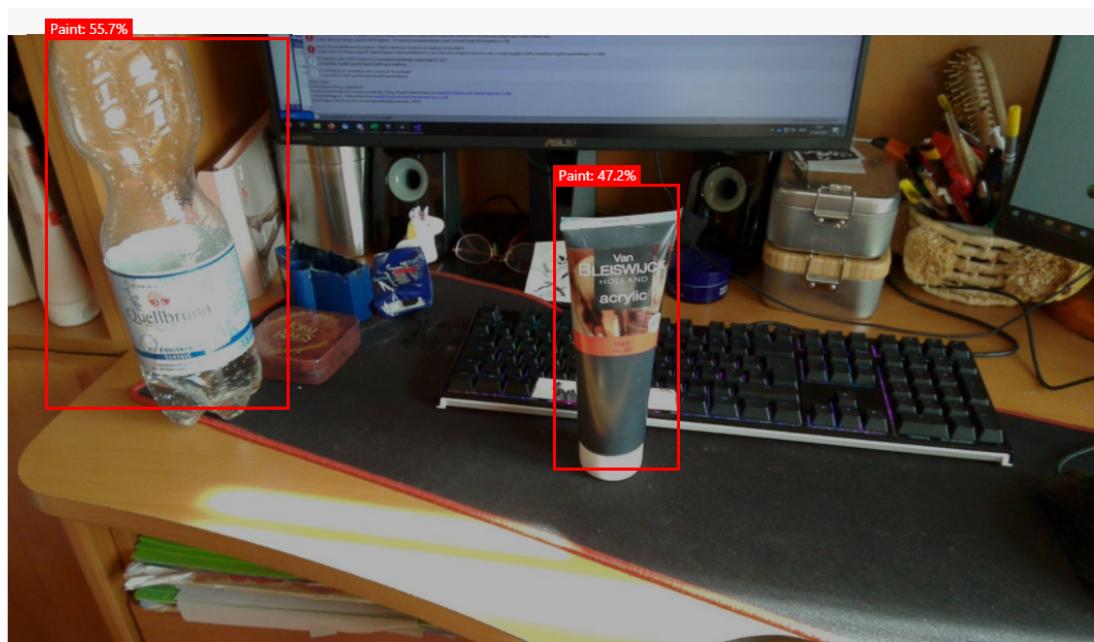


Abbildung 18: Beispieldfoto Iteration 1. Die Wasserflasche wurde als Farbe markiert mit 55.7 Prozent statistischer Konfidenz. Die tatsächliche Farbtube wurde mit einer Konfidenz von 47.5 Prozent erkannt.

**Iteration 2** In der zweiten Iteration wurde probiert das Modell darauf zu trainieren, eine blaue Dose von Nivea Hautcreme zu erkennen. Die Form und Farbe der Dose ist sehr simpel, daher wurde davon ausgegangen, dass sie leichter zu erkennen ist. Die berechnete Prediction Wahrscheinlichkeit während des Trainings lag bei 80 Prozent.

Trotzdem wurden in vielen Fotos fälschlicherweise Nivea Dosen erkannt.

**Iteration 3** In der dritten Iteration wurde versucht die vorherige Iteration zu verbessern. Es wurden ausgewählte Trainingsfotos entfernt, die die Dose von einem seitlichen Winkel zeigten. Die Erwartung war, dass die Detektion der Dose aus dem Blickwinkel von Oben konsistenter wird. Zusätzlich wurden mehr Fotos von der Dose auf unterschiedlich gefärbten und gemusterten Untergründen hinzugefügt.

Die Genauigkeit der Prediction sank auf 75 Prozent.

**Iteration 4** In der vierten Iteration wurden zwei Fotos von der Nivea Dose entfernt, was die Genauigkeit auf 100 Prozent steigen ließ. In der Umsetzung mit der Magic Leap Anwendung wurden trotzdem häufig Objekte fälschlicherweise als Nivea Dose markiert.

Neben der Dose wurde diese Iteration darauf trainiert eine bestimmte Holzhaarbürste zu erkennen. Aufgrund von dem komplexeren, und markanten Aussehen der Bürste wurde davon ausgegangen, das die Bürste besser von anderen Objekten zu unterscheiden ist. Die Bürste wurde nur mit den Borsten nach oben fotografiert.

Die Genauigkeit für die Bürste lag bei 100 Prozent. In der Umsetzung mit der Magic Leap Anwendung wird die Bürste häufig nicht erkannt, obwohl sie im Bild ist und mit den Borsten nach oben liegt. Es werden jedoch keine Objekte fälschlicherweise als Haarbürste erkannt.

#### 4.6 Entwicklung der Foto-Repräsentation

Um die u,v Foto-Koordinate eines gefundenen Objektes auf der Clipping Plane der Kamera zu lokalisieren, wurden ein paar Herangehensweisen ausprobiert.

Das Ziel ist das Setzen einer Markierung in dem 3D Raum, basierend auf der Foto-Koordinate. Das Foto beinhaltet keine Information über die Entfernung zu dem Objekt. Dafür muss ein Raycast durchgeführt werden.

Mit einer Repräsentation des Fotos in dem 3D Raum ist es möglich diesen Raycast durchzuführen. Dazu muss das Foto nicht tatsächlich in dem 3D Raum vorhanden sein. Es muss jedoch mit dem Input einer Foto-Koordinate ein Output einer Koordinate in dem 3D Raum erzeugt werden, mit dem der Raycast durchgeführt werden kann.

Die Position des Fotos hängt mit der Kamera zusammen, daher kann das Foto durch den Camera Space simuliert werden. Als erstes wurde probiert ein Sphären-Objekt an eine gezielte Koordinate des Camera Space zu bewegen.

Wenn die Kamera am Ursprung des globalen Koordinatensystem liegt und eine neutrale Rotation hat, stimmt der Camera Space mit dem globalen Koordinatensystem überein. Die Sphäre wurde in der Szene per Hand bewegt um markante Koordinaten des Camera Space abzulesen. Siehe Abbildung 19.

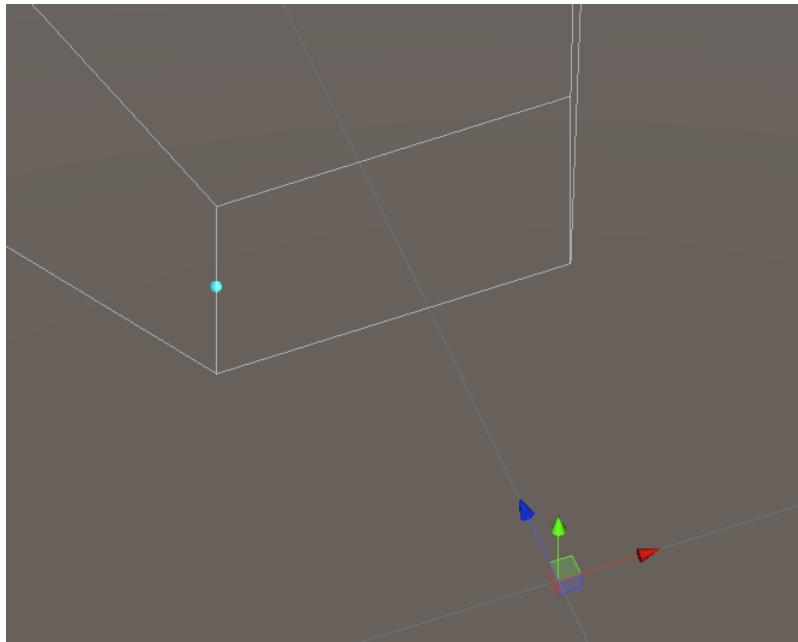


Abbildung 19: Die Blaue Sphäre liegt auf dem Rechten Rand der Clipping Plane.

Dabei wurden folgende Camera Space Koordinaten gefunden:

- Near Clipping Plane bei  $z = -0.37$
- linker Rand bei  $x = -0.153$
- rechter Rand bei  $x = 0.153$
- oberer Rand bei  $y = 0.1147$
- unterer Rand bei  $y = -0.1147$

Die x und y Koordinaten hängen von der u,v Koordinate des Fotos ab. Es wurden lineare Funktionen aufgestellt um u,v auf x,y abzubilden. Diese Abbildung dient als Repräsentation des Fotos im 3D Raum, unter Berücksichtigung der Position und Skalierung des Fotos im Verhältnis zu der Kamera.

Dann wurde getestet wie genau DetectedObjects in der AR Umgebung lokalisiert werden. Es wurden testweise Fotos aufgenommen, analysiert und die DetectedObjects markiert. Die entstandenen Markierungen lagen in Sichtfeld, jedoch nicht an den erwarteten Stellen.

Um dem Problem auf den Grund zu gehen, wurde ein UI Objekt erstellt, das ein aufgenommenes Foto bei Runtime anzeigt. Das Foto wurde dann mit dem Display verglichen. Dabei fiel auf, dass sie ein unterschiedliches Seitenverhältnis haben und das Display einen kleinen Bildausschnitt zeigt.

Es gibt zwei Möglichkeiten die Unterschiede zwischen Foto und Display auszugleichen. Entweder wird das Foto auf das Display zugeschnitten oder das gesamte Foto wird verwendet. Im zweiten Fall würden auch Objekte erkannt, die außerhalb des Sichtfeldes liegen. Es wurde die Entscheidung getroffen das Foto zuzuschneiden. Damit gibt es ein besseres Feedback für den Nutzer, wenn ein Objekt gefunden wurde.

Das Zuschneiden wurde realisiert, indem die Intervalle für u und v der Abbildungsfunktionen stärker eingegrenzt wurden. Alle Objekte die außerhalb der Intervalle liegen werden ignoriert. Um die Intervalle zu bestimmen wurde dem Fotoanzeige-UI-Element ein Gitter hinzugefügt. Mit dem Gitter kann die u,v Position von beliebigen Stellen des Fotos abgelesen werden. Durch Aufnehmen von Fotos und Vergleichen mit dem Sichtfeld des Displays wurde abgelesen, bei welcher u,v Position des Fotos die Ecken des Displays zu finden sind. Die Intervalle wurden dem entsprechend eingegrenzt.

Mit den durchführten Veränderungen der Intervalle konnten DetectedObjects korrekt in der Umgebung lokalisiert werden. Jedoch wurden sehr häufig Objekte nicht markiert, obwohl sie im Sichtfeld des Nutzers lagen, weil deren Mittelpunkt außerhalb eines Intervalls lag.

Daher wurde entschieden die zweite Möglichkeit zu implementieren und das gesamte Foto zu verwenden und Objekte auch zu markieren, wenn sie komplett außerhalb des Sichtfeldes liegen. Dafür wurden die Intervalle für u und v wieder auf die ursprünglichen Werte - [0,1920] und [0,1080] - gesetzt. Die Intervalle für x und y mussten vergrößert werden.

Um die x und y Intervalle bestimmen zu können, wurde das Fotoanzeige-UI-Element Parallel zu der ClippingPlane gelegt. Das Element folgt den Bewegungen der Kamera und liegt möglichst nah an der Near Clipping Plane. Das Display der Magic Leap Brille zeigt selbst solide Objekte leicht durchsichtig an. Das wurde genutzt, um Fotos aufzunehmen, mit dem UI Element anzuzeigen und mit der realen Welt zu vergleichen. Durch Ausprobieren wurde das UI Element so skaliert und verschoben, dass das angezeigte Foto mit der realen Welt soweit wie möglich übereinstimmt. Siehe Abbildungen 20 und 21.

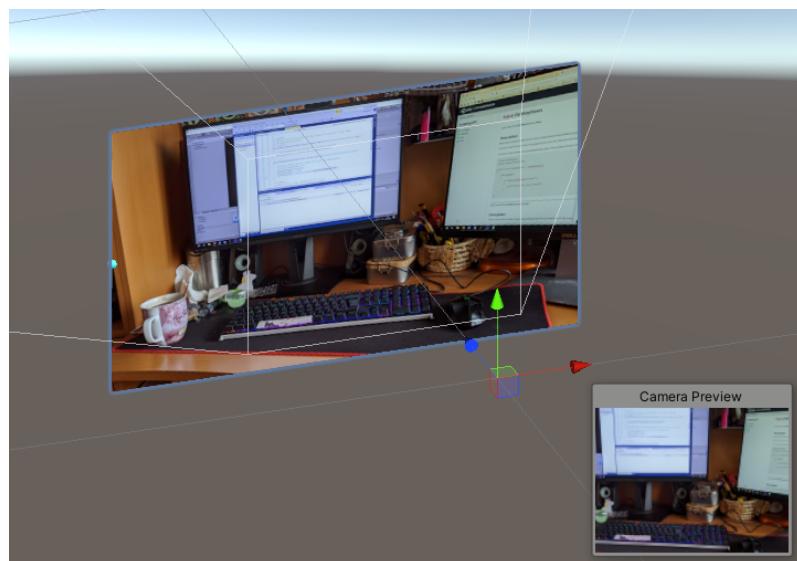


Abbildung 20: Das Aufgenommene Foto füllt das gesamte Display aus, wenn es angezeigt wird. Die Blaue Sphäre liegt auf dem Rechten Rand des Foto-Anzeige-Elementes.

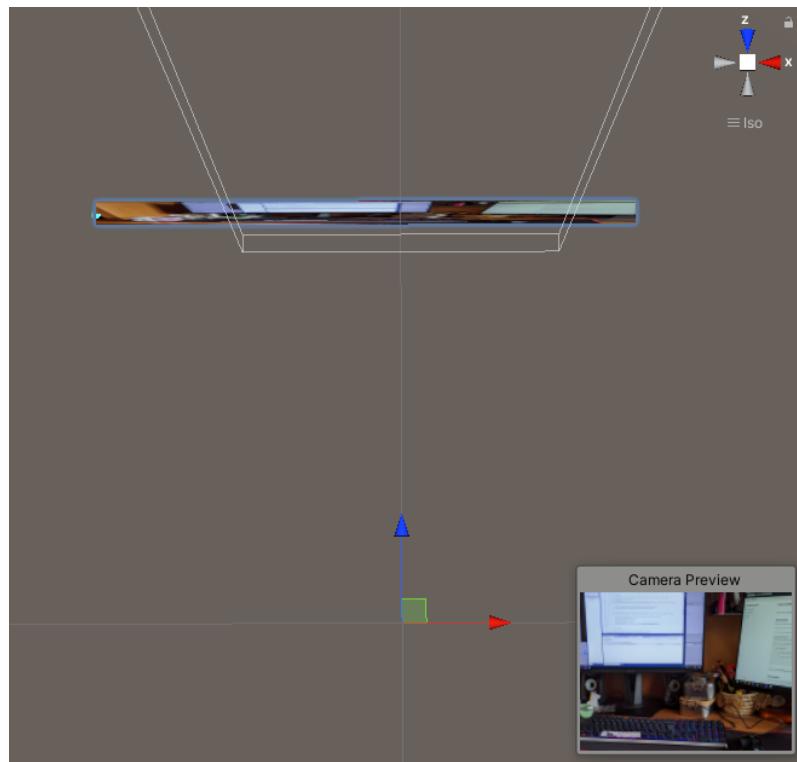


Abbildung 21: Die Blaue Sphäre befindet sich nicht mehr in dem View Frustum und das Foto-Anzeige-Element befindet sich ein wenig hinter der Near Clipping Plane.

Dann wurden die Ränder des UI Elementes genutzt um die Intervalle für x und y zu bestimmen.

- für x: [-0.295, 0.2281]
- für y: [0.1546, -0.1507]
- Zusätzlich wurde z = -0.4 gesetzt. Das UI Element musste ein wenig weiter von der Clipping Plane entfernt sein um angezeigt zu werden.

Mit diesen Intervallen für u,v,x und y konnten DetectedObjects gut lokalisiert werden und es wurden keine Objekte mehr weggelassen, von denen der Nutzer erwarten würden, dass sie markiert werden.

## 5 Auswertung

in welchem raum getestet. wie groß sind die fotos?

### 5.1 Laufzeitanalyse

Die Laufzeit von dem Aufnehmen und Analysieren eines Fotos bis hin zur Label Erstellung in der Szene wurde aufgezeichnet. Siehe Abbildung 22

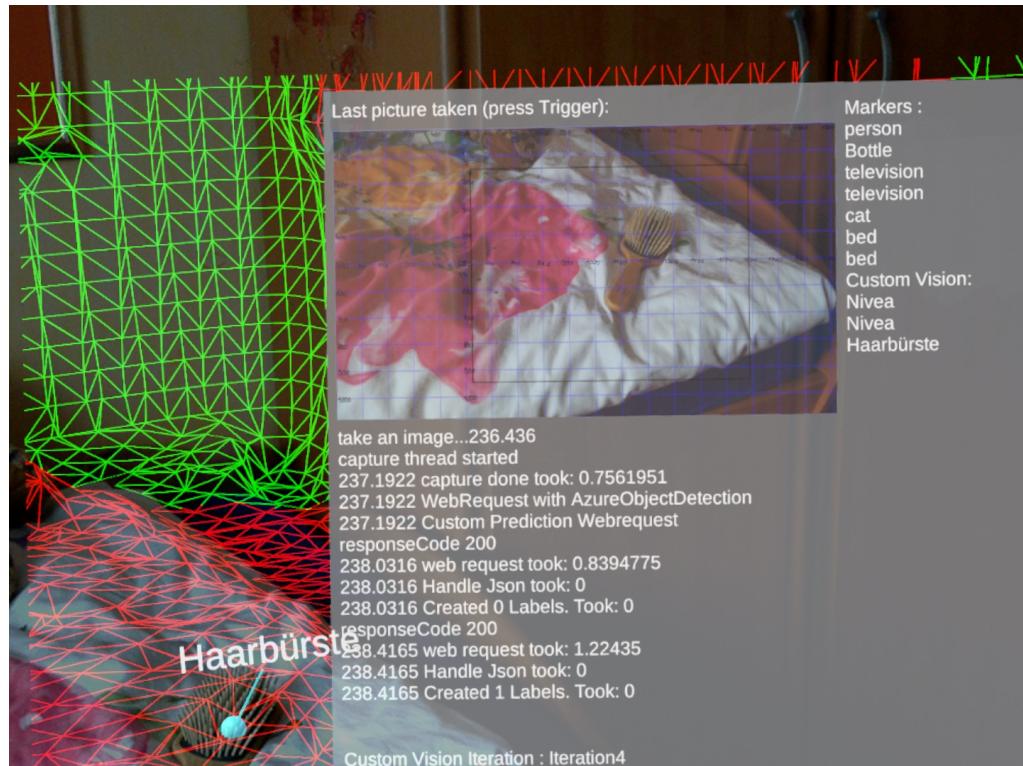


Abbildung 22: Durchlauf mit Laufzeit Aufzeichnung

Über 11 Aufgenommene Fotos wurden gefunden, das das Aufnehmen des Fotos im Durchschnitt 0,9 Sekunden dauert. Das Anfragen des Azure Object Detection Services, inklusive Netzwerk Response Time liegt durchschnittlich bei 1,01 Sekunden. Die Anfrage an den Azure Custom Vision Service, inklusive Netzwerk Response Time, dauert im durchschnitt 1,28 Sekunden.

Das auslesen der Json Antworten, lokalisieren der Objekten in der 3D Szene und das Label erstellen, benötigt weniger als eine Mikrosekunde.

Insgesamt dauert das Aufnehmen und Analysieren eines Fotos somit durchschnittlich 3,19 Sekunden. Siehe Abbildung 23 und 24.

	Foto 1	Foto 2	Foto 3	Foto 4	Foto 5	Foto 6	Foto 7	Foto 8	Foto 9	Foto 10	Foto 11	Average
image capture	0,82	1,18	0,89	0,75	1,57	0,80	0,74	0,89	0,77	0,75	0,76	0,90
obj det web request	0,86	0,84	1,11	1,30	1,04	1,22	1,12	0,93	0,95	0,84	0,84	1,01
handle json	0	0	0	0	0	0	0	0	0	0	0	0,00
created labels	0	0	0	0	0	0	0	0	0	0	0	0,00
custom v web request	1,38	1,18	1,42	1,30	1,04	1,28	1,12	1,54	1,34	1,27	1,22	1,28
hanlde json	0	0	0	0	0	0	0	0	0	0	0	0,00
created labels	0	0	0	0	0	0	0	0	0	0	0	0,00
<b>total</b>	<b>3,06</b>	<b>3,20</b>	<b>3,42</b>	<b>3,36</b>	<b>3,65</b>	<b>3,30</b>	<b>2,98</b>	<b>3,35</b>	<b>3,06</b>	<b>2,86</b>	<b>2,82</b>	<b>3,19</b>
Objekte durch obj det	4	3	0	0	2	0	2	4	1	2	0	1,64
Objekte durch custom	2	2	0	0	0	0	1	2	0	0	1	0,73

Abbildung 23: Laufzeitanalyse über 11 Bild-Analysen.

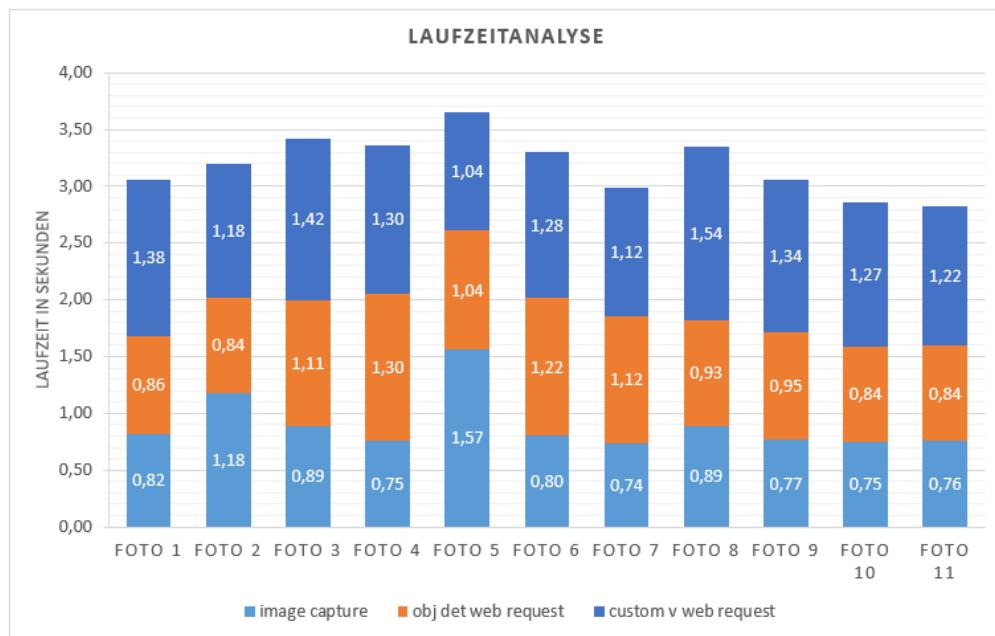


Abbildung 24: Diagramm Laufzeitanalyse

Gedanken dazu: Für die Laufzeit ist es besser ein einziges Network zu nehmen. Und das Image Capture so lange dauert ist eigentlich unacceptable. Das sollte schneller gehen. Dadurch statt einzelne Fotos aufzunehmen ein continuous video aufnehmen und frames davon zu analysieren wäre besser. Ist auf jeden Fall nicht real time.

Durchschnittlich wurden durch Objekt Detection 1.64 Objekte pro bild erkannt und durch custom detection 0.73.

## 5.2 Analyse durch Azure Objekt Detection

Wie sieht der Raum aus? Welche Objekte wurden erkannt. Wie wurden sie misinterpretiert. Objekte erkennen auch wenn sie nicht komplett im Screen sind.

## 5.3 Azure Custom Vision

Kommt halt darauf an wie es trainiert ist.

#### 5.4 Objekte in 3D Szene Lokalisieren

generell gut. Durch mehrere detectionen wird lokation verbessert, besonders wenn durch leicht unterschiedliche Blickwinkel. Aber auch bei nur einer Erkennung ist die Position in der 3D Szene ganz gut.

spatial mapping braucht manchmal ne weisse, das dazu führt, das das label nicht an der richtigen stelle ist. passiert bei objekten die sich oft bewegen, wie ein stuhl. Siehe Abbildung 25.

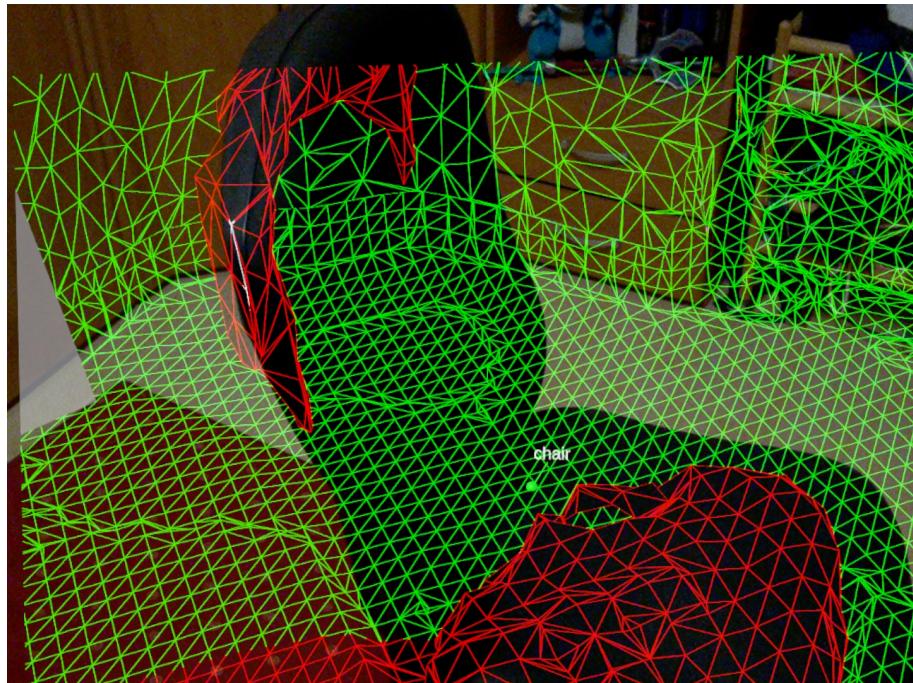


Abbildung 25: Laufzeitanalyse

Halb Trzparente Objekte sind auch eine challenge für sptaial mapping. werden nicht richtig gemapped. Siehe Abbildung 26.

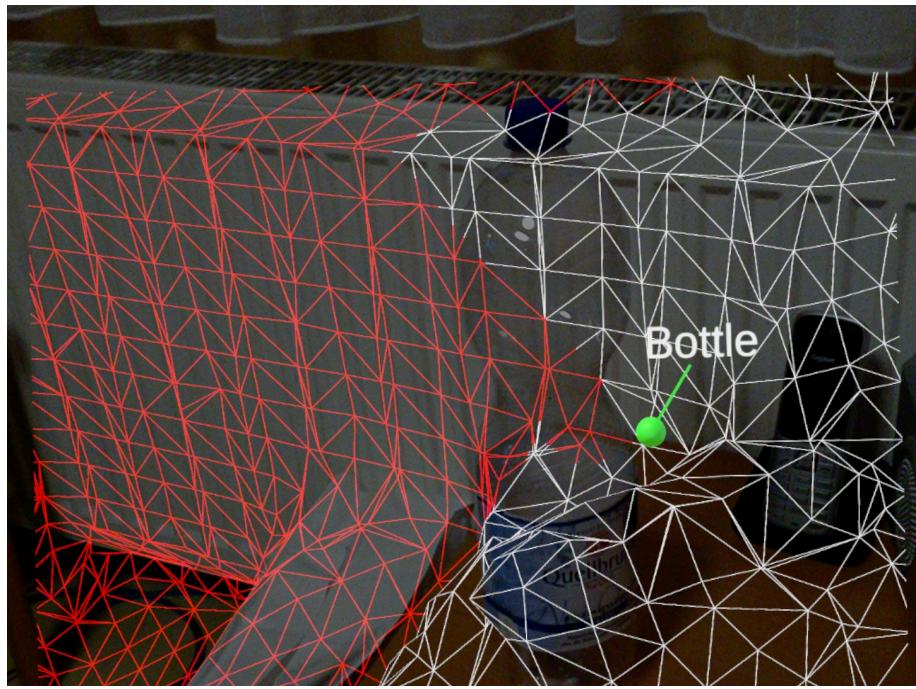


Abbildung 26: Label für Bottle liegt hinter dem Tatsächlichen Objekt. Die Flasche wurde nicht korrekt gemapped.

## 6 Zusammenfassung

In diesem Kapitel werden die bearbeiteten Themen kurz zusammengefasst.

### 6.1 Konzepte und Impelemntierte Funktionen

In dieser Arbeit wird Image based Objekt Detection in eine Augmented Reality Brille integriert, um eine automatisches Objekt Erkennen und Labeln zu ermöglichen.

Es werden Fotos von der Umgebung der AR Brille aufgenommen und mit einer Machine Learning Modell analysiert. Das Modell ist darauf trainiert Objekte und Lebewesen in einem Bild zu erkennen. Objekte auf den Fotos der Umgebung werden so auf den Fotos erkannt. Dann werden sie in der AR Umgebung lokalisiert und mit einem Label Markiert. Das automatische der Erkennen und Labeln von Objekten kann für große und dynamische Umgebungen verwendet werden. Es bietet somit eine gute Grundlage für AR Anwendungen die einer solchen Umgebung arbeiten oder eine ausgeprägteres Semantisches Verständnis der Umgebung benötigen. Beispielsweise eine Blindenführung in unbekannten Umgebungen, der Bereich des Autonomous Driving und Robotic.

### 6.2 Ausblick

Die Automatische Erkennung von Objekten in einer AR Umgebung ist funktionsfähig und effektiv.

Die Objekte, die erkannt werden hängen von dem verwendeten Machine Learning Modell ab. Es gibt die Arten der Objekte und die Genauigkeit der Detection an.

In Zukunft könnte man weitere Modelle verwenden um die Fotos der Umgebung zu analysieren. Beispielsweise können Kontextinformationen extrahiert werden. Daraus kann hervorheben, in welchem Raum eines Hauses die AR Brille sich befindet (Küche, Arbeitszimmer, Schlafzimmer).

Um die Lokalisierung eines Objektes in der AR Umgebung zu verbessern, könnte zusätzlich zu den Daten des Spatial Mappings, die Tiefenkamera der AR Brille verwendet werden. Die Spatial Map zu erstellen ist Arbeitsaufwändig. Das kann dazu führen Stellenweise die Map noch nicht aufgebaut ist, oder in einem Veralteten Zustand vorliegt, wenn ein Objekte lokalisiert werden soll. Die rohen Daten der Tiefenkamera könnten dazu verwendet werden, das Spatial Map zu ergänzen und zu korrigieren.

## 7 Literaturverzeichnis

### Azuma und Furmanski 2003

AZUMA, R. ; FURMANSKI, C.: Evaluating label placement for augmented reality view management. In: *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings.*, 2003, S. 66–75

### Bell et al. 2001

BELL, Blaine ; FEINER, Steven ; HÖLLERER, Tobias: View Management for Virtual and Augmented Reality. In: *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology*. New York, NY, USA : Association for Computing Machinery, 2001 (UIST '01). – ISBN 158113438X, 101–110

### Chen et al. 2018

CHEN, Long ; TANG, Wen ; JOHN, Nigel ; WAN, Tao R. ; ZHANG, Jian J.: *Context-Aware Mixed Reality: A Framework for Ubiquitous Interaction*. 2018

### Dörner et al. 2019

In: DÖRNER, Ralf ; BROLL, Wolfgang ; JUNG, Bernhard ; GRIMM, Paul ; GÖBEL, Martin: *Einführung in Virtual und Augmented Reality*. Berlin, Heidelberg : Springer Berlin Heidelberg, 2019. – ISBN 978-3-662-58861-1, 1–42

### Huynh et al. 2019

HUYNH, B. ; ORLOSKY, J. ; HÖLLERER, T.: In-Situ Labeling for Augmented Reality Language Learning. In: *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2019, S. 1606–1611

### Jiao et al. 2019

JIAO, L. ; ZHANG, F. ; LIU, F. ; YANG, S. ; LI, L. ; FENG, Z. ; QU, R.: A Survey of Deep Learning-Based Object Detection. In: *IEEE Access* 7 (2019), S. 128837–128868

### Jmour et al. 2018

JMOUR, N. ; ZAYEN, S. ; ABDELKRIM, A.: Convolutional neural networks for image classification. In: *2018 International Conference on Advanced Systems and Electric Technologies (IC ASET)*, 2018, S. 397–402

### Leap 2018

LEAP, Magic: *App Security*. <https://developer.magicleap.com/en-us/learn/guides/application-security-overview>. Version: 2018. – [Online; Stand 18. September 2020]

### MagicLeap

MAGICLEAP: *JsonUtility.FromJson*. <https://www.magicleap.care/hc/en-us>. – [Online; Stand 1. Oktober 2020]

### MagicLeap 2018

MAGICLEAP: *magic leap 1*. <https://www.magicleap.com/en-us/magic-leap-1>. Version: 2018. – [Online; Stand 18. September 2020]

### MagicLeap 2019a

MAGICLEAP: *Lumin OS Overview*. <https://developer.magicleap.com/en-us/learn/guides/lumin-os-overview>. Version: 2019. – [Online; Stand 18. September 2020]

### MagicLeap 2019b

MAGICLEAP: *World Rekonstruktion*. <https://developer.magicleap.com/en-us/>

learn/guides/world-reconstruction-overview-landing. Version: 2019. – [Online; Stand 18. September 2020]

### **MagicLeap 2020a**

MAGICLEAP: *1.4 Spatial Meshing - Unity*. <https://developer.magicleap.com/en-us/learn/guides/meshing-in-unity>. Version: 2020. – [Online; Stand 18. September 2020]

### **MagicLeap 2020b**

MAGICLEAP: *Glossary and Usage*. <https://developer.magicleap.com/en-us/learn/guides/glossary>. Version: 2020. – [Online; Stand 18. September 2020]

### **MagicLeap 2020c**

MAGICLEAP: *Magic Leap Features*. <https://developer.magicleap.com/en-us/learn/guides/magic-leap-features>. Version: 2020. – [Online; Stand 18. September 2020]

### **Micosoft 2018**

MICOSOFT: *Mr und Azure 302b benutzerdefinierte Vision*. <https://docs.microsoft.com/de-de/windows/mixed-reality/mr-azure-302b>. Version: 2018. – [Online; Stand 17. September 2020]

### **Microsoft**

MICROSOFT: *Microsoft Azure Computer Vision*. <https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/>. – [Online; Stand 17. September 2020]

### **Microsoft 2018a**

MICROSOFT: *Mr und Azure 302 Maschinelles Sehen*. <https://docs.microsoft.com/de-de/windows/mixed-reality/mr-azure-302>. Version: 2018. – [Online; Stand 17. September 2020]

### **Microsoft 2018b**

MICROSOFT: *Spatial Mapping*. <https://docs.microsoft.com/de-de/windows/mixed-reality/spatial-mapping>. Version: 2018. – [Online; Stand 17. Septmber 2020]

### **Microsoft 2019a**

MICROSOFT: *Detect common objects in images*. <https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/concept-object-detection>. Version: 2019. – [Online; Stand 17. September 2020]

### **Microsoft 2019b**

MICROSOFT: *Erkennen von alltäglichen Objekten in Bildern*. <https://docs.microsoft.com/de-de/azure/cognitive-services/computer-vision/concept-object-detection>. Version: 2019. – [Online; Stand 24. September 2020]

### **Microsoft 2020**

MICROSOFT: *What is Computer Vision*. <https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/home>. Version: 2020. – [Online; Stand 17. September 2020]

### **O'Shea und Nash 2015**

O'SHEA, Keiron ; NASH, Ryan: An Introduction to Convolutional Neural Networks. In: *ArXiv e-prints* (2015), 11

**Unity 2020a**

UNITY: *Camera.cameraToWorldMatrix.* <https://docs.unity3d.com/ScriptReference/Camera-cameraToWorldMatrix.html>. Version: 2020. – [Online; Stand 18. September 2020]

**Unity 2020b**

UNITY: *JsonUtility.FromJson.* <https://docs.unity3d.com/ScriptReference/JsonUtility.FromJson.html>. Version: 2020. – [Online; Stand 24. September 2020]

**Unity 2020c**

UNITY: *Matrix4x4.MultiplyPoint.* <https://docs.unity3d.com/ScriptReference/Matrix4x4.MultiplyPoint.html>. Version: 2020. – [Online; Stand 18. September 2020]

## 8 Anhang