Roboter-Perzeption mittels lokaler planarer Volumenmodelle

Von der Universität Bayreuth
zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat)
genehmigte Abhandlung

von
Dorian Till Joscha Rohner
aus Schweinfurt

1. Gutachter: Prof. Dr. Dominik Henrich
2. Gutachter: Prof. Dr. Markus Vincze

Tag der Einreichung: 10.07.2025
Tag des Kolloquiums: 01.10.2025

Danksagung

Vielen Dank an Prof. Dr. Dominik Henrich für die Möglichkeit zu Promovieren, die Betreuung in der Zeit und das Setzen entsprechender Ziele. Besten Dank auch Prof. Dr. Markus Vincze für die zügige Erstellung des Zweitgutachtens, sowie den weiteren Mitgliedern des Prüfungsausschusses.
Darüber hinaus gilt mein Dank allen Freunden sowie (ehemaligen) Kollegen und Mitarbeitern am Lehrstuhl, die einen Teil der Strecke mit mir gegangen sind, im Besonderen: Anke, David, Johannes, Jonathan, Josua, Kim, Maximilian, Sascha, Simon und Tobias. Ohne die vielseitige Unterstützung über die Jahre hinweg wäre diese Arbeit in der aktuellen Form nicht möglich gewesen.
Weiterhin herzlichen Dank meinen Eltern, Helga und Eduard, die diesen Weg ermöglicht und begleitet haben. Abschließend sind mein Bruder Tristan zusammen mit Laura und Thea hervorzuheben, die mich alle auf unterschiedliche Art begeistert haben.

Abstract

Robot Perception using Local Planar Solid Models

In order to ensure successful interaction of robots in unknown surroundings, it is necessary to perceive these environments. Different sensors can be used to achieve an understanding of the surroundings, from a reconstruction of the scene, to recognizing objects and identifying interrelations. However, different challenges arise, like obstructions in the scene or changes in the environment.
This work aims to investigate the different sub-steps involved in understanding the environment. As a working hypothesis a representation of the sensor data directly mapping vertices, edges and faces is used (*Boundary Representation Models*, B-Reps). Especially advantages and disadvantages of using this representation are of interest. The used sensor in this work is limited to a robot-mounted depth-sensor (*Eye-in-Hand*) in order to reduce hardware cost and informative value.
A first step to understanding the surroundings is recognising all objects in the scene. The surface-based, segmentation-independent approach developed here uses only geometric features of B-Reps. Based on an object database, several hypotheses are generated and the best-fitting subset is selected. During evaluation, obstructions and poor views on the scenes are identified as limiting factors of the recognition rate. To tackle this, an *Active Vision* approach is developed, exploring the scene further and simultaneously validating existing hypotheses. Different kinds of edges in the reconstruction are used to determine new views scene-dependent to improve the object recognition. As the scene can be modified, for example by a human co-worker, detecting and handling these changes is necessary. For the geometric elements of a B-Rep all possible changes are defined and recognized within the global world representation. The changes are processed using the existing object hypotheses. As a last step towards understanding the environment the meaning of *Semantics* in robotics is investigated and how it can be systematically defined and modelled.
In addition to individual evaluations, different tasks and applications are discussed: On the one hand Pick-and-Place tasks, which can maintain a valid world representation after every step due to the used B-Reps. On the other hand a way to automatically generate precedence graphs for the human-robot-collaboration is discussed. Finally, an approach to automatically generate new models for the object database is presented.

Zusammenfassung

Roboter-Perzeption mittels lokaler planarer Volumenmodelle

Für die Verwendung von Robotern in unbekannten Umgebungen ist eine Wahrnehmung der Umwelt erforderlich, um eine erfolgreiche Interaktion mit dieser sicherzustellen. Unter Nutzung unterschiedlicher Sensorik kann ein Verständnis der Welt erlangt werden, beginnend mit einer allgemeinen Rekonstruktion über erkannte Objekte bis hin zu Zusammenhängen in der Welt. Dabei treten unterschiedliche Herausforderungen auf, beispielsweise durch Verdeckungen in der Szene oder Veränderungen an der Umwelt.
Diese Arbeit setzt sich zum Ziel, die unterschiedlichen Schritte bis zum Verständnis der Umgebung zu untersuchen. Als Arbeitshypothese wird dabei eine Repräsentationsform der Sensordaten verwendet, die direkt Knoten, Kanten und Flächen verwaltet (*Boundary Representation Models*, B-Reps). Dabei ist von Interesse, welche Vor- und Nachteile sich durch die Nutzung dieser Umweltmodellierung ergeben. Die verwendete Sensorik wird in dieser Arbeit auf eine robotergehaltene Tiefenkamera (*Eye-in-Hand*) eingeschränkt, um Hardwarekosten und die Mächtigkeit der verfügbaren Eingabedaten zu reduzieren.
Als erster Schritt zum Verständnis der Umwelt müssen die Objekte in der Szene wiedererkannt werden. Der entwickelte oberflächenbasierte, segmentierungsunabhängige Ansatz verwendet dazu ausschließlich die geometrischen Elemente der B-Reps. Ausgehend von einer Objektdatenbank werden Hypothesen erzeugt und die am besten geeigneten ausgewählt. In der Evaluation dieses Ansatzes zeigt sich, dass Verdeckungen von Objekten sowie ungünstige Positionen die Wiedererkennungsrate negativ beeinflussen. Dafür wird ein *Active Vision* Ansatz vorgestellt, der sowohl die Szene weiterhin exploriert als auch bestehende Hypothesen validiert. Durch unterschiedliche Typen von Kanten in der Umweltrepräsentation können diese Sichten szenenspezifisch zur Verbesserung der Wiedererkennung bestimmt werden. Da eine Veränderung der Szene möglich ist, beispielsweise durch menschliche Koarbeiter, ist eine Erkennung und Verarbeitung dieser Änderungen erforderlich. Für die geometrischen Elemente von B-Reps werden alle möglichen Arten von Änderungen definiert und in der globalen Umweltrepräsentation erkannt. Unter Einbeziehung potentieller Hypothesen werden die Änderungen schließlich verarbeitet. Als letzter Schritt zum Verständnis der Umwelt wird untersucht, was unter *Semantik* im Rahmen der Robotik zu verstehen ist und wie diese systematisch definiert und beschrieben werden kann.
Neben der individuellen Evaluation jedes Ansatzes werden schließlich mehrere Aufgaben und Anwendungen vorgestellt: Zum einen Pick-and-Place Aufgaben, bei denen die Umweltrepräsentation aufgrund der verwendeten B-Reps nach jeder Aktion automatisch konsistent gehalten wird. Zum anderen die Erzeugung von Präzedenzgraphen für die Mensch-Roboter-Kollaboration. Abschließend wird ein Ansatz zur automatischen Erzeugung neuer Einträge für die Objektdatenbank vorgestellt.

Inhaltsverzeichnis

Abkürzungsverzeichnis

B-Rep	Boundary-Representation (Model)
CAD	Computer Aided Design
ICP	Iterative Closest Points
LBR	Leichtbauroboter
LIDAR	Light Detection and Ranging
SIFT	Scale Invariant Feature Transform
SLAM	Simultaneous Localization And Mapping
TCP	Tool Center Point
NRR	Nicht reduzierbare Relationen
RFR	Referenzfreie Relationen
RBR	Referenzbehaftete Relationen
MR	Mengenrelationen

Kapitel 1

Einleitung

Inhalt

Das Einsatzgebiet der Robotik hat sich in den letzten Jahren neben der Industrie auch für den Haushalt und klein- und mittelständische Unternehmen entwickelt. Für den Einsatz in diesen neuen Domänen werden spezifische Anforderungen an die Umweltwahrnehmung des Roboters gestellt (Abschnitt 1.1). Ein Überblick, welche Information auf welche Art erfasst werden kann (Abschnitt 1.2), führt zu der Vision, dass Roboter ihre Umwelt vollständig erfassen und verstehen (Abschnitt 1.3). An die Ansätze für das Umweltverständnis werden Anforderungen und Evaluationskritierien diskutiert, aus denen sich unterschiedliche wissenschaftliche Fragestellungen für diese Arbeit ableiten lassen (Abschnitt 1.4).

1.1 Motivation: Wahrnehmen der Umwelt

Der Einsatzzweck sowie das Bild von Robotern hat sich in den letzten Jahren gewandelt: Ursprünglich wurden diese als Industrieroboter wahrgenommen und genutzt, die eine fest vorgegebene Bahn abfahren und denen jeder Arbeitsschritt präzise vorgegeben ist. Dadurch sind diese Systeme vergleichsweise fehleranfällig, da auf Unregelmäßigkeiten im Betrieb nicht reagiert werden kann. Ebenso ist die Zusammenarbeit mit dem Menschen aufgrund von Sicherheitsvorkehrungen meist nicht möglich. Dies wandelt sich durch die Verfügbarkeit von kollaborativen und kostengünstigen Robotern, die immer mehr in Klein- und mittleren Unternehmen eingesetzt werden, aber auch perspektivisch Aufgaben im privaten Haushalt übernehmen sollen.
In diesen neuen Domänen ergeben sich somit unterschiedliche Herausforderungen: Beispielsweise muss mit dem Menschen interagiert oder der Roboter muss vom Menschen instruiert beziehungsweise programmiert werden. Weiterhin kann man nicht von einer statischen Umwelt ausgehen. So kann während der Ausführung einer einzelnen Aufgabe durch den Roboter sich die Umwelt dynamisch ändern, beispielsweise durch einen Menschen der die Szene manipuliert. Weiterhin kann nicht von identischen Ausgangssituationen für die gleiche Aufgabe ausgegangen werden, was an den zugrundeliegenden Anwendungsorten wie zum Beispiel dem Haushalt liegt, da diese weniger strukturiert sind als Industrieumgebungen. Abschließend soll der Roboter eine Vielzahl unterschiedlicher Aufgaben erfüllen können, die unbekannte Werkstücke, dynamische Aufgabenaufteilung sowie variierende Ausgangs- und Endsituationen umfassen. Damit nicht jeder Zustand explizit modelliert werden muss, ist es notwendig, dass der Roboter seine Umwelt selbstständig erfassen kann. Somit kann dieser die Ausführung der Aufgabe den entsprechenden Begebenheiten anpassen.
Insgesamt existiert somit unterschiedliches Wissen über die Umwelt, welches wahrgenommen und erlangt werden muss. Zunächst ist der Roboter selbst Teil der Umwelt, der mit diversen Sensoren seinen eigenen Zustand beschreiben kann. Vorerst ist eine präzise Positions- und Orientierungsangabe (zusammenfasst *Pose*) möglich, die über die Gelenkwinkel und das dazugehörige Robotermodell gegeben ist. Über zusätzliche Kraftsensorik kann der Roboter Kollisionen erkennen oder Aussagen über gegriffene Werkstücke (zum Beispiel das Gewicht) treffen. Diese Art der Sensorik kann auch verwendet werden, um Wissen über eine vorliegende Szene bestehend aus Werkstücken zu erzeugen. So kann mittels kraftgeregelter Bewegungen die Arbeitsoberfläche abgetastet werden. Mit der Verwendung weiterer Sensorik erfasst der Roboter die Umwelt. Mithilfe von Kameras kann ein Bild der Szene aufgezeichnet werden, woraus Informationen sowohl über den Roboter als auch über die Anordnung von Werkstücken gewonnen werden können. Andere mögliche Sensorik umfasst optische Systeme, wie beispielsweise Lidar- beziehungsweise Radar-Sensoren. Als drittes Wissen über die Umwelt ist die Anwesenheit und Pose des Menschen von Interesse. Auch dafür eignen sich optische Sensoren. Spezialisierte Systeme verwenden Marker, die am Menschen platziert und erkannt werden. Darüber hinaus existieren weitere Sensoren, die in spezialisierten Anwendungen verwendet werden. Dazu gehören beispielsweise Thermal-Kameras, akustische oder elektromagnetische Sensorik. Auch in der mobilen Robotik, wie beispielsweise dem autonomen Fahren, findet sich diese Unterteilung. Sowohl der Zustand der mobilen Einheit (zum

Beispiel eines Autos), der Umwelt als auch weiterer Aktoren (wie Menschen) ist von Interesse und wird mit unterschiedlicher Sensorik erfasst.
Somit existieren für die Wahrnehmung der Umwelt unterschiedliche Ziele sowie eine Vielzahl an Möglichkeiten, diese wahrzunehmen. Verschiedene Sensoren können für mehrere Aufgaben genutzt werden, wobei der Nutzen variiert. So kann ein Roboter mittels kraftgesteuerter Abtastung eine Szene wahrnehmen, was im Vergleich zu einem Kamerabild aber mehr Zeit beansprucht und invasiv gegenüber der Szene ist. Im Weiteren werden zunächst die unterschiedlichen Arten von Sensoren diskutiert als auch die mögliche Verwendung.

1.2 Arten der Wahrnehmung

Basierend auf den unterschiedlichen Zielen werden im weiteren Verlauf verschiedene Sensoren erläutert. Dabei wird zum einen auf den Nutzen der Information eingegangen, zum anderen auf mögliche Repräsentationsarten. Dabei ist der Überblick nicht abschließend, sondern soll eine Zusammenfassung über bestehende und verbreitete Sensorik geben, auf welche im weiteren Verlauf aufgebaut werden kann. Die verwendete Sensorik in dieser Arbeit wird im Detail in Kapitel 3 für die Evaluation dargestellt.

Interne Robotersensorik

Neben diversen internen Zuständen, die für eine optimale Nutzung notwendig sind, verfügt der Roboter über unterschiedliche Sensorik, die Information über seine Interaktion mit der Umwelt gibt.

Position und Orientierung Für diese Interaktion ist allem voran die genaue Pose (entspricht der Position zusammen mit der Orientierung) des Roboters im Raum notwendig. Dazu wird ein *Weltkoordinatensystem* definiert, wozu die Pose des Roboters relativ angegeben werden kann. Die genaue Pose sowohl der räumlichen Ausdehnung des Roboters und der Gelenke als auch des *Tool-Center-Points* (TCP) wird mit Hilfe der Gelenkwinkel bestimmt. Das zugrundeliegende Modell des Roboters (was im Allgemeinen bekannt und gegeben ist) wird durch die Gelenkwinkel (welche durch geeignete Sensorik abgefragt werden) im Raum transformiert. Diese Information ist beispielsweise notwendig für Kollisionsberechnung sowohl mit der Umwelt als auch Selbstkollisionen des Roboters. Weiterhin ist eine genaue Pose hilfreich für eine erfolgreiche Manipulation. Abschließend kann die Pose des Roboters Einfluss auf den Nutzen von weiterer Sensorik haben, beispielsweise aufgrund von Verdeckungen und der Umrechnung von Sensordaten in das Weltkoordinatensystem. Siehe dazu auch die Ausführungen im Unterkapitel zu Kamerasystemen.

Manipulator Neben dem eigentlichen Roboter kann ein zusätzlicher Manipulator zur Verfügung stehen, klassischerweise eine Art Greifer, um mit Objekten zu interagieren. Je nach Art des Manipulators können unterschiedliche Rückschlüsse über die manipulierten Objekte getroffen werden. Dabei ist es aber auch möglich, dass keine Information erlangt wird, wenn

der Manipulator kein Feedback gibt. Ansonsten erfolgt im einfachsten Fall eine binäre Aussage, ob aktuell ein Objekt gegriffen wird oder nicht (beispielsweise Sauggreifer). Darüber hinaus hat die Wahl des Manipulators Einfluss auf die mögliche Information: So kann eine Schätzung der Größe mit Hilfe eines Backen- oder Mehr-Finger-Greifers erfolgen. Außerdem kann beispielsweise die vom Manipulator ausgewirkte Kraft gemessen oder das Gewicht des Objekts bestimmt werden.

Kraft-Momenten-Sensorik

Durch die steigende Popularität von Leichtbaurobotern in den letzten Jahren wurde die Verbreitung von Kraft-Momenten-Sensorik erhöht. Zunächst als Kraft-Mess-Dose verfügbar, ist diese Art von Sensorik stellenweise bereits in den Robotern an den Gelenken verbaut. Dadurch ist es zum einen möglich, die Kraft zu bestimmen, die der Roboter auf die Umwelt auswirkt. Dies ist notwendig für kraft-geführte sowie kraft-gesteuerte Bewegungen als auch zum Erkennen von Kollisionen. Zum anderen kann ein Roboter mit Hilfe der Gravitationskompensation mit der Hand geführt werden, ohne dass das Gewicht des Roboters gestützt werden muss.

Optische Sensorik

Um gezielter Informationen über die Umwelt zu erfassen, wird auf optische Sensorik zurückgegriffen. Kamerasysteme gehören im klassischen Sinn auch zu dieser Gruppe, auf die aber gesondert eingegangen wird. Optische Sensoren können je nach Dimension ihrer Daten unterteilt werden: von einer binären Aussage (0D) über einzelne Linien (1D) und Ebenen (2D) bis zu einer kompletten Raumrepräsentation (3D). Anwendung finden diese optischen Sensoren beispielsweise als Lichtschranken, um die An-/Abwesenheit von Objekten zu erkennen, als Abstandssensoren, um eine grobe Position zu schätzen, oder als Winkelmesser (beispielsweise für Gelenkwinkel). Technische Grundlagen können dabei unter anderem Lidar- und Radar-Systeme sein, die in den letzten Jahren vermehrt in der mobilen Robotik und dem autonomen Fahren Verwendung finden.

Kamerasysteme

Aufgrund der Beliebtheit und Mächtigkeit von Kamerasystemen wird auf diese gesondert eingegangen. Der Vorteil dieser ist, dass durch einen Sensor viel Information auf einmal aufgenommen werden kann, da ganze Raumbereiche direkt wahrgenommen werden. Nachteil dieser großen Menge an Informationen ist die vergleichsweise geringe Wiederholrate sowie die Anfälligkeit gegenüber unterschiedlichen Umwelteinflussen, allen voran Beleuchtung.

2D-Kameras Am weitesten verbreitet (vor allem durch den privaten Bereich) sind 2D-Kameras, die entweder ein Graustufenbild oder ein Farbbild aufnehmen können. Dieses Bild kann auf unterschiedliche Arten weiterverarbeitet werden. Falls die Pose der Kamera relativ zum Roboter bekannt ist, kann neben einer Aussage über die Umwelt diese auch räumlich zugeordnet werden. Die Information der Kamera kann aufgrund der bekannten Transformation

zwischen Roboter und Sensor in das Weltkoordinatensystem umgerechnet werden. Je nach Sichtbereich sind weiterhin Informationen über den Roboter ableitbar. Da Kameras ein sehr allgemeiner Sensor sind, liegt die Herausforderung bei der Verwendung darin, die relevante Information für eine spezifische Situation zu erhalten, was im Allgemeinen durch vergleichsweise aufwändige Analysen geschieht. Durch den Einsatz mehrerer 2D-Kameras kann zu einem gegebenen Zeitpunkt räumliche Information über die Umwelt an verschiedenen Orten gesammelt werden.

3D-Kameras Als Erweiterung von 2D-Kameras existieren 3D-Kameras, die zu jedem Pixel des Bildes noch den Abstand zur Kamera mitbestimmen (Tiefenwert). Der Tiefenwert kann dabei auf unterschiedliche Arten berechnet werden, beispielsweise über Korrespondenzen zwischen mehreren 2D-Bildern (Stereokamera) oder über Laufzeitmessungen. Die gewonnenen Daten liegen somit zunächst als Punktwolke vor. Ein möglicher Weiterverarbeitungsschritt ist zunächst eine Tesselierung, primär in Dreiecke (Triangulation). Von diesen Repräsentationen kann weiter abstrahiert werden, beispielsweise zu CAD-Daten.

Sonstiges

Neben diesen Sensoren existieren diverse weitere, die für unterschiedliche Anwendungsgebiete der Robotik von Interesse sind. Dazu gehören beispielsweise Temperatursensoren, Magnet- und Elektrofeldmessgeräte oder Mikrofone. Dabei handelt es sich um spezialisierte Messgeräte, die in bestimmten Anwendungsfällen notwendig sind. Am geläufigsten sind dabei Mikrofone, die durch Assistenzgeräte im Haushalt in den letzten Jahre steigende Beliebtheit erfahren haben und für die Roboter-Instruierung sowie -Programmierung genutzt werden können.

Positionierung der Sensorik

Neben den unterschiedlichen Arten von Sensoren ist die Positionierung entscheidend für den Nutzen. Im Allgemeinen kann ein Sensor entweder am Roboter oder in der Umwelt montiert werden. Je nach Art des Sensors ist dabei eine *extrinsische Kalibrierung* notwendig. Das bedeutet, dass der Ort des Sensors zu jedem Zeitpunkt relativ zum Weltkoordinatensystem bekannt ist. Falls der Sensor am Roboter befestigt ist, wird der Sensor relativ zum Roboter kalibriert und über die Pose des Roboters zur Welt weiter gerechnet. Das Anbringen der Sensorik in der Umwelt hat den Vorteil einer großen Flexibilität, allerdings verbunden mit dem Aufwand, bei einem Ortswechsel des Aufbaus erneut kalibrieren zu müssen. Zudem kann es zum Problem von Verdeckungen führen, wenn sich Agenten in der Welt bewegen und so den Sensor blockieren. Ein Vorteil der robotermontierten Sensorik ist dabei die Mobilität durch den Roboter und ein leichterer Einrichtungsprozess. Allerdings kann aufgrund von Größe, Gewicht oder sonstigen bauartbedingten Einschränkungen nicht jede Sensorik am Roboter angebracht werden.

Anzahl der einzelnen Sensorik

Abschließend kann man unterschiedliche Sensorik miteinander kombinieren und von einem Sensor mehrere Instanzen gleichzeitig verwenden. Die Kombination mehrerer, unterschiedlicher Sensoren stellt dabei meist kein besonderes Problem dar, es sei denn die Messprinzipien interferieren. Beispielweise können optische Sensoren (und damit auch Tiefensensoren) auf ähnlichen Wellenlängen arbeiten, wodurch die Messergebnisse verfälscht werden. Je nach Sensor und Messprinzip kann dieses Problem jedoch dann auftreten, wenn von einem Sensortyp mehrere verwendet werden. Allerdings wird die *Fusion* von Sensordaten stark vereinfacht, wenn ein Typ mehrmals benutzt wird. Durch die Verwendung mehrerer Sensoren lässt sich im Allgemeinen ein größerer Raum abdecken, was aber auch einen größeren Verwaltungsaufwand bezüglich Kalibrierung und Wartung bedeutet. Bei der Verwendung von Sensoren, die unterschiedliche Daten erzeugen, spricht man von *Multimodalität*. Beispielsweise kann im Rahmen der Robotik ein System gleichzeitig eine Kamera (Gestenerkennung), Mikrofon (Spracheingabe) und Kraftsensorik (Gravitationskompensation, Kraftsteuerung) für die Steuerung nutzen.

Fehlsignale

Abschließend sei auf Fehlsignale von unterschiedlichen Sensoren hingewiesen. Bedingt durch Messprinzipien und die Bauart sind einzelne Sensoren anfällig für unterschiedliche Fehlsignale und *Rauschen*. Diese können an unterschiedlichen Stellen im Aufnahmeprozess entstehen. Bei Kamerasystemen kann bei der Aufnahme und dem Auslesen der Information aus dem Sensor beispielsweise der *Bloom-* oder *Smear-* Effekt auftreten. Ein weiteres Beispiel ist die Quantisierung im Rahmen der Analog-Digital-Wandlung. Abschließend kann beispielsweise die Übertragung der digitalen Information gestört sein oder es entstehen Kompressionsartefakte durch eine En- und anschließende Dekodierung.

1.3 Vision: Verständnis der Umwelt

Insgesamt ist das Ziel der Perzeption ein vollständiges Verständnis der beliebig komplexen Umwelt. Die Umwelt setzt sich dabei aus der physikalischen Umgebung, den zu untersuchenden Objekten und sonstigen möglichen Agenten zusammen. Dabei ist es das Ziel, das Wissen über alle Komponenten der Umwelt zu maximieren. Welches Wissen notwendig ist, richtet sich dabei nach der Anwendung und den betrachteten Komponenten. Um Wissen über die Umwelt zu erzeugen, ist der Einsatz von Sensorik notwendig. Da das Einrichten und der Betrieb von Sensoren allerdings aufwendig und ressourcenintensiv ist, gilt es gleichzeitig, die notwendige Anzahl zu reduzieren. Darüber hinaus soll die Verarbeitung der Sensordaten echtzeitfähig erfolgen und eine minimale Zeit in Anspruch nehmen. Die Verarbeitungsdauer skaliert mit der Menge und Art der verwendeten Sensorik sowie deren Repräsentationsformen und der tatsächlichen Anwendung und deren Anforderungen.

1.4 Wissenschaftliche Fragestellungen und Zielstellung

Basierend auf der Vision wird im Weiteren ein Aspekt des Verständnisses der Umwelt als Ziel für diese Arbeit ausgewählt. Dazu wird zunächst das Interessensgebiet des Szenenverständnisses als in sich geschlossene Problemstellung näher betrachtet und eine Liste an Anforderungen vorgestellt. Szenenverständnis bedeutet in diesem Fall, dass die Anwesenheit und Zusammenhänge zwischen Werkstücken untersucht werden, mit denen der Roboter interagieren soll. Basierend darauf werden wissenschaftliche Fragestellungen für die Arbeit erstellt.

1.4.1 Anforderungen und Herausforderungen

Für das Ziel der Wahrnehmung der Umwelt des Roboters wurde in [Bennamoun02] eine Menge an Anforderungen und Herausforderungen vorgestellt. Sie werden in Grundzügen dargestellt und um einige Problemstellungen erweitert und lassen sich in drei Gruppen aufteilen: Herausforderungen im Zusammenhang mit dem konkreten Wiedererkennungsansatz, mit den zu untersuchenden Objekten und der individuellen Szene. Zunächst zu den Schwierigkeiten mit den verwendeten Objekten:

Objektkomplexität Eine Herausforderung für das Erkennen von Objekten ist deren Komplexität. In [Bennamoun02] wird dabei primär auf die geometrische Form der Objekte eingegangen. Dabei sind sowohl geometrisch primitive Objekte problematisch, da diese wenig Oberflächeninformation tragen und häufig Ähnlichkeiten zu verwandten Objekten haben. Dies ist unter anderem bekannt als das Problem der *Bricks World*. Auf der anderen Seite müssen geometrisch komplexe Objekte repräsentiert werden können, was allem voran eine Herausforderung für abstraktere Repräsentationen darstellt. Die Schwierigkeit der Objektkomplexität kann weiterhin auf die Texturierung analog angewandt werden. Objekte mit einer komplexen Textur können dabei leichter erkannt werden als Objekte vollständig ohne Textur.

Veränderliche Objekte Manche Objekte sind in einem konkreten Zustand grundsätzlich starr, können durch externe Kräfte allerdings in einen anderen Zustand versetzt werden (beispielsweise Scheren, Bücher, Verpackungsmaterial). Zudem können Objekte deformierbar sein. Da diese Gruppe von Gegenständen in bis zu unendlich vielen Konfigurationen vorliegen kann, ist die Repräsentation sowohl dieser als auch einer Referenzdarstellung eine Herausforderung.

Neben diesen Herausforderungen kann der Wiedererkennungsansatz unterschiedlich mächtig sein:

Größe der Objektdatenbank Eine häufige Annahme für die Verwendung von Objektwiedererkennungsverfahren ist die Verfügbarkeit einer Objektdatenbank. Das bedeutet, dass alle möglichen Objekte in einer Szene mit ihren vollständigen Merkmalen vorab bekannt sind. Je nach Anwendung muss die dazugehörige Objektdatenbank eine Vielzahl von Objekten verwalten, woraus sich mehrere Probleme ergeben: Zum einen kann es ab einer gewissen Größe

zu aufwendig sein, gegen jeden Eintrag in der Objektdatenbank zu testen. Zum anderen ist es mit steigender Anzahl an Objekten wahrscheinlich, dass sich mehrere Objekte in bestimmten Eigenschaften ähneln (beispielsweise eine ähnliche Form oder Textur).

Aufbau der Objektdatenbank Neben der Größe der Objektdatenbank ist eine Schwierigkeit, wie bisher unbekannte Objekte neu hinzugefügt werden können (beispielsweise wenn ein neues Werkstück benötigt wird). Eine Problem ist es dabei, das neue Modell mit allen notwendigen Merkmalen vollständig zu erfassen. Weiterhin muss das neue Modell dem Wiedererkennungsansatz zur Verfügung gestellt werden. Je nach Methodik reicht es, das neue Objekt in der Datenbank zu hinterlegen (beispielsweise wenn alle Objekte aus der Datenbank in der Szene gesucht werden) oder das Wiedererkennungsverfahren vollständig neu zu trainieren. Diese Herausforderung verschärft sich bei veränderlichen Objekten.

Objektkategorien Die Wiedererkennung konkreter Objektinstanzen kann dahingehend erweitert werden, dass in einem ersten Schritt zunächst übergeordnete Klassen wiedererkannt werden, um bei Bedarf dann eine konkrete Instanz zu identifizieren. Beispielsweise reicht es oft, dass eine Tasse vorhanden ist - welche genau spielt je nach Anwendung eine untergeordnete Rolle. In manchen Fällen kann aber eine genaue Identifikation notwendig sein.

Darüber hinaus können im Aufbau einer zu untersuchenden Szene mehrere Herausforderungen existieren.

Verdeckungen In einer Szene können die Objekte so positioniert sein, dass diese sich gegenseitig gegenüber dem Sensor verdecken. Dadurch ist ein Objekt nicht vollständig erfassbar. Für Ansätze zur Objektwiedererkennung ist es somit notwendig, auch mit einem geringen Erfassungsgrad der Objekte diese trotzdem noch korrekt in der Szene zu erkennen.

Szenenkomplexität Als Erweiterung der Herausforderungen in [Bennamoun02] wird hier die Szenenkomplexität analog zur Objektkomplexität eingeführt. Die Problematik liegt darin, dass Szenen zu komplex sein können, als dass diese aus einem einzelnen Blickwinkel der Kamera zu erfassen sind. Einerseits verlangt das die Möglichkeit, mehrere lokale Sichten aus unterschiedlichen Posen konsistent zu einem globalen Weltmodell zu verschmelzen. Andererseits müssen sowohl die Umweltrepräsentation als auch der Ansatz zum Erkennen von Objekten eine größere Menge Daten (als nur eine einzelne lokale Sicht) verarbeiten können.
Neben diesen Herausforderungen wurden in [Bennamoun02] mehrere Bewertungskriterien an die **Umweltrepräsentation** vorgestellt, die aus den beschriebenen Anforderungen folgen: Allem voran soll die Repräsentation **effizient** sein - sowohl im Rechenaufwand, als auch im Speicherbedarf. Darüber hinaus **robust**, was in diesem Zusammenhang die Invarianz bezüglich affinen Transformationen und Rauschen beziehungsweise Fehlsignalen bedeutet. In diesem Zusammenhang soll die Repräsentation möglichst **genau** sein, sodass die tatsächlichen Objekte und die in der rekonstruierten Welt möglichst wenig voneinander abweichen, und die ursprüngliche Szene aus der rekonstruierten wiederhergestellt werden kann. Abschließend soll die **Domäne** an möglichen Objekten groß genug sein und dementsprechend *ausreichend*

für die Repräsentation der notwendigen Informationen - je nach Wiedererkennungsansatz. Schließlich soll die Repräsentation **lokale** Sichten unterstützen, sodass auch teilweise unvollständige Objekte wiedererkannt werden.

1.4.2 Zielstellung

Basierend auf den Arten der Wahrnehmung, der Vision sowie den diskutierten An- und Herausforderungen wird im Weiteren die Zielsetzung dieser Arbeit dargestellt. Der Fokus liegt dabei auf dem Verständnis der Umwelt. Die weiteren Aspekte der Perzeption (Roboterzustand und Wahrnehmung von weiteren Agenten) stellen gesonderte Herausforderungen dar und sind Bestandteil eigener Forschung. Das Ziel dieser Arbeit ist es, dem Roboter mit Hilfe geeigneter Sensorik ein Weltmodell zur Verfügung zu stellen, das die aktuelle Szene beschreibt. Dazu ist es zunächst notwendig, eine geeignete Repräsentation basierend auf den eben dargestellten Kriterien auszuwählen. Ausgehend von dieser Repräsentation werden anschließend mehrere Schritte der Perzeption untersucht. Ein erster Schritt ist dabei die Wiedererkennung von bekannten Objekten in einer Szene. Dabei sollen die Objekte bezüglich ihrer Komplexität möglichst wenig Information tragen: sowohl geringe Oberflächen- als auch Texturinformation. Zudem soll das Problem der lokalen Sicht gelöst werden. Dazu ist es notwendig, basierend auf einer gegebenen lokalen Sicht und dem dazugehörigen Weltmodell, neue Sichten auf die Szene zu erzeugen und korrekt einzuarbeiten. Diese Sichten müssen sowohl von der Umweltrepräsentation als auch dem Wiedererkennungsansatz integriert werden können. Dabei ist es möglich, dass Objekte in der Szene bewegt werden, allem voran im potentiellen Kontext der Mensch-Roboter-Kollaboration. Somit muss die Umweltrepräsentation zum einen die weiterhin gültige Information halten, diese aber entsprechend den Änderungen anpassen können. Abschließend muss eine Möglichkeit gegeben werden, um zusätzliches Wissen zu generieren, was über die Position und Orientierung einzelner Objekte hinausgeht (beispielsweise Zusammenhänge zwischen Objekten oder die konkrete Bedeutung eines Objektes in der Szene oder für eine Aufgabe). Dadurch kann mehr Verständnis über die Szene erzeugt werden, was im Rahmen von Manipulationen durch unterschiedliche Agenten notwendig ist.
Von den vorgestellten Herausforderungen werden in dieser Arbeit nicht alle betrachtet. Wie oben beschrieben soll eine niedrige Objektkomplexität Ausgangslage für die Objektwiedererkennung sein. Allerdings wird in dieser Arbeit davon ausgegangen, dass alle Objekte nur in einem definierten, nicht veränderlichen Zustand (betreffend Form und Farbe) vorliegen. Bei der Verwendung einer Objektdatenbank ist zu untersuchen, inwieweit die Größe einen Einfluss auf die Ergebnisse der Wiedererkennung hat und wie leicht diese bei Bedarf ergänzt werden könnte. Die Verwendung von Objektkategorien wird vom verwendeten Wiedererkennungsansatz abhängen. Da Verdeckungen in der Robotik eine bekannte Herausforderung sind, ist zu untersuchen, inwieweit mit diesen umgegangen werden kann. Ebenso ist es je nach gewählter Sensorik nicht möglich, mit nur einem Sensoreindruck die vollständige Szene zu erfassen. Von besonderer Bedeutung ist die Wahl der Umweltrepräsentation, da diese Auswirkungen auf die weiteren Herausforderungen hat. Das Ziel dieser Arbeit ist es, möglichst viele der gegebenen Anforderungen zu erfüllen. Die Diskussion und Auswahl der Umweltrepräsentation erfolgt gesondert in Abschnitt 2.2.

1.4.3 Wissenschaftliche Fragestellungen

Insgesamt ergeben sich so unterschiedliche wissenschaftliche Fragestellungen, die im Hinblick einer neuen Repräsentationsform beantwortet werden müssen. Dabei ist es zunächst notwendig, Objekte sicher wiedererkennen zu können, auch wenn diese wenig Information tragen:

F1 Inwieweit können geometrisch primitive, untexturierte Objekte wiedererkannt werden?

Da für den Wiedererkennungsschritt einzelne Sichten oft nicht ausreichend sind, stellt sich die Frage, wie sehr neue Sichten auf die Szene die Wiedererkennungsrate verbessern können. Dazu ist zum einen zu untersuchen, auf welche Art und Weise neue Sichten generiert werden können, die die aktuell gesammelten Informationen berücksichtigen. Zudem muss beachtet werden, dass die neu gesammelten Informationen im Weltmodell mit hinterlegt werden.

F2 Inwieweit unterstützen zusätzliche lokale Sichten die Objektwiedererkennung? Inwieweit können neue Sichten szenenspezifisch bestimmt werden?

Da zwischen zwei Aufnahmen die Szene sich aufgrund externer Faktoren ändern kann, ist es notwendig, diese Änderungen zu erkennen.

F3 Inwieweit können Änderungen an der Szene zwischen zwei Aufnahmen erkannt und verarbeitet werden?

Sobald die drei vorangegangenen Fragen beantwortet sind, ist das grundlegende Problem der Objektwiedererkennung in einigen Aspekten gut untersucht. Da allerdings nicht nur Objektinstanzen als Ergebnis der Perzeption relevant sind, sondern auch deren Bedeutung, muss weiteres Wissen bestimmt werden.

F4 Inwieweit kann semantische Information definiert und extrahiert werden?

Hierbei sind die ersten drei Fragestellungen direkt voneinander abhängig, da ohne eine Objektwiedererkennung keine neuen Sichten szenenspezifisch bestimmt werden können. Ohne neue Aufnahmen nach der Bewegung eines Roboters ist der Umgang mit dynamischen Szenen hinfällig. Daher existiert ein Zusammenhang zwischen der zweiten und dritten Fragestellung. Für eine Spezifizierung der wissenschaftlichen Fragestellungen auf die Aufgabenstellung dieser Arbeit siehe Kapitel 2.2.3.

1.5 Kapitelübersicht

Um diese Fragen zu beantworten, werden in Kapitel 2 notwendige Grundlagen dargelegt. Dazu wird zunächst der allgemeine Stand der Forschung bezüglich Roboterperzeption diskutiert. Zusammen mit dem Bewertungskriterium aus Kapitel 1 wird anschließend der Grundansatz, die gewählte Umweltrepräsentation und die Arbeitshypothese vorgestellt. Abschließend wird als Grundlage für die folgenden Kapitel vorgestellt, wie diese Repräsentationsform online erzeugt und verwaltet werden kann.

Die darauf folgenden Kapitel beschäftigen sich jeweils mit der Beantwortung der einzelnen wissenschaftlichen Fragestellungen. Der Aufbau ist im Allgemeinen ähnlich. Ausgehend vom spezifischem Stand der Forschung zu dem aktuellen Thema wird die Methodik erläutert und anschließend evaluiert. In Kapitel 3 wird der spezifische Objektwiedererkennungsansatz vorgestellt. Ausgehend von einer Objektdatenbank und der erzeugten Umweltrepräsentation wird eine Menge an möglichen Hypothesen bestimmt, woraus die bestmöglichen, überschneidungsfreien ausgewählt werden. Dieses Wissen wird in Kapitel 4 genutzt, um neue Sichten zu generieren, die sowohl den Arbeitsraum explorieren und so neue Objekte erkennen als auch bestehende Hypothesen validieren, um Verwechslungen und uneindeutige Erkennungen zu beheben. In Kapitel 5 werden anschließend die unterschiedlichen Arten von Änderungen zwischen zwei Sensoreindrücken klassifiziert, und es wird erläutert, wie diese jeweils erkannt werden können. Für jede Art der Änderung wird ein möglicher Umgang bezüglich der gewählten Umweltrepräsentation vorgeschlagen. Abschließend wird in Kapitel 6 die Extraktion von weiterführendem Wissen diskutiert. Dazu wird eine formale Beschreibung von semantischem Wissen erarbeitet, welche genutzt wird, um ausgewählte Informationen zu extrahieren. In Kapitel 7 erfolgt die Diskussion unterschiedlicher Aufgaben, die mit den hier vorgestellten Methoden gelöst werden können. Die Zusammenfassung und Diskussion der Ergebnisse dieser Arbeit erfolgt ebenso wie ein Ausblick zu weiterführenden Arbeiten in Kapitel 8.

Kapitel 2

Grundlagen

Inhalt

In diesem Kapitel werden zusammen mit den wissenschaftlichen Fragestellungen und dem Stand der Forschung die Grundlagen dargelegt, und die Arbeitshypothese wird mit den einzelnen Teilschritten erarbeitet. Dazu werden zunächst verwandte Gesamtsysteme untersucht (Abschnitt 2.1). Zusammen mit einer Festlegung auf eine *Eye-in-Hand*-Kamera und *Boundary Representation Models (B-Reps)* als Repräsentation der 3D-Daten wird die Arbeitshypothese definiert (Abschnitt 2.2). Als notwendige Grundlage wird abschließend dargestellt, wie aus der Ausgabe des Sensors B-Reps rekonstruiert werden können (Abschnitt 2.3).

2.1 Stand der Forschung

Der Stand der Forschung in diesem Kapitel beschränkt sich auf Gesamtsysteme und verwendete Repräsentationsformen der Umwelt. Jedes weitere Kapitel diskutiert den jeweiligen Forschungsstand detailliert für das zugehörige Thema. Ziel dieses Abschnittes ist es, einen Überblick bezüglich aufgeworfener und gelöster Fragestellungen, genutzter Sensoren, gewählter Repräsentionsformen der Umwelt sowie Möglichkeiten und Grenzen aufzuzeigen. Dazu wird zunächst auf regel- und modellbasierte Ansätze eingegangen (zum Teil auch als *klassische* Bildverarbeitung aufgefasst) und anschließend auf Ansätze mit künstlichen neuronalen Netzen. Aus diesem Überblick wird schließlich der zu verwendende Sensor sowie die Repräsentationsform für diese Arbeit abgeleitet.

2.1.1 Regel- und modellbasierte Ansätze

Die Analyse der Roboterumwelt ist in unterschiedlichen Teildisziplinen notwendig, in denen die Position von Objekten und Agenten in der Szene nicht fest definiert ist.
In [Rusu09a] wird ein globales Weltmodell bestehend aus Punktwolken aufgebaut. Dazu werden Punktwolken aus unterschiedlichen Posen in eine gemeinsame Repräsentation eingefügt. In dieser Repräsentation werden Objekte ausgehend von Oberflächenmerkmalen wiedererkannt. Die Semantik umfasst dabei unter anderem die Klassifikation von polygonalisierten Punktnetzen gemäß deren Winkeln zueinander, um Geländetypen zu erkennen.
Für autonome Roboter im Haushalt stellt [Klasing10] Ansätze unter anderem für die Erkennung von Möbeln im Haushalt dar. Komplette Objekte werden dabei durch einzelne Teile beschrieben, die aus einer Objektdatenbank heraus in mehrere Komponenten zerlegt werden können. Diese Teile stehen in festen, räumlichen Beziehungen, die für die Wiedererkennung genutzt werden.
Die Arbeit [Buchholz15] beschäftigt sich mit der Perzeption einer Szene mit dem Ziel ein wiedererkanntes Objekt in der Szene zu greifen. Als Eingabe werden 3D-Punktwolken verwendet, die mit Hilfe eines oberflächenbasierten Ansatzes wiedererkannt werden sollen. Die Objektmodelle liegen dabei als CAD-Daten vor, werden für die Objektwiedererkennung allerdings abgetastet.
In [Fischer15] wurden auf Tiefendaten, die aus mehreren Sensoren bestimmt wurden, beliebige haushaltsübliche Objekte wiedererkannt. Die Anwendung war dabei ein Haushalts- beziehungsweise Pflegeroboter. Dabei wird zwischen der Wiedererkennung von texturierten und texturlosen Objekten unterschieden, wobei der Nutzer beim Hinzufügen der Modelle zu einer Datenbank diese Entscheidung trifft. Als Repräsentation der Szene werden direkt die Farbbilder und Punktwolken verwendet.
Für einen mobilen Roboter entwickelte [Grotz21] ein System zur Wahrnehmung der Umwelt. Der verwendete Sensor erzeugt Tiefendaten, in die geometrische Primitive angepasst werden, um Gegenstände zu repräsentieren. Für diese Objekte werden Nachbarschaftsbeziehungen und Stabilitätseigenschaften bestimmt. Unterstützt wird das System durch eine Active Vision Komponente, die den Blick auf spezifische Szenen während der Bewegung der Kamera oder des Roboters stabil halten kann.

Auch für einen mobilen Roboter verwendet [Kollmitz21] eine Farbkamera, um mittels neuronaler Netze Personen wiederzuerkennen und den Abstand zu diesen zu schätzen. Um mögliche Kollsisionen des Roboters mit der Umwelt zu verarbeiten, wird Kraftsensorik verwendet, die den kompletten Roboter überwacht. Das Ziel ist die Unterstützung von Menschen mit Mobilitätshilfen.
Die Arbeit [Ding22] kombiniert unterschiedliche Sensorik; zum einen Tiefenmessungen über Laufzeit, zum anderen Näherungssensoren mit dem zugrundeliegenden Ziel der Kollisionsvermeidung ohne Erkennung von konkreten Objektinstanzen.

2.1.2 Neuronale Netze

Für einzelne Aufgaben in der Robotik und der Bildverarbeitung werden neuronale Netze schon seit längerem verwendet. In den letzten Jahren sind aber auch vollständige Systeme bedeutender geworden, in denen Szenenerfassung, Bewegungsplanung und Ausgabenausführung in einem neuronalen Netz geplant werden. Zu erwähnen sind dabei allem voran *Foundation Models* und *Large Language Models*. Unter *Foundation Models* versteht man im Allgemeinen künstliche neuronale Netze, die mit einer sehr großen Datenmenge sowie einer großen Anzahl an Parametern im Netz trainiert wurden. Eine Spezialisierung davon sind *Large Language Models*, die gezielt für den Einsatz mit natürlicher Sprache entwickelt wurden. In der Robotik gibt es mehrere solcher Systeme.
In den Arbeiten [Brohan23b] [Brohan23a] wurde mit *RT-1* und *RT-2* (*Robotic Transformer*) ein neuronales Netz der *Transformer*-Architektur [Vaswani17] trainiert, welches zu gegebenen Anweisungen und Farbbildern entsprechende Anweisungen im Konfigurationsraum erzeugt. Das Netz wird dabei zum einen auf frei verfügbaren Daten aus dem Internet trainiert. Zum anderen finden spezifische Aufgabenausführungen mit entsprechenden niedrig-leveligen Anweisungen an eine Robotersteuerung Verwendung. Auf diese Weise können natürlichsprachliche Anfragen beantwortet und durch den Roboter ausgeführt werden. Das resultierende Training zeigt in der Evaluation große Generalisierungsfähigkeiten über unterschiedliche Objekte sowie Szenen auf.
Im Modell *PaLM-SayCan* (*Pathways Language Model Say Can*) [Ahn22] (als Erweiterung von *PaLM* (*Pathways Language Model*) [Chowdhery23]) wird eine natürlichsprachliche Eingabe mittels des neuronalen Netzes zunächst interpretiert. Eine gegebene Menge an möglichen Aktionen wird dahingehend bewertet, wie gut diese die Anfrage erfüllen würde. Für den Roboter und die aktuelle Umwelt wird danach bewertet, ob die unterschiedlichen Aktionen aktuell möglich sind (*Affordanzen*). Durch die Gewichtung der Bewertungen kann eine Aktion ausgeführt werden, welche während der Ausführung immer wieder evaluiert wird, um sich an mögliche Änderungen anzupassen.
Ein weiteres *Large Language Model* zu Fragen wie Aufgaben- und Bewegungsplanung, Bildanalyse und -beschreibung sowie Manipulation von Objekten ist *PaLM-E* (*Pathways Language Model Embodied*) [Driess23]. Dieses Modell ist aus Kombination der zwei Modelle *PaLM* (*Pathways Language Model*) [Chowdhery23] und *ViT* (*Vision Transformer*) [Dehghani23] hervorgegangen. Als Eingabe wird ein natürlichsprachlicher Satz verwendet, der durch unterschiedliche Sensoreindrücke (wie Farbbilder) ergänzt wird. Ausgegeben wird eine Antwort auf die

Anfrage, was einerseits eine rein sprachliche Beantwortung sein kann, andererseits aber auch eine Anweisung an den Roboter, die beispielsweise mittels [Brohan23a] auf tatsächliche Aktionen abgebildet wird. Solange die Aktion nicht erfolgreich ausgeführt wurde, wird die Kontrollschleife wiederholt, indem die neu aufgezeichneten Sensorwerte erneut mittels PaLM-E ausgewertet und durchgeführt werden.
Ein abschließendes Modell ist *RobotCAT* [Bousmalis23], welches ähnlich zu den bisherigen Ansätzen versucht, sprachliche Eingabe mit Sensoreindrücken zusammen auf Roboteraktionen abzubilden. Dabei stehen vor allem Verbesserungen beim Training und den verwendeten Daten im Vordergrund, wodurch sich der Bedarf an tatsächlichen Demonstrationen reduzieren soll und die Generalisierung steigt.
Das Alleinstellungsmerkmal dieser Ansätze ist die Umsetzung aller notwendigen Schritte zur erfolgreichen Aufgabenbearbeitung in einem Modell. Während bei den regel- und modellbasierten Ansätzen jeder Schritt einzeln durchgeführt wird, ist hier die Analyse der Eingabedaten, das Berechnen der Bewegung sowie die Überwachung der Aufgabenausführung in eine Komponente integriert.

2.2 Grundansatz und Arbeitshypothese

Ausgehend von diesem Stand der Forschung wird im Weiteren zunächst die Sensorik ausgewählt, die in dieser Arbeit verwendet wird. Für ebendiese Sensorik wird ausgehend vom bisherigen Stand der Forschung, einem weiteren Überblick über mögliche Umweltmodellierungen sowie der konkreten Aufgabenstellung eine Umweltrepräsentation festgelegt.

2.2.1 Wahl der Sensorik zur Umwelterfassung

Wie in der Einleitung beschrieben, kann die Umwelt mit unterschiedlichen Sensoren auf diverse Arten erfasst werden. Je nach Art des Sensors kann eine andere Art von Wissen und Umweltrepräsentation erzeugt werden. Und je nach Art des Sensors ist die Erfassung invasiver im Umgang mit der Umwelt. Für die dargelegte Vision - den Roboter als kooperativen Arbeiter im Haushalt oder Klein- und mittleren Unternehmen - müssen unterschiedliche Kriterien an die Wahl der Sensorik gestellt werden.
Zum einen ist eine geringe Anzahl an Sensoren erstrebenswert, um eine niedrigschwellige Verfügbarkeit sowie Skalierbarkeit größerer Installationen sicherzustellen. Das reduziert allem voran den Einrichtungsbetrieb, da Sensorik oft auf unterschiedliche Arten kalibriert werden muss (beispielsweise eine extrinsische Kalibrierung für Kameras). Zum anderen ist die Geschwindigkeit entscheidend, mit der eine Szene erfasst werden kann, da für manche Aufgaben eine erfasste Umwelt Voraussetzung ist und eine niedrige Geschwindigkeit somit die Ausführung von Aufgaben verzögern würde. Manche Sensoren erfassen mit einer Messung einen Bereich der Szene (beispielsweise Kameras), während andere Sensoren nur einen kleinen Teil der Szene mit einer Messung erfassen können (zum Beispiel Kraft-Momenten-Sensorik zur Umweltrekonstruktion). Weiterhin ist für ein möglichst aussagekräftiges Umweltverständnis der Informationsgehalt von Interesse, der pro erfasstem Bereich erlangt wird. So erfasst eine Tiefenkamera bei gleichem Szenenausschnitt und gleicher Auflösung mehr Information als

eine 2D-Kamera, da zusätzlich die Tiefeninformation für jeden Pixel vorliegt.
Abschließend können weitere subjektive Einschränkungen existieren. Beispielsweise kann viel Kontakt des Roboters mit der Umwelt unangenehm für den menschlichen Koarbeiter sein. Weitere mögliche Einschränkungen können allgemeine Verfügbarkeit der Sensorik sein (beispielsweise aufgrund der Kosten) oder die Anwendbarkeit in der Robotik (beispielsweise durch Gewicht oder Größe).
Ausgehend von der vorgestellten Sensorik ist als Sensor eine Kamera naheliegend. Zum einen kann eine Kamera einen Bereich der Szene erfassen und mit wenigen Sichten eine Szene vollständig aufnehmen (in Abhängigkeit von Verdeckungen und allgemeiner Größer der Szene). Ebenso kann die Anzahl der benötigten Kameras reduziert werden, indem diese am Roboterarm direkt befestigt werden, meist in der Nähe des TCP (*Eye-in-Hand*). Da die Kamera somit im Raum bewegt werden kann, reicht eine einzelne aus, um große Teile der Szene zu rekonstruieren. Die Verwendung der Kamera als *Eye-in-Hand* verhindert zudem das Problem von Verdeckungen durch den Roboter oder einen Menschen, die bei einer *Overhead*-Kamera auftreten würden. Abschließend ist eine Vorkalibrierung des Sensors möglich, wenn dieser fest am Roboterarm verbaut ist (siehe auch *Plug and Produce* Paradigma [Cencen18]). Da der Informationsgehalt von Kameras vergleichsweise groß ist und die Montage als Eye-in-Hand sowohl die Verdeckungsproblematik als auch die Kalibrierungsprobelmatik löst, wird eine einzelne Kamera als Sensor verwendet.
Für Kameras muss abschließend noch zwischen 2D- und 3D-Kameras unterschieden werden. 2D-Kameras haben den Vorteil, dass das Messprinzip weniger anfällig ist als das von 3D-Kameras. Je nach Messprinzip zur Erzeugung der 3D-Daten können unterschiedliche Fehler wie stark reflektierende Objekte, ungünstige Farben oder schlechte Lichtverhältnisse auftreten, die die Qualität der Tiefenmessung senken oder unmöglich machen. Unter geeigneten Bedingungen hingegen erzeugen 3D-Kameras mehr Informationen über die Szene, wodurch die Pose von Objekten leichter berechnet werden und eine dreidimensionale Repräsentation erzeugt werden kann.
Weiterhin können 3D-Tiefenkameras mit Farbbildern fusioniert werden, womit zu jedem Tiefenwert auch eine Farbrepräsentation verfügbar ist. Das kann beispielsweise als Punktwolke repräsentiert werden oder auch als 2D-Farbbild, wobei in jedem Pixel noch die Tiefe hinterlegt (*RGB-Depth*) wird. Das Erzeugen dieser kombinierten Repräsentation ist wiederum anfälliger für Fehler, da der Farb- und Tiefensensor aufeinander kalibriert werden müssen. Im Rahmen dieser Arbeit wird allerdings auf die Farbdaten verzichtet, um einen leichten Einrichtbetrieb sicherzustellen und die Kalibriervorgänge zu minimieren und um zu untersuchen, wie viel Wissen über eine Szene nur basierend auf Tiefendaten erzeugt werden kann.

2.2.2 Wahl der Umweltrepräsentation

Ausgehend von der Wahl der Sensorik muss eine 3D-Umweltrepräsentation ausgewählt werden. Für diese existieren unterschiedliche Anforderungen, auf deren Basis eine mögliche Repräsentation vorgestellt und ausgewählt wird.
Für ein Szenenverständnis ist es notwendig, dass die Repräsentation online auf Basis der Sensoreindrücke bestimmt werden kann. Eine nachträgliche Verarbeitung offline reicht nicht

aus, da während des Betriebs aus der aufgenommenen Szene Informationen für den Roboter zur Verfügung gestellt werden müssen. Darüber hinaus sollte die Integration neuer Sichten unterstützt werden, ohne dass der Speicher oder die Berechnungszeiten so sehr anwachsen, dass Berechnungen nicht mehr online durchgeführt werden können. Die Registrierung der unterschiedlichen Sensoreindrücke aus mehreren Perspektiven kann durch die bekannte Pose (aufgrund der extrinsischen Kalibrierung) der Kamera in der Szene unterstützt werden. Abschließend soll der Speicher- und Rechenaufwand möglichst gering sein.
Das Aufgabengebiet der 3D-Rekonstruktion aus mehreren Sichten lässt sich dabei in mehrere Gruppen aufteilen [Sand19] (siehe auch ebenda für einen ausführlicheren Stand der Forschung), teilweise historisch bedingt durch unterschiedliche Anwendungen beispielsweise im autonomen Fahren, in der Konstruktion oder in der Computergrafik. Eine Methodik, die durch Verbreitung von Tiefenkameras sowie autonomen Robotern größere Verbreitung erlangte, ist *Simultaneous Localizazion and Mapping* (*SLAM*) [Younes17,Taketomi17]. Das Ziel dabei ist, gleichzeitig eine Karte der Umwelt aus den Sensoreindrücken aufzubauen als auch die Pose des Sensors in ebendieser Karte zu bestimmen. Dabei ist zu beachten, dass die Problemstellung der Lokalisierung im Rahmen dieser Arbeit aufgrund des kalibrierten Sensors entfallen würde. SLAM Ansätze lassen sich weiterhin in graph- sowie modellbasierte Ansätze unterteilen. Graphbasierte Ansätze verwalten dabei die unterschiedlichen Sichten als einzelne Knoten in einem Graphen. Die Kanten stellen den zeitlichen Verlauf dar und enthalten zudem die Pose, die die beiden Sichten zueinander registriert. Dabei ist hervorzuheben, dass nicht direkt eine einzelne vollständige Weltrekonstruktion existiert. Dies ist bei modellbasierten Ansätzen der Fall, die in einer einzigen Repräsentation die Welt verwalten. Neue Aufnahmen werden direkt in die bestehende Repräsentation eingefügt. Diese können beispielsweise als Dreiecksnetz, Voxel [Curless96, Newcombe11, Steinbrucker13], Surfels [Keller13, Stückler14] oder sonstige 3D-Strukturen [Taguchi13,Ataer-Cansizoglu13] erfolgen. Neben den SLAM-Ansätzen sind Ansätze der *Digital Shape Reconstruction* verbreitet (stellenweise auch als *Reverse Engineering* bekannt). Ziel dieser Ansätze ist es, ein einzelnes Modell mit hoher Qualität, vor allem im Bereich der Konstruktion, zu erzeugen. Neben der eigentlichen Abbildung sind weiterführende Eigenschaften des Werkstücks von Interesse, beispielsweise Beweglichkeit von Achsen [Varady08]. Die Eingabe ist meist ein bereits vollständig aufgekommenes Werkstück, aus dem ein Modell rekonstruiert werden soll. Einerseits kann ein Modell in mehreren Schritten bestimmt werden, wobei eine oberflächenbasierte Segmentierung [Theologou15] der entscheidende Schritt ist, zum Beispiel durch Regionenwachstum [Besl88, Denker13], Clustering [Garland01, Katz03, Miandarhoie17] oder Primitivenanpassung [Fischler81, Li11, Le17]. Andererseits gibt es komplette Systeme, die automatisch Flächen erkennen oder funktional zusammengehörige Komponenten segmentieren können [Benkő01, Huang03, Várady07, Bénière13]. Entscheidender Nachteil dieser Ansätze ist die mangelnde online Fähigkeit, da für die Rekonstruktion meist ein vollständiges Modell notwendig ist oder Teilprozesse manuell erfolgen.
Die Arbeit [Sand19] verspricht daher einen Ansatz, der die Vorteile der unterschiedlichen Verfahren kombiniert und nach jeder Aufnahme ein valides Modell online erzeugen kann. Einschränkungen dieses Ansatzes ist die Limitierung auf planar rekonstruierbare Objekte sowie die Abwesenheit von Farbinformation. Als Repräsentationsform werden *Boundary Represen-*

tation Models (*B-Reps*) verwendet. Da im Rahmen dieser Arbeit der Fokus auf geometrisch primitiven, untexturierten Objekten als komplexeste Objektklasse liegt, sind diese Einschränkungen hinnehmbar. Das genaue Verfahren wird in Kapitel 2.3 näher vorgestellt.
Ein B-Rep lässt sich als 4-Tupel (F, E, V, B) auffassen. Dabei kann aus der Menge an Knoten V die Menge an Kanten E definiert werden. Aus mehreren Kanten können Begrenzungen abgeleitet werden, zusammengefasst in der entsprechenden Menge B. Eine Fläche aus der Menge F besteht aus einer Menge an äußeren Begrenzungen, um die tatsächliche räumliche Ausdehnung zu beschreiben, sowie optional einer Menge an inneren Begrenzungen um Löcher in der Fläche darstellen zu können. Zur internen Repräsentation von Kanten wird in [Sand19] eine *half-edge* Datenstruktur [Mäntylä87] verwendet. Durch eine definierte Drehrichtung kann über diese abgeleitet werden, ob ein Volumen inner- oder außerhalb des B-Reps ist. Für eine ausführliche Definition siehe Kapitel 3.2 in [Sand19]. Weitere notwendige Eigenschaften werden in dieser Arbeit bei Bedarf eingeführt.
Von den aufgestellten Anforderungen in Abschnitt 1.4.1 in Anlehnung an [Bennamoun02] werden diverse erfüllt: Die Repräsentation ist sehr **effizient**, da anstatt der Punktmengen direkt geometrische Formen verwaltet werden, wodurch auch die **Robustheit** gegenüber Transformationen und Fehlsignalen gegeben ist, da einzelne Ausreißer bei der B-Rep-Erzeugung entfernt werden. Die geometrische Information der Repräsentation wird **genau** wiedergegeben, allerdings wird die Texturinformation nicht erfasst und kann damit auch nicht repräsentiert werden. Die Domäne der Objekte ist teilweise eingeschränkt, da nur planar rekonstruierte Flächen als Ausgabe vorliegen. Die **Lokalität** der Repräsentation ist wiederum vollständig gegeben, da mehrere Sichten aus unterschiedlichen Posen in einem globalen Modell zusammengefügt und verwaltet werden.

2.2.3 Arbeitshypothese und Teilschritte

Durch die Festlegung auf einen Sensor und eine Umweltrepräsentation wird zunächst die zugrundeliegende Arbeitshypothese beschrieben und die wissenschaftlichen Fragestellung bezüglich der gewählten Repräsentation präzisiert.

Arbeitshypothese

Ausgehend von der Wahl der Umweltrepräsentation ist die Arbeitshypothese beziehungsweise Grundannahme für diese Arbeit, dass eine abstraktere Repräsentation der 3D-Daten einen Vorteil gegenüber detaillierten Repräsentationen wie beispielsweise Punktwolken besitzt. Mögliche Vorteile sind dabei die Verfügbarkeit von Zusammenhängen, allem voran die Zuweisung von Flächen, Kanten und Knoten, wodurch der Punktwolke eine geometrische Bedeutung zugewiesen wird. Da die 3D-Daten mit erheblich weniger Deskriptoren beschrieben werden können als durch eine Menge Punkte, sind aufwändigere Suchen auf dieser Repräsentation möglich.

Objektwiedererkennung

Die Herausforderung in Fragestellung F1 liegt in der geringen Information, die über die zu untersuchenden Objekte gegeben ist. Da diese weder markante Textur noch eine eindeutige Form haben, ist zu untersuchen, ob die geometrischen Zusammenhänge von B-Reps helfen, mögliche Objektkandidaten (Hypothesen) in der Umweltrepräsentation zu finden. In diesem Zusammenhang ist von Bedeutung, dass B-Reps ein Volumen mit einer geringeren Anzahl an geometrischen Objekten beschreiben können als Punktwolken, da mehrere Punkte in einer Fläche beziehungsweise Kante zusammengefasst werden. Dabei ist zu beachten, dass die abstrakte Repräsentation von B-Reps im gesamten Prozess beibehalten wird und nicht durch zufälliges Abtasten (*sampling*) auf eine Suche auf Punktebene zurückfällt.

Active Vision

Im Rahmen der wissenschaftlichen Fragestellung F2 ist im Hinblick auf B-Reps von Interesse, wie neue partielle Sichten die Umweltrepräsentation verbessern. Dabei ist zu beachten, dass bei der Rekonstruktion von Punktwolken als B-Rep Information über die Szene verloren gehen kann. Beispielsweise ist dies der Fall, wenn nur eine geringe Anzahl an Punkten für eine Fläche vorliegt und diese nicht rekonstruiert werden kann. Dem gegenüber steht der Nutzen der geometrischen Information durch Flächen und Kanten, die ein gezieltes Betrachten von geometrischen Elementen ermöglicht. Gleichzeitig ist darüber eine Schätzung möglich, wie sich die Szene fortsetzt (beispielsweise geben Kanten einen Hinweis auf eine angrenzende Fläche).

Dynamische Szenen

Die wissenschaftliche Fragestellung F3 teilt sich in zwei Unterfragen: Das ist zunächst der Aspekt, wie Änderungen zwischen zwei Aufnahmen erkannt werden können. Dabei ist durch B-Reps eine genaue Zuweisung zwischen zwei Sichten möglich, wenn diese durch das gleiche geometrische Element wie beispielsweise eine Fläche dargestellt werden. Der zweite Teil der Frage umfasst die Verarbeitung von Änderungen. Zentral ist hier die Notwendigkeit, dass die Umweltrepräsentation dabei ihre Gültigkeit behält, da B-Reps eine interne Struktur vorgeben. Dabei ist gleichzeitig zu beachten, dass die Umweltrepräsentation die tatsächliche Szene möglichst wahrheitsgetreu abbildet, insbesondere Bereiche, die nicht immer sichtbar sind.

Semantik

Für die Beantwortung der letzten Fragestellung bezüglich Semantik weißt der entsprechende Stand der Forschung auf Uneindeutigkeiten hin, was unter Semantik zu verstehen ist. Darüber hinaus muss untersucht werden, inwieweit B-Reps in die Definition von Semantik einstrukturiert werden können. Für die bestehende Beschreibung von Semantik können die Vorteile der abstrakteren Datenstruktur diskutiert werden. Analog zur Definition muss für die Extraktion von Semantik geprüft werden, ob die Nutzung von B-Reps einen Mehrwert bietet.

2.3 Erstellen von B-Reps aus Punktwolken

Da in der Arbeitshypothese eine spezielle Form der Umweltrepräsentation gewählt wurde, erläutern die nächsten Absätze, wie diese erzeugt werden kann. Der Gesamtprozess wird im Detail in [Sand19] (für Teilschritte siehe auch [Sand16, Sand17, Rohner20b]) dargestellt. Dazu wird zunächst beschrieben, wie eine einzelne Punktwolke in ein B-Rep umgewandelt wird. Ergänzend werden die wichtigsten Schritte und die Anpassung für eine robotermontierte Tiefenkamera diskutiert. Darauf aufbauend kann ein Roboter eine vollständige Szene rekonstruieren. Abschließend wird die Parametrierung des Systems vorgestellt.

2.3.1 Rekonstruktion eines einzelnen B-Reps

Die Rekonstruktion eines einzelnen B-Reps aus einer Sicht besteht zunächst aus einer Segmentierung für jedes planare Teilstück und anschließend einer Polygonalisierung ebendieser Segmente. Abschließend werden alle Segmente zu einem B-Rep vereint.

Segmentierung

Als Eingabe zur Rekonstruktion eines einzelnen (*partiellen*) B-Reps wird eine organisierte Punktwolke verwendet. Diese muss zunächst in einzelne Regionen segmentiert werden, welche anschließend als Flächen repräsentiert werden können. Als erster Schritt der Segmentierung werden die gegebenen Punkte vernetzt, wodurch sich neue Nachbarschaftsbeziehungen ergeben, die Tiefenunterschiede berücksichtigen. Eine direkte Nachbarschaft ist bereits durch die Organisiertheit der Punktwolke gegeben. Auf Basis dieser Vernetzung kann zudem für jeden Punkt der Normalenvektor geschätzt werden.
Als zweiter Schritt wird ein Regionenwachstum durchgeführt mit dem Ziel, zusammengehörige, planare Segmente zu bestimmen. Zu einer bestehenden Region wird dann ein Punkt hinzugefügt, wenn sowohl der Abstand des Punktes zur bisherigen Ebene der Region als auch der Winkel zwischen Punkt- und Flächennormalen kleiner als ein Schwellwert ist. Dabei wird ein *greedy* Ansatz verfolgt, der auf Basis eines zufällig gewählten Startpunktes alle kompatiblen Punkte zu dieser Region hinzufügt. Dies wird solange wiederholt, bis alle Punkte genau einer Region zugeordnet sind. Da in diesem Fall das Ergebnis von der (zufällig bestimmten) Reihenfolge abhängt, wird in einem zweiten Schritt die Segmentierung erneut durchgeführt, allerdings in umgekehrter Reihenfolge der betrachteten Regionen. Die Ergebnisse dieser beiden Segmentierungen werden verglichen, indem für alle benachbarten Punkte einer Region betrachtet wird, ob diese gemäß dem Segmentierungskritierium besser zur aktuell betrachteten Region passen.
Abschließend werden die berechneten Segmente gefiltert, sodass das Rauschen reduziert wird und kleine Segmente entfernt werden, die für den weiteren Verlauf irrelevant sind.

Polygonalisierung

Die bestimmten Segmente werden im Weiteren in Polygone umgewandelt, welche anschließend Eingabe für die B-Rep-Erzeugung sind. Gemäß den Annahmen sowie Einschränkungen

der Welt müssen diese Polygone unterschiedliche Eigenschaften erfüllen: Ein Polygon kann Löcher enthalten (beispielsweise wenn einzelne Punkte in der Punktwolke fehlen). Aufgrund physikalischer Einschränkungen dürfen sich Kanten und Ecken allerdings nicht schneiden. Da ausschließlich planare B-Reps rekonstruiert werden sollen, müssen alle Eckpunkte eines Polygons in einer Ebene liegen (wodurch das gesamte Polygon in dieser Ebene liegt). Die Ausgabe erfolgt im Dreidimensionalen.
Zunächst wird die Kontur des Polygons bestimmt, wobei ein knotenbasierter Ansatz verfolgt wird. Da eine organisierte Punktwolke vorliegt, wird dieser Schritt zunächst in 2D durchgeführt. Dazu wird ein Segment als 2D-Binärbild repräsentiert und die Kontur von einem Ausgangspunkt aus verfolgt, bis dieser wieder erreicht ist. Diese 2D-Kontur wird anschließend mit Hilfe der bekannten Kameraparameter zurück ins Dreidimensionale projiziert.
In den weiteren Schritten werden die Kanten der Polygone bestimmt. Dafür werden zunächst die Kanten berechnet und vereinfacht, die im Dreidimensionalen benachbart sind. Dazu werden Ecken und Kanten, die sehr nah aneinander liegen, zu einem Kantensegment zusammengefasst. Anschließend werden Kanten betrachtet, die durch einen Tiefensprung entstanden sind. Hier werden allerdings keine benachbarten Kanten vereinfacht, sondern nur die Kante selbst, wenn mehrere Eckpunkte zusammen durch eine Kante dargestellt werden können (beispielsweise wenn mehrere Eckpunkte auf einer Kante liegen). Abschließend werden auf die gleiche Weise Kanten vereinfacht, die am Rand der Messung existieren. Diese drei Kantentypen werden im Active Vision Kapitel (Abschnitt 4.2.1) weiter betrachtet.

B-Rep-Erzeugung

Die Polygone können nun in ein B-Rep umgewandelt werden. Diese Transformation erfolgt direkt, da die Unterscheidung der Kanten im bevorstehenden Schritt auf die Datenstruktur der B-Reps abgebildet werden kann. Als Nachverarbeitungsschritt werden die Ecken angepasst, wenn mehr als zwei Flächen in dieser zusammenlaufen. Anschließend werden die Ecken und Ebenengleichungen iterativ weiter optimiert. Die so entstehenden Flächen erhalten ein Bewertungsmaß basierend auf der Anzahl der zugrundeliegenden Punkte.

2.3.2 Rekonstruktion einer Szene aus mehreren Sichten

Mit dem bisherigen Verfahren kann eine einzelne Punktwolke in ein B-Rep transformiert werden. Im Weiteren wird dieser Prozess ergänzt, sodass mehrere B-Reps aus unterschiedlichen Sichten fusioniert werden können. Dazu wird zunächst der Ansatz der Fusion unter Verwendung einer handgehaltenen Tiefenkamera beschrieben. Abschließend folgt eine Erläuterung, wie eine vollständige Szene erfasst werden kann und mit Nachverarbeitungschritten verfeinert wird.

Fusion von zwei partiellen B-Reps

Für die Fusion sind zwei B-Reps mit bekannter Pose im Raum gegeben. Um diese zu fusionieren, ist es zunächst erforderlich, alle Flächen zu bestimmen, die in beiden B-Reps repräsentiert werden. Ob zwei Flächen korrespondieren, wird anhand des Abstands, Winkels

und der überlappenden Flächen entschieden. Falls alle Bedingungen zutreffen, existiert eine Korrespondenz. Diese Flächen werden nun fusioniert, wobei dies analog zur Polygonalisierung im Zweidimensionalen durchgeführt wird. Dazu werden die beiden Flächen im Dreidimensionalen gemittelt und auf dieser neuen Ebene unter Anwendung des Map-Overlay-Algorithmuses [de Berg08] fusioniert. Das Ergebnis dieses Verfahrens erzeugt valide Flächen und Ecken, wobei die Kanten allerdings noch nicht der B-Rep Repräsentation genügen. In einem weiteren Schritt werden alle möglichen Probleme bei der Kantenerstellung behoben, indem zunächst Begrenzungsintervalle bestimmt und durch entsprechende Kanten ersetzt werden. Abschließend erfolgt eine Optimierung der fusionierten Ecken und Ebenen.
Für alle Flächen, die keine Korrespondenz haben, erfolgt dieses Verfahren analog, nur dass keine Mittelung über mehrere Ebenengleichungen notwendig ist.

Registrierung von partiellen B-Reps

Für die Fusion ist es notwendig, die Pose der beiden B-Reps zu kennen. In der zugrundeliegenden Literatur [Sand19] wird ein Verfahren zur Registrierung vorgeschlagen, wenn die Pose nicht bekannt ist. Da für diese Arbeit von einer robotermontierten Tiefenkamera ausgegangen wird, ist die Pose der beiden B-Reps bereits bei der Aufnahme der Punktwolke bekannt. Der entwickelte Ansatz findet im Rahmen dieser Arbeit trotzdem Anwendung im nächsten Kapitel zur Objektwiedererkennung. Für die Registrierung mittels Pose des Roboters ist es zunächst nötig, die Kamera extrinsisch auf den TCP des Roboters zu kalibrieren, was sich beispielsweise mit Hilfe eines Schachbrettmusters realisieren lässt. Die Transformation des Roboters zwischen der Kamera und seiner Basis (in diesem Anwendungsfall gleichzeitig auch der Ursprung des Weltkoordinatensystems) ist mit Hilfe der bauartbedingten und somit bekannten Kinematik bestimmbar. Die B-Reps werden in das Weltkoordinatensystem transformiert und in diesem fusioniert.

Rekonstruktion einer Szene

Um eine vollständige Szene zu rekonstruieren, wird eine Punktwolke an einer Pose aufgenommen und mit oben beschriebener Methodik in ein B-Rep umgewandelt. Dieses wird in Weltkoordinaten transformiert und die entsprechende Pose hinterlegt. Der Roboter bewegt sich nun an eine neue Pose, beispielsweise durch den Menschen geführt oder autonom (siehe Kapitel Active Vision, Abschnitt 4.2.1). An dieser neuen Pose wird wieder eine Punktwolke aufgenommen, in ein B-Rep umgewandelt und in das Weltkoordinatensystem transformiert. Dort werden beide B-Reps fusioniert. Dieser Prozess kann nun beliebig wiederholt werden, wobei neue Sichten in die stetig wachsende Umweltrepräsentation integriert werden. Damit ist die Anforderung, dass nach jeder aufgenommenen Punktwolke eine gültige Umweltrepräsentation vorhanden sein muss, erfüllt.

Nachverarbeitungsschritte

Nach jeder Fusion von zwei B-Reps kann die resultierende Repräsentation verbessert werden, indem Löcher automatisch geschlossen werden. Dabei werden diese Unvollständigkei-

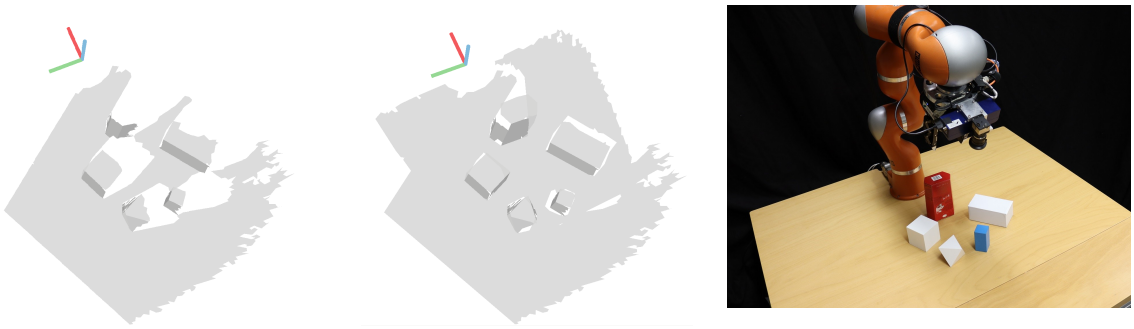


Abbildung 2.1: Rekonstruktion einer Szene in mehreren B-Reps: Aus einer Sicht (links) und fusioniert mit einer zweiten Sicht (mitte), rekonstruiert von oben (rechts). Das Weltkoordinatensystem (im Roboterurspung) ist in den Rekonstruktionen zusätzlich abgebildet.

ten geschlossen, die in einer Fläche oder entlang einer Kante liegen, sowie fehlende Ecken ergänzt. Weiterhin existieren Methoden um Nutzerrückmeldungen auf Basis der aktuellen Rekonstruktion zu bestimmen, wodurch ein Hinweis gegeben wird, wo sich Unvollständigkeiten befinden. Da diese für einen menschlichen Nutzer und nicht für eine robotermontierte Tiefenkamera ausgelegt sind, können sie nicht direkt angewandt werden. Für eine Vervollständigung der Umweltrepräsentation sei auf Kapitel 4 (Active Vison) verwiesen.

2.3.3 Parametrierung

Im gesamten Prozess der Rekonstruktion von B-Reps aus Punktwolken sind diverse Parameter erforderlich, beispielsweise verschiedene Schwellwerte. Dazu wird in [Sand19] die *Strukturgröße* als einzig notwendiger Parameter eingeführt. Die Strukturgröße ist ein eindimensionaler Parameter, der beschreibt, welche Größe geometrische Elemente aufweisen müssen (beispielsweise Kanten und Flächen), damit diese rekonstruiert werden. Dieser Parameter beschreibt grundsätzlich eine Länge, kann aber auf Flächen und Volumen analog angewendet werden. Alle notwendigen Parameter zur Rekonstruktion und Fusion von B-Reps hängen direkt von diesem ab, sodass zur Parametrierung nur dieser Wert bestimmt werden muss. Die Wahl dieser Länge hängt unter anderem von der Größe der zu untersuchenden Objekte, der verwendeten Tiefenkamera (und dem damit verbundenen Rauschen) und der gewünschten Qualität der Rekonstruktion ab. Diese Aspekte sind untereinander nicht unabhängig, da die Qualität der Rekonstruktion von der Tiefenkamera abhängt und diese durch die Art der Objekte eingeschränkt ist.

2.4 Zusammenfassung

Ausgehend von dem Stand der Forschung zum Thema Umweltwahrnehmung in der Robotik wurde die Arbeitshypothese formuliert, dass eine abstrakte Umweltrepräsentation Vorteile im Verarbeitungsprozess aufweist. Als verwendete Sensorik wurde ein minimaler Aufbau in Form einer einzelnen, vollständig kalibrierten Eye-in-Hand Kamera festgelegt, montiert am Endeffektor des Roboterarms. Als verwendete Umweltrepräsentation werden im Rahmen dieser Arbeit planar-rekonstruierbare B-Reps genutzt. Zum einen, da diese online und vollau-

tomatisiert mit geringem Parametrierungsaufwand erstellt werden können. Zum anderen, da eine texturlose, planare Rekonstruktion für die Ziele dieser Arbeit ausreichend ist. Als Grundlage für die weiteren Kapitel wurde erläutert, wie aus einzelnen Punktwolken sowohl partielle B-Reps als auch eine einzelne Umweltrepräsentation erzeugt werden können.

Kapitel 3

Objektwiedererkennung mit B-Reps

Inhalt

Dieses Kapitel beschreibt den ersten Schritt zur Umwelterfassung mit B-Reps, die Objektwiedererkennung. Dazu wird der Stand der Forschung dargestellt, mit welchen Methoden und welchen Eingabedaten Objekte wiedererkannt werden können (Abschnitt 3.1). Davon ausgehend wird ein Ansatz entwickelt, um mögliche Objektinstanzen (Hypothesen) zu erzeugen. Aus dieser Menge an Hypothesen wird anschließend eine möglichst gut geeignete Teilmenge ausgewählt (Abschnitt 3.2). Für die Evaluation werden zunächst die verwendete Hardware sowie verwendete Testobjekte vorgestellt (Abschnitt 3.3). Der Ansatz wird auf Basis der Ergebnisse abschließend diskutiert. Dieses Kapitel erweitert die Vorarbeit [Rohner19b].

3.1 Stand der Forschung

Da im Rahmen der Objektwiedererkennung Begriffe ähnlich verwendet werden, wird zunächst ein kurzer Überblick über die Namensgebung und nahe Forschungsfelder dargestellt, um eine einheitliche Nomenklatur sicherzustellen. Anschließend folgt ein ausführlicherer Stand der Forschung explizit zu den unterschiedlichen Gruppen und Ansätzen der Objektwiedererkennung.

3.1.1 Nomenklatur

Im Bereich der Analyse von Bildern gibt es primär drei Gruppen an Verfahren, die zum Verständnis einer Szene beitragen: Das sind die (semantische) Segmentierung, die Objekterkennung (*Object detection*) und die Objektwiedererkennung (*Object Recognition*).
Bei einer Segmentierung ist das Ziel, die gegebene Szene in zusammenhängende Bereiche zu unterteilen. Die weiterführende Bedeutung der Segmente hängt dabei von den gewählten Kriterien ab. So wird in der Vorarbeit [Sand19] segmentiert, um planare Flächen zu entdecken. Alternativ ist das Ziel einer Segmentierung, komplette Objekte zu identifizieren. Dabei ist das Ergebnis nicht, welches Objekt vorliegt, sondern nur die Aussage, welcher Bildbereich ein einzelnes Objekt repräsentieren könnte. Die semantische Segmentierung ist insofern eine Erweiterung, da jedem Segment zusätzlich noch eine semantische (beispielsweise natürlichsprachliche) Klasse zugewiesen wird. Dabei handelt es sich nicht um individuelle Objektinstanzen, sondern um die Zugehörigkeit zu einer allgemeinen Klasse (wie zum Beispiel *Wand*, *Tisch* oder *Boden* für Innenräume). Als Alternative werden bei der Objekterkennung *Bounding Boxen* so in das Bild eingefügt, dass jede Bounding Box einem Objekttyp entspricht. Analog werden hier keine Objektinstanzen erzeugt, und die genaue Postion und Orientierung des Objektes durch eine Bounding Box ist zudem nicht präzise beschrieben. Als letzter Ansatz wird bei der Objektwiedererkennung eine Szene so analysiert, dass für jedes Objekt in ihr eine eigene Instanz erzeugt wird. Jede Instanz verfügt dabei zumindest über eine eigene Pose, die die Position und Orientierung im Raum genau beschreibt. Darüber hinaus können ein Modell des Objektes sowie weitere Informationen vorliegen.

3.1.2 Verwandte Arbeiten

Wie in der Einleitung dargelegt, ist es das Ziel, Objektinstanzen zu erzeugen. Daher beschränkt sich diese Arbeit im weiteren Verlauf primär auf Methoden der Objektwiederkennung. Diese kann man anhand von drei Kriterien unterscheiden: Zum einen danach, welche Information der Eingabe genutzt wird. In Kamerabildern kann sowohl Farbinformation vorliegen als auch Oberflächeninformation. Wissen über die Oberfläche ist entweder explizit gegeben, da 3D-Daten verwendet werden, oder kann aus 2D-Bildern geschätzt werden. Darüber hinaus können beide Merkmale gleichzeitig als hybrider Ansatz verwendet werden. Zum anderen ist für manche Ansätze eine objektbasierte Segmentierung als Vorverarbeitungsschritt notwendig. Diese Verfahren benötigen explizit einen Ausschnitt der Szene als Eingabe, der ein einzelnes Objekt repräsentiert. Dies hat den Nachteil, dass eine objektbasierte Segmentierung kein leicht zu lösendes Problem und fehleranfällig ist; andererseits ist das Wissen, dass genau

ein Objekt vorliegen muss, hilfreich für die Objektwiedererkennung. Abschließend kann die Methodik der Ansätze unterschieden werden: Vor allem in den letzten Jahren sind lernbasierte Ansätze populärer geworden, meist in Verbindung mit neuronalen Netzen. Die Alternative sind modellbasierte Verfahren, die auf expliziten Merkmalen der Objekte arbeiten. Weitere Kriterien sind zudem die Dimension der Eingabedaten (2D oder 3D) sowie die Repräsentationsform. Da im Rahmen dieser Arbeit die Dimension (3D) und die Repräsentation (B-Reps) bereits diskutiert und festgelegt wurden, wird diese Unterscheidung im Stand der Forschung nicht weiter betrachtet. Im Weiteren wird zwischen 2D- und 3D-Bildern sowie lernbasierten Verfahren unterschieden.

Als bekanntester Ansatz mit 2D-Farbbildern als Eingabe hat sich *SIFT* [Lowe99, Lowe04] etabliert - mit einigen Abwandlungen wie beispielsweise *SURF* [Bay06], *PCA-SIFT* [Ke04] und *Affine-SIFT* [Yu09]. Dabei werden *Keypoints* sowohl in der Szene als auch auf Einträgen in einer Objektdatenbank bestimmt, verglichen und gruppiert, woraus sich Hypothesen für die einzelnen Objekte bestimmen lassen. Eine andere Möglichkeit, um in Farbbildern Objekte wiederzuerkennen, ist das sogenannte *Template Matching* [Kim07, Korman13, Park19]. Dieser Ansatz vergleicht den aktuellen Szeneneindruck mit der Objektdatenbank und versucht über unterschiedliche Ähnlichkeitsmaße, Objekte wiederzuerkennen. Die entscheidende Information ist in beiden Fällen die Farbinformation. Die eigentliche 3D-Pose kann über dazugehörige Tiefendaten berechnet werden.

Andere Ansätze verwenden ausschließlich die Oberflächeninformation (beispielsweise Normalen und Krümmung) der 3D-Daten. Für zwei gegebene Punkte und deren Normalen kann das sogenannte *Point Feature Histogram (PFH)* [Rusu08] mit der Erweiterung des *Fast Point Feature Histogram* [Rusu09b] berechnet werden. Dieser Ansatz wird im sogenannten *SHOT*-Deskriptor weiterverwendet [Tombari10]. Andere Weiterentwicklungen umfassen das *Viewpoint Feature Histogram* [Rusu10], *Clustered Viewpoint Feature Histogram (CVFH)* [Aldoma11], *Camera Roll Histogram* [Aldoma12b] und das *Oriented, Unique and Repeatable CVFH* [Aldoma12a]. Anstatt der ausschließlichen Nutzung der Normalen kann zusätzlich der Abstand zwischen den Punkten verwendet werden [Buchholz13]. Neben Punktwolken als Repräsentationsform können auch CAD-Daten verwendet werden. Eine Möglichkeit umfasst das Abtasten der Punktwolken, um anschließend punktbasierte Verfahren darauf anzuwenden [Somani13]. Alternativ können Objektprimitive an die CAD-Daten angepasst werden, womit eine Objektwiedererkennung ermöglicht wird [Somani15]. Diese Ansätze können beispielsweise für mobile Roboter im Haushalt verknüpft werden [Koppula11] [Wu14].

Vor allem im letzten Jahrzehnt wurden neben den modell-und regelbasierten Methoden Ansätze mit künstlichen neuronalen Netzen entwickelt. Dabei sind vor allem *Convolutional Neural Networks* von Interesse (neben den bereits vorgestellten Transformer-Netzen aus Kapitel 2.). Eine Möglichkeit sind neuronale Netze für die Objektdetektion, beispielsweise im 2D mit [Redmon17] (und Erweiterungen [Terven23]) aufbauend auf [Szegedy15] (siehe auch [Chen23a] für einen Überblick). Alternativ kann mit neuronalen Netzen eine semantischen Segmentierung berechnet werden, sowohl mit 2D-Daten [Yu15, Badrinarayanan17, Takikawa19] als auch 3D beziehungsweise multimodal [Silberman12, Valada17, McCormac16]. Eine weitere Alternative ist Objektklassifikation [Krizhevsky12]. Eine erste Arbeit für die Objektwiedererkennung mit

3D-Daten war *VoxNet* [Maturana15]. Als Eingabe wurde eine teilweise segmentierte Punktwolke mit vordefinierten Objektinstanzen klassifiziert. Für diese Aufgabe existieren diverse Methoden mit unterschiedlichen Architekturen, siehe [Chen23a, Muzahid24].
Als Ergebnis des Stands der Forschung und als Zielstellung für dieses Kapitel sind mehrere Aspekte festzuhalten: Die größte Herausforderung für die Objektwiedererkennung sind geometrisch einfache, untexturierte Objekte, da diese kaum Information tragen (siehe auch Herausforderungen in Kapitel 1). Der zu entwickelnde Ansatz soll mit dieser Herausforderung umgehen können. Dabei ist es möglich, sich ausschließlich auf die Oberflächeninformation zu konzentrieren, um zu untersuchen, inwieweit B-Reps ohne zusätzliche Information für die Objektwiedererkennung geeignet sind. Darüber hinaus sollte aufgrund des damit verbundenen Fehlerpotentials eine objektspezifische Segmentierung vermieden werden. Da das Ziel die Verwendung des Roboters im Haushalt ist, ist es notwendig, dass neue Objekte mit geringem Aufwand hinzugefügt werden können, ohne dass der Erkennungsansatz neu gelernt werden muss. Unter diesem Aspekt ist ein modellbasierter Ansatz vielversprechend, da B-Rep Modelle von Werkstücken entweder durch den Hersteller vorhanden sind oder nachträglich erzeugt werden können. Weiterhin ist ein Ansatz notwendig, der mit einer stetig wachsenden Umwelt umgehen kann. Da im Laufe der Zeit mehr Sichten und somit mehr Objekte in der Umweltrepräsentation hinterlegt werden, muss die Methodik in dieser Größe skalieren. Zusätzlich soll mit Verdeckungen rein visuell umgegangen werden können. Dabei ist zu beachten, dass bei zu großen Verdeckungen einzelne Objekte unmöglich wiedererkannt werden können, da diese vom Sensor nicht erfassbar sind. Das bedeutet für den Wiedererkennungsansatz, dass auch mit wenig Information (beispielsweise einer Ecke) ein Objekt korrekt erkannt werden soll. Hier existiert eine weitere mathematische Einschränkung, da zumindest drei Punkte gegeben sein müssen, die nicht paarweise in einer Ebene liegen, um eine Ecke durch drei Flächen genau zu beschreiben.

3.2 Beschreibung des Verfahrens

Im Weiteren wird das gesamte Verfahren beschrieben. Dazu werden der allgemeine Ansatz und die Teilschritte vorgestellt. Der erste Schritt dabei ist, eine Menge an Objekthypothesen zu erzeugen. Dafür werden ausschließlich geometrische Merkmale verwendet. Aufgrund des hohen Abstraktionsniveaus durch B-Reps kann eine große Menge an Hypothesen generiert werden. Da diese sich teilweise untereinander widersprechen (beispielsweise wenn zwei Hypothesen sich schneiden, wodurch die Objektvolumina überlappen), muss in einem zweiten Schritt eine Menge an geeigneten Hypothesen ausgewählt werden. Auf diese Weise wird die erste wissenschaftliche Fragestellung

F1 Inwieweit können geometrisch primitive, untexturierte Objekte wiedererkannt werden?

hinsichtlich der Arbeitshypothese untersucht.

3.2.1 Ansatz

Die Hypothesenerzeugung basiert auf dem Ansatz in [Sand19]: Zwei B-Reps mit unbekannter Pose sollen zueinander registriert werden. Dieser Ansatz wird gewählt, da er einigen der Anforderungen genügt und ebenda bereits positiv evaluiert wurde. Zunächst basiert der Registrierungsansatz ausschließlich auf der Oberflächeninformation, die durch die B-Reps gegeben wird, allem voran die Flächennormale. Weiterhin ist keine objektspezifische Segmentierung notwendig. Anstatt dass zwei rekonstruierte B-Reps registriert werden sollen, wird hier eine Umweltrepräsentation mit einer Objektdatenbank abgeglichen. In der Objektdatenbank werden alle Modelle von möglichen Objekten als B-Rep hinterlegt. Dadurch genügt der Ansatz ebenfalls dem Ziel der schnellen Erweiterung um neue Objekte. Aus diesem Grund wird im Weiteren davon ausgegangen, dass alle Objekte in der Szene auch in der Objektdatenbank hinterlegt sind und somit keine unbekannten Objekte auftreten können. Die Arbeitsoberfläche ist bekannt und kann aufgrund des registrierten Sensors bei Bedarf aus der Umweltrepräsentation entfernt werden. Inwieweit dieser Ansatz mit Verdeckungen umgehen kann und auch bei steigender Szenengröße performant bleibt, wird in der Evaluation untersucht.
Zur Auswahl der Objekthypothesen können *Hypothesenverifikationsverfahren* herangezogen werden (siehe beispielsweise [Aldoma16, Bauer20]). Dabei ist zunächst wichtig, dass physikalisch unmögliche Situationen aufgelöst werden, insbesondere dann, wenn sich Hypothesen schneiden. Dabei ist allerdings zu beachten, dass aufgrund von Rauschen und Ungenauigkeiten bei der Hypothesenerzeugung eine geringe Überschneidung auftreten kann, ohne dass dadurch Hypothesen falsifiziert werden sollen (vor allem wenn sich die Objekte in der Realität berühren). Ein weiteres Problem ist die Auswahl der Hypothesen: Falls diese gierig nach ihrer Qualität sortiert werden, können vor allem bei ähnlichen Objekten lokale Minima auftreten. In diesem Fall wird zwar eine Hypothesenauswahl getroffen die ein sehr hohes Qualitätsmaß erreicht, diese ist aber nicht die optimale Auswahl. Da eine komplette Optimierung aufwendig ist, ist ein *anytime* fähiger Ansatz erforderlich, der bei Bedarf abgebrochen werden kann, aber trotzdem ein valides Ergebnis generiert hat.

3.2.2 Hypothesenerzeugung

Für die Hypothesenerzeugung steht zunächst eine Umweltrepräsentation zur Verfügung, die aus mindestens einer Sicht erstellt wurde. Die Umwelt ist dabei ein B-Rep $W = (F, E, V, B)$, welches aus Flächen F, Halb-Kanten E, Knoten V und Begrenzungen B besteht. Für weiterführende Formalisierung und Zusammenhänge zwischen den einzelnen Komponenten wird an dieser Stelle auf Kapitel 3 in [Sand19] verwiesen. Grundsätzlich werden aus den Knoten eines B-Reps die Kanten gebildet und aus den Kanten die planaren Flächen. Zu den Flächen gehören noch die Begrenzungen, da Flächen in diesem Zusammenhang Löcher beinhalten können. Neben der Szene existiert eine Objektdatenbank $\Omega = \{M_1, ..., M_o\}$, in der insgesamt o Objektmodelle gehalten werden. Dabei besteht ein Eintrag M_i mindestens aus einem B-Rep, welches das zugrundeliegende Objekt repräsentiert. Um nun mögliche Hypothesen für die unterschiedlichen Einträge in der Objektdatenbank zu bestimmen, muss ein Deskriptor aufgestellt werden. Dieser wird sowohl auf der Szene W bestimmt als auch für jeden Eintrag $M_i \in \Omega$. In [Sand19] werden dazu drei linear unabhängige Flächen $f_1, f_2, f_3 \in F_W, f_1 \neq f_2 \neq f_3 \neq f_1$ genutzt, wobei F_W

die Menge aller Flächen aus der Szene W beschreibt. Die drei Flächen sind deshalb wichtig, da sonst keine eindeutige Pose für das Modell aus der Objektdatenbank bestimmt werden könnte. Somit lässt sich der Deskriptor beschreiben als

$$D(f_1, f_2, f_3) := \begin{pmatrix} \alpha_{23} & \vec{p}_1^T \\ \alpha_{13} & \vec{p}_2^T \\ \alpha_{12} & \vec{p}_3^T \end{pmatrix} \tag{3.1}$$

wobei $\vec{p}_i^T = (\vec{n}_i^T z_i)$ die Ebenenkoeffizienten der Fläche f_i sind, bestehend aus dem normalisierten Normalenvektor $\vec{n}_i \in \mathbb{R}^3$ von f_i und dem Abstand $z_i \in \mathbb{R}$ der Ebene zum Ursprung (die Szene W ist über die Roboterkinematik in das Weltkoordinatensystem transformiert). Der Wert von $\alpha_{ij} \in [0, 360]$ beschreibt den Innenwinkel zwischen f_i und f_j mit $i, j \in \{1, 2, 3\}$ für $i \neq j$. Für einen Deskriptor, der auf Basis der Szene erstellt wurde, und einem von den Objektmodellen generierten, muss geprüft werden, ob diese korrespondieren und sich so für eine Hypothese eignen. Dabei wird die Korrespondenz über die Ähnlichkeit der berechneten Innenwinkel bestimmt. Dies stellt sicher, dass mindestens die Form (welche über die Flächennormalen beschrieben wird) übereinstimmt. Die Normalen $\vec{n}_1, \vec{n}_2, \vec{n}_3$ bilden hierbei ein rechtshändiges Koordinatensystem. Die Ähnlichkeit von zwei Deskriptoren D_M, D_W von einem B-Rep des Modells $M \in \Omega$ und der Umweltrepräsentation W berechnet sich als das Minimum der maximalen Differenz zwischen allen möglichen Winkelpermutationen. Da die Reihenfolge der Winkel im Deskriptor willkürlich gewählt wurde und keine Ordnung vorab bekannt ist, müssen alle Permutationen geprüft werden. Die Ähnlichkeit bestimmt sich somit als

$$s(D_M, D_W) = \min\{s_0, s_1, s_2\} \tag{3.2}$$

wobei s_0, s_1, s_2 die drei möglichen Permutationen der Innenwinkel darstellt:

$$s_0 = \max\{\alpha_{23} - \beta_{23}, \alpha_{13} - \beta_{13}, \alpha_{12} - \beta_{12}\} \tag{3.3}$$

$$s_1 = \max\{\alpha_{23} - \beta_{13}, \alpha_{13} - \beta_{12}, \alpha_{12} - \beta_{23}\} \tag{3.4}$$

$$s_2 = \max\{\alpha_{23} - \beta_{12}, \alpha_{13} - \beta_{23}, \alpha_{12} - \beta_{13}\} \tag{3.5}$$

Die Winkel α_{ij} entsprechen den Winkeln aus dem Deskriptor D_W und β_{ij} den Winkeln aus D_M. Das resultierende Ähnlichkeitsmaß s wird abschließend noch mit einem Schwellwert verglichen, um zu entscheiden, ob die zwei Deskriptoren sich ähnlich sind.
Für zwei gegebene Deskriptoren in einer Hypothese kann schließlich die Transformation zwischen diesen berechnet werden. Im Fall der Objektwiedererkennung entspricht dies der Pose im Weltkoordinatensystem. Der translatorische Teil kann direkt bestimmt werden. Der rotatorische Anteil kann über die drei linear unabhängigen Winkel bestimmt werden. Dabei ist zu beachten, dass eine Starrkörpertransformation bestimmt wird, was über eine Singulärwertzerlegung sichergestellt ist (siehe [Sand17], [Sand19]).
Die berechnete Transformation bildet zusammen mit dem Modell des Objekts die Grundlage für eine Hypothese. Diese muss aber noch um ein Qualitätsmaß erweitert werden, damit Hypothesen untereinander verglichen werden können, was für die Hypothesenauswahl ent-

scheidend ist. Das Qualitätsmaß richtet sich danach, wieviel Fläche der Welt W durch das Modell erklärt wird, wenn dieses ebenfalls in Weltkoordinaten transformiert wird. Dazu wird zunächst bestimmt, welche Flächen des transformierten Objektmodells mit welchen Flächen der Szene übereinstimmen, indem der Abstand und Winkel zwischen den Ebenengleichungen verglichen wird und eine Mindestgröße der Überschneidungsfläche sichergestellt wird (siehe auch die Definition `explainedby` in Kapitel 4). Somit kann für zwei Flächen das Qualitätsmaß $q(f_k, f_l)$ als der Anteil des Schnittfläche gegen die Fläche der Vereinigung bestimmt werden (auch bekannt als *Jaccard-Koeffizient*). Eine der Flächen gehört dabei zu dem aktuell untersuchten Eintrag der Objektdatenbank und die andere Fläche zur globalen Umweltrepräsentation W. Diese Berechnung wird für jedes Paar an korrespondierenden Flächen bestimmt und ergibt so das Gesamtqualitätsmaß

$$q(f_k, f_l) := \frac{\text{area}(f_k \cap f_l)}{\text{area}(f_k \cup f_l)} \text{ und } q := \sum q(f_k, f_l). \tag{3.6}$$

Das Ziel des Qualitätsmaßes ist es, dass möglichst alle vorhandenen Flächen gut durch die Hypothese erklärt werden und nicht nur eine einzelne, große, korrespondierende Fläche das Qualitätsmaß dominiert. Aus diesem Grund wird nicht das aufsummierte, absolute Überschneidungsmaß über alle korrespondierenden Flächen verwendet, sondern die Summe über die einzelnen, relativen Überschneidungsmaße. Somit werden mehrere kleine Flächenstücke gegenüber wenigen großen bevorzugt. Das hat allerdings zur Folge, dass das Qualitätsmaß nicht auf 1 normalisiert werden kann. Da im Bewertungsmaß die tatsächliche, metrische Größe des Flächeninhaltes eingeht, ist dieser Ansatz ebenfalls robust gegenüber Objekten in der Datenbank mit identischen Deskriptoren, aber unterschiedlicher Größe.
Insgesamt kann man so eine einzelne Hypothese definieren als $h = (M, T_W^M, q)$, welche aus einem Modell $M \in \Omega$ aus der Objektdatenbank, der Starrkörpertransformation $T_W^M \in SE(3)$ zwischen M und W und dem Qualitätsmaß $q \in \mathbb{R}^+$ besteht.
Dieser Prozess muss für jedes Objekt aus der Objektdatenbank durchgeführt werden. Darüber hinaus können pro Objekt unterschiedlich viele Hypothesen erzeugt werden. Um auch im Fall von wenig sichtbarer Oberflächeninformation pro Objekt diese korrekt wiederzuerkennen, ist es notwendig, alle möglichen Hypothesen zu erzeugen. Somit ist der Aufwand zur Hypothesenerzegung quadratisch, da jede Fläche der Welt mit jeder Fläche der Objektdatenbank verglichen werden muss (siehe auch [Sand19]). Die Hypothesen können basierend auf dem Qualitätsmaß bereits vor der Hypothesenauswahl gefiltert werden. Da für den Deskriptor nur die Objektform beziehungsweise die Winkel zwischen den einzelnen Flächen, aber nicht die Größe relevant ist, werden vor allem für symmetrische Objekte (beispielsweise Quader) viele Hypothesen erzeugt, die ein geringes Qualitätsmaß haben. Durch die erschöpfende Suche über alle möglichen Deskriptoren wird gleichzeitig sichergestellt, dass ein Objekt auch dann erkannt werden kann, wenn es mit mehreren Instanzen in einer Szene vertreten ist. Insgesamt erhält man eine Menge $\Lambda = \{h_1, ..., h_m\}$ von m möglichen Hypothesen, aus der im nächsten Schritt geeignete Hypothesen ausgewählt werden.

3.2.3 Hypothesenauswahl

Um die Objekte in der Szene W bestmöglich zu erklären, muss eine Teilmenge $H \subseteq \Lambda$ bestimmt werden. Das grundsätzliche Kriterium ist dabei das Qualitätsmaß, was addiert über alle Hypothesen in H maximiert werden soll, da auf diese Art am meisten Fläche der Szene durch die Hypothesen erklärt wird. Zusätzlich soll der Schnitt der ausgewählten Hypothesen bis auf einen Toleranzwert leer sein, um mit ungenauen Transformationen aufgrund von Rauschen umgehen zu können. Weiterhin soll stets eine gültige Lösung verfügbar sein sowie das Problem von lokalen Maxima gelöst werden können.
Dazu bietet sich ein *Jackknife* [Duda01] basierter Ansatz an (auch bekannt unter *leave one out*), welcher beispielsweise für das Generieren von Test-, Trainings- und Validierungsdatensätzen verwendet wird. Dazu werden zunächst alle Hypothesen in Λ nach ihrem Qualitätsmaß sortiert. Die Teilmenge H wird nun nach einem *greedy* Verfahren befüllt. Zunächst wird die Hypothese mit dem höchsten q hinzugefügt, da diese Hypothese am meisten von der Szene erklärt. Die nachfolgenden Hypothesen werden nun auf Schnitt mit den bereits vorliegenden Hypothesen in H getestet. Falls ein Schnitt vorliegt, wird diese Hypothese übersprungen. Falls keine vorliegt, wird diese Hypothese zu H hinzugefügt. Dieser Prozess wird solange wiederholt, bis alle Hypothesen in Λ geprüft wurden. Insgesamt kann man nun das Gesamtqualitätsmaß für H bestimmen. Damit liegt eine erste Auswahl an möglichen Hypothesen vor. Um lokale Maxima zu umgehen, wird dieser Prozess wiederholt, wobei die erste Hypothese (mit dem größten q) übersprungen wird. Sobald wieder alle Hypothesen getestet wurden, wird erneut das Gesamtqualitätsmaß bestimmt. Nun wird iterativ erhöht, wieviele Hypothesen übersprungen werden. Für dieses Vorgehen sind unterschiedliche Terminierungskriterien möglich, beispielsweise eine feste Dauer oder Iterationen. Ein Mindestqualitätsmaß kann hier nicht angegeben werden, da dieses nicht relativ ist und somit der notwendige Schwellwert nicht begründet parametriert werden kann. Abschließend wird die Menge H als Objektinstanzen verwendet, die das größte Qualitätsmaß aufweisen, um einen möglichst hohen Grad an erklärten Flächen zu erzielen.
Für den Test, ob sich zwei Objekte schneiden, existieren unterschiedliche Möglichkeiten: In der zugrundeliegenden Arbeit [Sand19] werden achsenorientierte Bounding Boxen verwendet, welche einen schnellen Kollisionstest ermöglichen, aber nur eine Approximation darstellen. Dies erlaubt zudem nur eine binäre Aussage. Alternativ kann das Maß der *Durchdringungstiefe* verwendet werden. Dies gibt eine Distanz an, wie weit ein Objekt bewegt werden muss, sodass es sich mit einem anderen Objekt nicht mehr schneidet. Für diese Distanz gibt es diverse Berechnungsvorschriften; eine schnelle Approximation kann über die Minkowskidifferenz bestimmt werden [Kim02]. Dies führt für konvexe Objekte stets zum korrekten Ergebnis, für nicht konvexe stellt es eine obere Schranke dar und kann somit als konservative Schätzung genutzt werden.

3.3 Experimentelle Auswertung

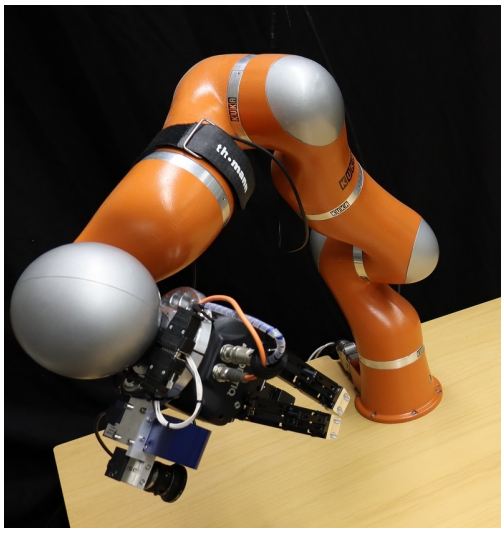
In diesem Kapitel wird der entwickelte Ansatz evaluiert. Dazu werden zunächst der verwendete Prototyp und Hardware-Aufbau beschrieben. Zur Evaluation wurde ein eigener Datensatz aufgebaut, der unterschiedliche Domänen umfasst. Anschließend werden die Experimente und Bewertungskriterien vorgestellt, ausgewertet und diskutiert.

3.3.1 Prototyp

Für die Evaluationen der ersten Kapitel wird ein Kuka LBR IV mit einer montierten *Eye-in-Hand*-Kamera am letzten Gelenk verwendet. Die Kamera wird extrinsisch kalibriert, indem ein Kalibriermuster an einer festen Position im Arbeitsraum platziert wird. Mehrere Roboterposen werden angefahren, und die Pose des *Tool Center Points* (TCP) sowie die entsprechende Punktwolke gespeichert. Da die Position des Kalibriermusters in Weltkoordinaten aus jeder Sicht identisch sein muss, kann über die Paare von Posen und Punktwolken optimiert werden. Die resultierende Transformation rechnet somit von TCP zum Kamerakoordinatensystem um. Als Kamera wird eine ENSENSO N-10 Stereo-Tiefenkamera mit einem konfigurierten Arbeitsabstand von 23.6 cm bis 90.6 cm verwendet. Die Strukturgröße für die B-Rep-Erzegung wurde auf 4mm gesetzt, basierend auf den Ergebnissen in [Sand19]. Ausgehend davon wurde das minimale Qualitätsmaß für Hypothesen auf 0.15 festgelegt.

3.3.2 Zusammensetzung der Objektdatenbank

Um die entwickelten Ansätze in dieser Arbeit zu evaluieren, ist es notwendig, einen festen Satz an Versuchsobjekten zu definieren. Da zur Rekonstruktion planare B-Reps verwendet werden, müssen alle Testobjekte stückweise planar rekonstruierbar sein. Darüber hinaus sollen mit den Testobjekten unterschiedlich komplexe Szenarien abgebildet werden können. Abschließend ist ein Bezug zu tatsächlichen, haushalts- oder industrieüblichen Objekten erforderlich. Alle Testobjekte gemeinsam sind in Abbildung 3.2 zu sehen, wo auch ein Größenvergleich möglich ist. Die Objekte einzeln mit Namen finden sich in Abbildung 3.3.
Die Objektdatenbank setzt sich insgesamt aus fünf unterschiedlichen Gruppen zusammen (Platonische Körper, *Brick World*, Haushalt, ähnliche Objekte, komplexe Objekte), die mit jeweils fünf Objekten vertreten sind. Die erste Gruppen bilden dabei die fünf platonischen Körper, da diese klassische planare Objekte darstellen und einen unterschiedlichen Grad an Komplexität mitführen (Abbildung 3.3, Reihe 1). Zum einen steigt die Anzahl der Flächen, zum anderen die Komplexität der Oberfläche. Dabei besitzt der Tetraeder mit vier Flächen die wenigsten, der Ikosaeder mit 20 die meisten.
Neben den platonischen Körpern werden einige Objekte aus der *Bricks World* betrachtet (Abbildung 3.3, Reihe 2). Hier wurden Bausteine ausgewählt, bei denen vier von fünf als Grundform Quader sind (und somit sechs Flächen haben) und sich ausschließlich in der Ausdehnung in den drei Raumdimensionen unterscheiden. Das fünfte Objekt ist ein dreieckiges Prisma. Die Herausforderung liegt bei diesen Objekten allem voran darin, da sie bezüglich ihrer Oberflächennormalen sehr ähnlich zueinander sind und ausschließlich über ihre Größe unterschieden werden können.



(a) Insgesamt verwendeter Hardwareaufbau, bestehend aus KUKA LBR IV mit Tiefenkamera und Greifer

(b) ENSENSO 10 Tiefenkamera, bestehend aus zwei Graustufenkameras sowie IR-Emitter und -Empfänger

Abbildung 3.1: Verwendete Hardware zur Evaluation

Die dritte Gruppe bilden dabei Gegenstände, die dem Haushalt entnommen werden (Abbildung 3.3, Reihe 3). Dabei handelt es sich primär um Verpackungsmaterial alltäglicher Gegenstände, wie beispielsweise Seife, Tee oder Süßigkeiten. Diese Objektgruppe ist im Vergleich zur vorangegangenen Gruppe heterogener. Allerdings ist Verpackungsmaterial oft ebenfalls ein Quader, womit Ähnlichkeit zu bestehenden Objekten existiert.
Die Herausforderung von untereinander ähnlichen Objekten wird mit der vierten Gruppe intensiviert. Diese besteht aus Objekten, die eine hohe Ähnlichkeit (sowohl in Größe als auch Oberflächennormalen) zu bereits bestehenden Objekten aufweisen (Abbildung 3.3, Reihe 4).
Abschließend wird mit der letzten Gruppe die Komplexität der Objektdatenbank erhöht, indem Gegenstände hinzugefügt werden, die eine hohe Flächenzahl haben, nicht konvex sind oder eine ungewöhnliche Form besitzen (Abbildung 3.3, Reihe 5). Manche dieser Objekte wurden bereits in der Vorarbeit [Sand19] als Benchmarkobjekte verwendet.
Von Objekten, die anfällig für Verformungen während eines Griffs durch den Roboter sind, wurde mit Hilfe eines 3D-Druckes ein Abbild geschaffen.

3.3.3 Aufbau

In diesem Abschnitt werden die unterschiedlichen Evaluationsszenarien vorgestellt, angelehnt an die in Kapitel 1 aufgestellten Kriterien: allem voran die Objektkomplexität, Verdeckungen und Szenenkomplexität. In jedem Schritt wird die komplette Objektdatenbank berücksichtigt.

Einzelne Objekte aus einer Sicht

Um zunächst die grundlegende Leistungsfähigkeit zu untersuchen, wird ein einfacher Anwendungsfall betrachtet, indem die zu untersuchenden Objekte sich alleine in der Szene befinden und der Sensor nur eine Sicht einnimmt. Dabei ist von Interesse, ob eine Sicht genügt,

Abbildung 3.2: Alle verwendeten Testobjekte

um alle Objekte korrekt wiederzuerkennen. Weiterhin können Verwechslungen zwischen den Objekten in einfachen Szenen näher untersucht werden. Als Evaluierungskriterium dient die Wiedererkennungsrate.

Rotierende, einzelne Objekte aus einer Sicht

In dieser Evaluation wird die Bedeutung der einzelnen Sicht auf die Szene näher untersucht. Da für das Aufstellen des Deskriptors drei linear unabhängige Flächen notwendig sind, ist die einzelne Sicht entscheidend. Dafür bleibt die Sicht des Sensors identisch, die ausgewählten Objekte werden um ein festes Inkrement entlang ihrer z-Achse rotiert. Die Umweltrekonstruktion wird vor jeder weiteren Aufnahme zurückgesetzt, sodass stets nur eine Sicht für die Objektwiedererkennung verwendet wird. Für jedes untersuchte Objekt wird dabei gemessen, in wie vielen Fällen das Objekt korrekt erkannt wurde.

Einzelne Objekte aus mehreren Sichten

Anschließend wird jedes Objekt, welches in der ersten Versuchsreihe nicht korrekt erkannt wurde, aus drei festen Sichten betrachtet, wobei diese so angeordnet sind, dass mehrere Flächen pro Objekt wahrnehmbar sind. Die Aufnahmen aus jeder Sicht werden in die bestehende Umweltrepräsentation fusioniert, wodurch mit jeder Sicht weiter Flächen rekonstruiert werden. Auf diese Weise wird die grundsätzliche Leistung des Ansatzes evaluiert. Auf Objekte, die bereits mit einer Sicht erkannt wurden, wird hier verzichtet, da für diese die Mächtigkeit des Ansatzes schon gezeigt wurde.

(a) Tetrahedron (b) Cube1 (c) Octahedron (d) Dodecahedron (e) Icosahedron

(f) Cube2 (g) Cuboid (h) Flat cuboid (i) Long cuboid (j) Triangle

(k) Soap (l) Tea1 (m) Sweets1 (n) Pen (o) Filter

(p) Peg (q) Tea2 (r) Tea3 (s) Sweets2 (t) Sweets3

(u) Anvil (v) Crane (w) Bench (x) Hole (y) Sponge

Abbildung 3.3: Alle Testobjekte individuell mit Namen, gruppiert in jeder Reihe: Platonische Körper, Bricks World, Haushalt, Ähnlichkeit, Komplexität

Mehrere Objekte aus mehreren Sichten

Als eine Komplexitätssteigerung der Szene befinden sich in dieser Evaluation mehrerer Objekte gleichzeitig in der Szene. Die gewählten Sichten bleiben identisch zu den bisherigen Experimenten, ebenso das Fusionieren der lokalen Repräsentationen. Anhand der Wiedererkennungsrate wird untersucht, inwiefern die räumliche Ausdehnung der Szene oder mögliche Verdeckungen einen Einfluss haben. Für die komplexesten Szenen berühren sich die Objekte in der Szene. Damit reduziert sich einerseits die räumliche Ausdehnung (im Fall eines Stapels an Objekten), allerdings erhöht sich andererseits der Verdeckungsgrad. Diese Einschränkung wird vor allem im Hinblick auf die Verfügbarkeit von linear unabhängigen Flächen untersucht. Insgesamt werden fünf Szenen mit jeweils fünf Objekten näher betrachtet. Jedes Objekt wird dabei verwendet, und die unterschiedlichen Objektklassen werden zufällig kombiniert.

Zeitmessung

Während dieser Experimente wird stets die Zeit bestimmt, die für den kompletten Wiedererkennungsschritt notwendig ist (das bedeutet vor allem, dass die Aufnahme und Rekonstruktion in ein B-Rep nicht Teil der Messung ist; für eine Evaluation dessen siehe [Sand19]). Anhand dieser Daten wird der quadratische Zusammenhang zwischen der Anzahl der Flächen und der Gesamtlaufzeit untersucht.

3.3.4 Ergebnisse

In diesem Abschnitt werden die Ergebnisse der unterschiedlichen Szenarien vorgestellt, die Mächtigkeit des entwickelten Ansatzes diskutiert und Einschränkungen erläutert.

Einzelne Objekte aus einer Sicht

Von den insgesamt 25 Objekten wurden 13 aus einer Sicht erkannt, zwölf Stück nicht. Bei acht der nicht erkannten Objekte wurden aus der einzelnen Position nur zwei Flächen rekonstruiert. Nach Definition des Deskriptors kann bei dieser Anzahl von Flächen keine Hypothese berechnet werden. Besonders zu erwähnen ist dabei der Tetraeder. Wenn dieses Objekt flach auf dem Tisch steht, müssen alle übrigen Flächen rekonstruiert werden. Dies ist mit der verwendeten Tiefenkamera aufgrund des Messprinzips aus einer Sicht nicht möglich. Von den verbleibenden vier Objekten wurden drei (*Sponge*, *Cuboid*, *Tea2*) aufgrund von Verwechslungen zwar erkannt, aber mit einem falschen Objekt verwechselt. Diese Objekte besitzen alle sehr ähnliche Deskriptoren sowohl untereinander als auch zu weiteren Objekten, da deren Flächen nahezu im 90° Winkel zueinander stehen. Das letzte nicht richtig erkannte Objekt (*peg*) wurde mit vier Flächen rekonstruiert, wobei allerdings jeweils zwei dieser Flächen in einer Ebene lagen. In diesem Fall ist es ebenfalls nicht möglich, eine Hypothese zu bestimmen. Ausgewählte Szenen und entsprechende Rekonstruktionen sind für ausgewählte Beispiele in Abbildung 3.4 dargestellt.

Rotierende, einzelne Objekte aus einer Sicht

Für diese Experimente wurden insgesamt drei Objekte mit unterschiedlicher Oberflächenkomplexität ausgewählt: *Cube1* (gering), *Sweets2* (mittel), und *Icosahedron* (hoch). Das Objekt *Cube1* wurde dabei aus fünf unterschiedlichen Sichten aufgenommen, wobei es in zwei davon korrekt wiedererkannt wurde, in drei davon nicht. In den positiven Fällen wurden jeweils drei Flächen rekonstruiert, in den restlichen nur zwei Flächen erkannt. Auch hier ist zu beachten, dass rein physikalisch maximal drei Flächen aus einer Sicht rekonstruiert werden können. Im Fall von *Sweets2* wurde dieser in drei Fällen korrekt erkannt (mit drei, drei und fünf Flächen). Wie beim *peg* erklären sich die fünf Flächen dadurch, dass diese bei der Rekonstruktion zerfallen sind. Das bedeutet, dass durchgehende Flächen in der Welt durch mehrere Flächen in der Rekonstruktion abgebildet werden. Mögliche Ursachen dafür sind Verdeckungen, wodurch die Teilflächen nicht zusammengefügt werden können. Physikalische Einschränkungen aufgrund des Messprinzips oder Rauschen haben ähnlichen Einfluss. In den nicht erkannten

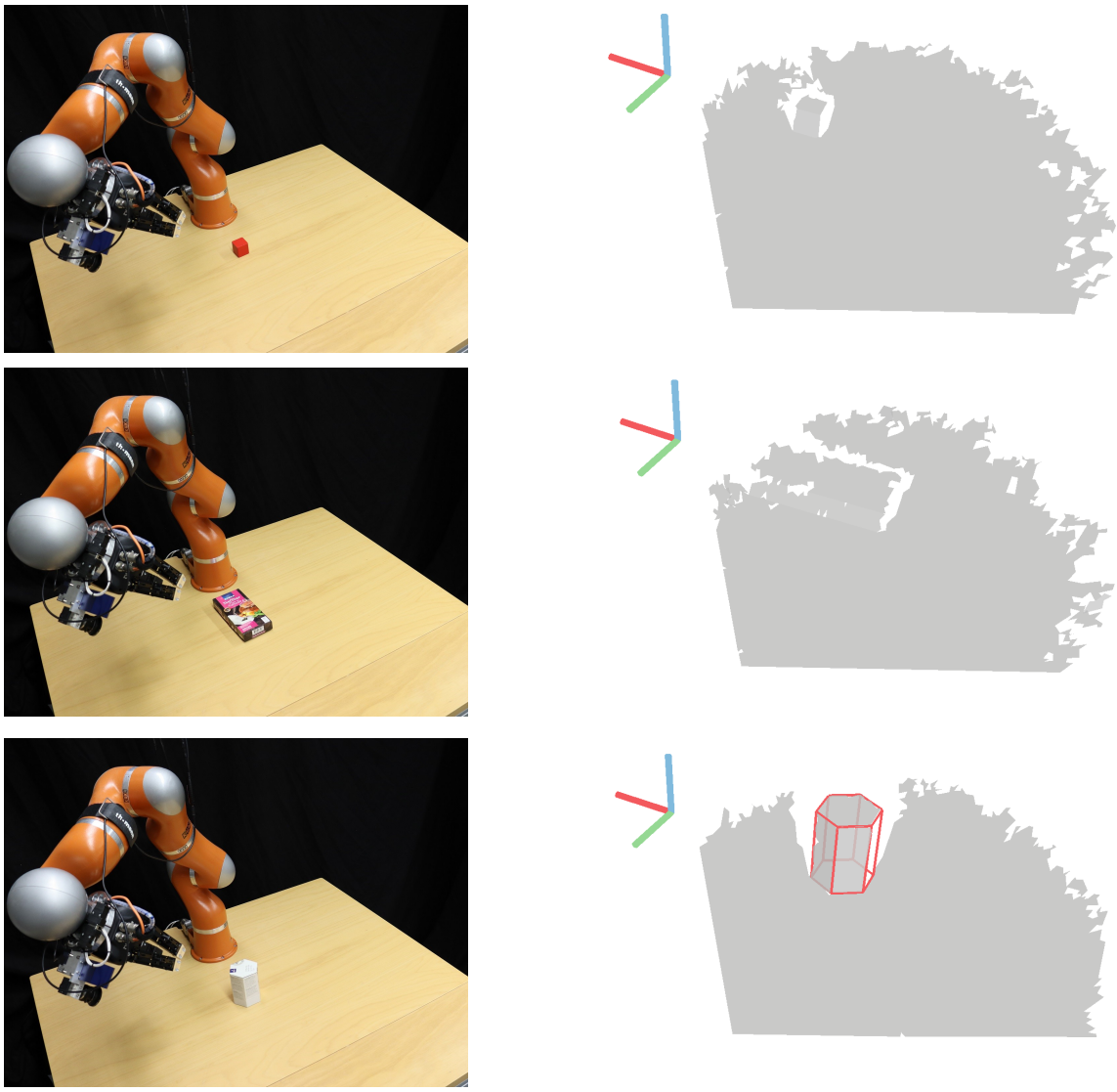


Abbildung 3.4: Wiedererkennung aus einer Sicht für die Objekte *Cube2*, *Filter* und *Pen*. Links die Szene, rechts die Rekonstruktion mit gegebenenfalls erfolgreicher Wiedererkennung (roter Rahmen), sowie eingezeichnetem Weltkoordinatensystem in der Roboterbasis

Fällen wurden jeweils zwei Flächen rekonstruiert. Der *Icosahedron* wurde aus allen fünf Sichten korrekt mit jeweils sechs rekonstruierten Flächen erkannt. Das Ergebnis für drei Sichten von *Sweets2* ist in Abbildung 3.5 zu sehen.

Einzelne Objekte aus mehreren Sichten

Von den zwölf nicht erkannten Objekten aus der ersten Versuchsreihe wurden nach maximal drei Sichten alle korrekt wiedererkannt. In neun Fällen haben zwei Sichten genügt, in den verbleibenden drei Fällen waren alle drei Sichten notwendig. Diese waren *Sponge*, *Tetrahedron* und *Triangle*. Im Fall des *Sponge* haben zusätzliche Sichten die Rekonstruktion verfeinert und die Flächen vervollständigt, wodurch das Bewertungsmaß der korrekten Hypothese gestiegen ist. Der Tetraeder bedarf wie oben erläutert einer vollständigen Rekonstruktion, ähnlich dazu das Dreieck. Für den Tetraeder sind die Aufnahmen in Abbildung 3.6 dargestellt.

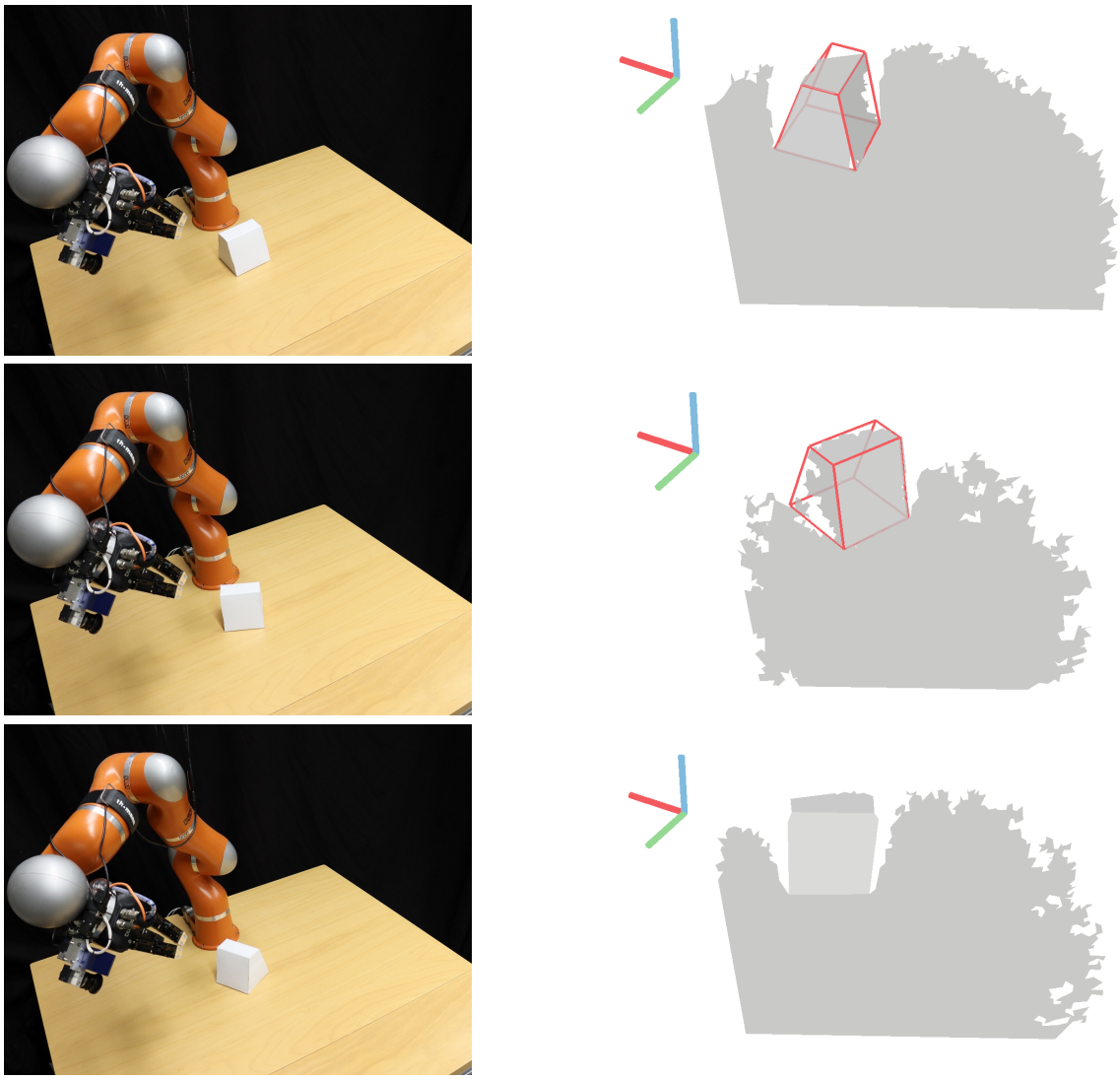


Abbildung 3.5: Rotation eines Objekts in der Szene bei gleichbleibender Roboterpose. Links die Szene, rechts die Rekonstruktion mit gegebenenfalls erfolgreicher Wiedererkennung (roter Rahmen)

Mehrere Objekte aus mehreren Sichten

Von den insgesamt 25 platzierten Objekten wurden 17 korrekt erkannt. Alle anderen Objekte wurden nach drei Sichten nicht wiedererkannt. Während der Rekonstruktion kam es mehrmals zu falsch-positiven Erkennungen, die aber alle in der Rekonstruktion nach drei Sichten nicht mehr auftraten. Für alle Objekte, die nicht erkannt wurden, haben in der finalen Rekonstruktion nicht genügend Flächen vorgelegen, was sowohl auf ungeeignete Sichten als auch Verdeckungen durch andere Objekte zurückzuführen ist. Zwei Szenen mit den dazugehörigen Aufnahmen sind in Abbildung 3.7 und 3.8 zu sehen.
Als Besonderheit ist festzuhalten, dass für Objekte mit einer einzigartigen Anordnung der Oberflächennormalen bereits ein sehr geringer Rekonstruktionsgrad genügt, um das Objekt korrekt zu erkennen. Beispielsweise für Icosahedron liegen einzigartige Winkel zwischen den Flächen vor, wodurch das Objekt im Allgemeinen korrekt erkannt wird, wenn drei Flächen davon rekonstruiert werden (unabhängig vom Rekonstruktionsgrad). Insgesamt lässt sich

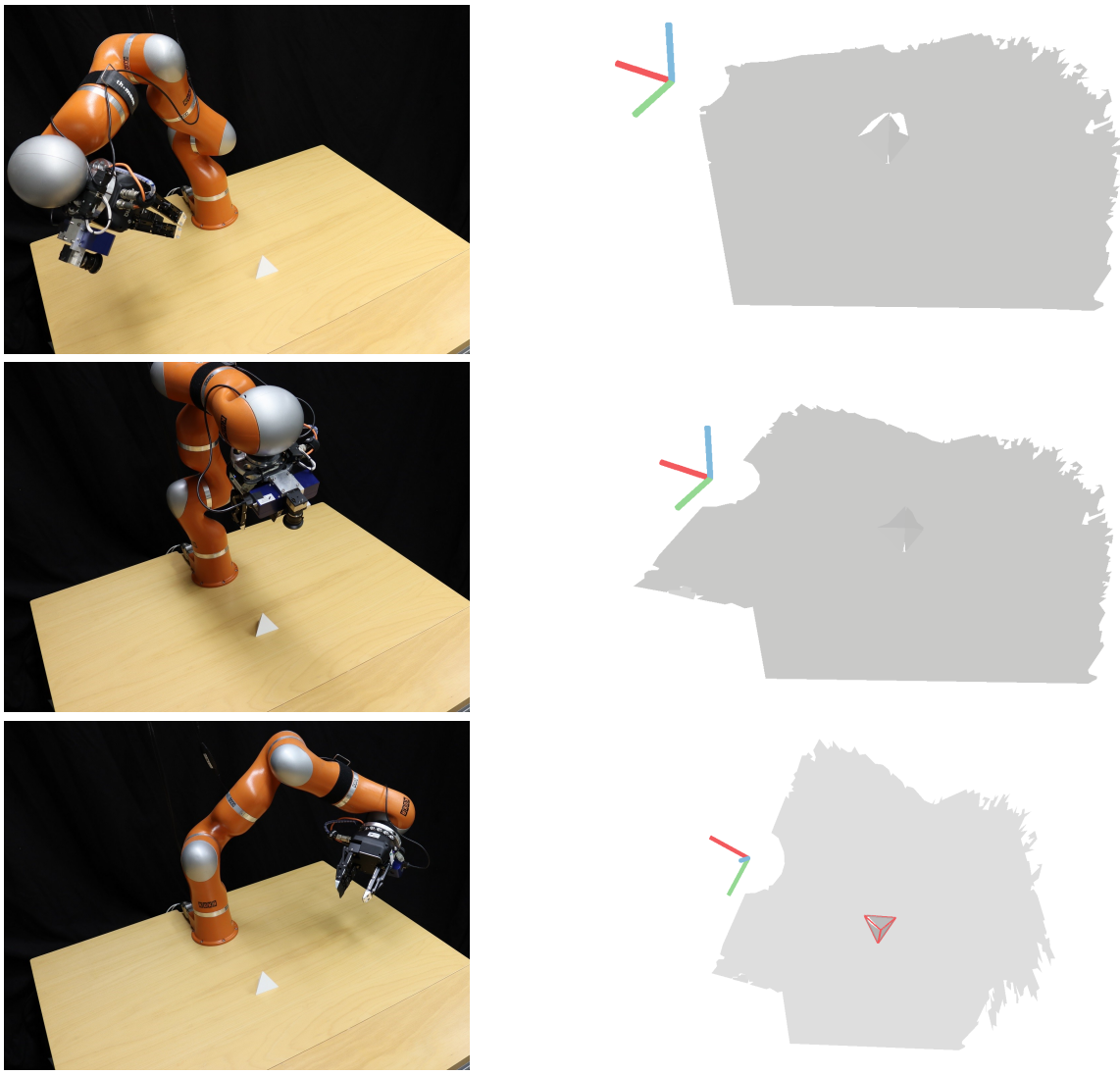


Abbildung 3.6: Rekonstruktion und Wiedererkennung eines einzelnen Objekts aus mehreren festen Sichten: links die Szene, rechts die Rekonstruktion mit gegebenenfalls erfolgreicher Wiedererkennung (roter Rahmen)

schlussfolgern, dass der Wiedererkennungsansatz auch bei komplexeren Szenen den Großteil der Objekte korrekt wiedererkennt, falsch-positive Erkennungen durch mehrere Sichten behebt und den Aufbau einer globalen Umweltrepräsentation aus lokalen Sichten unterstützt.

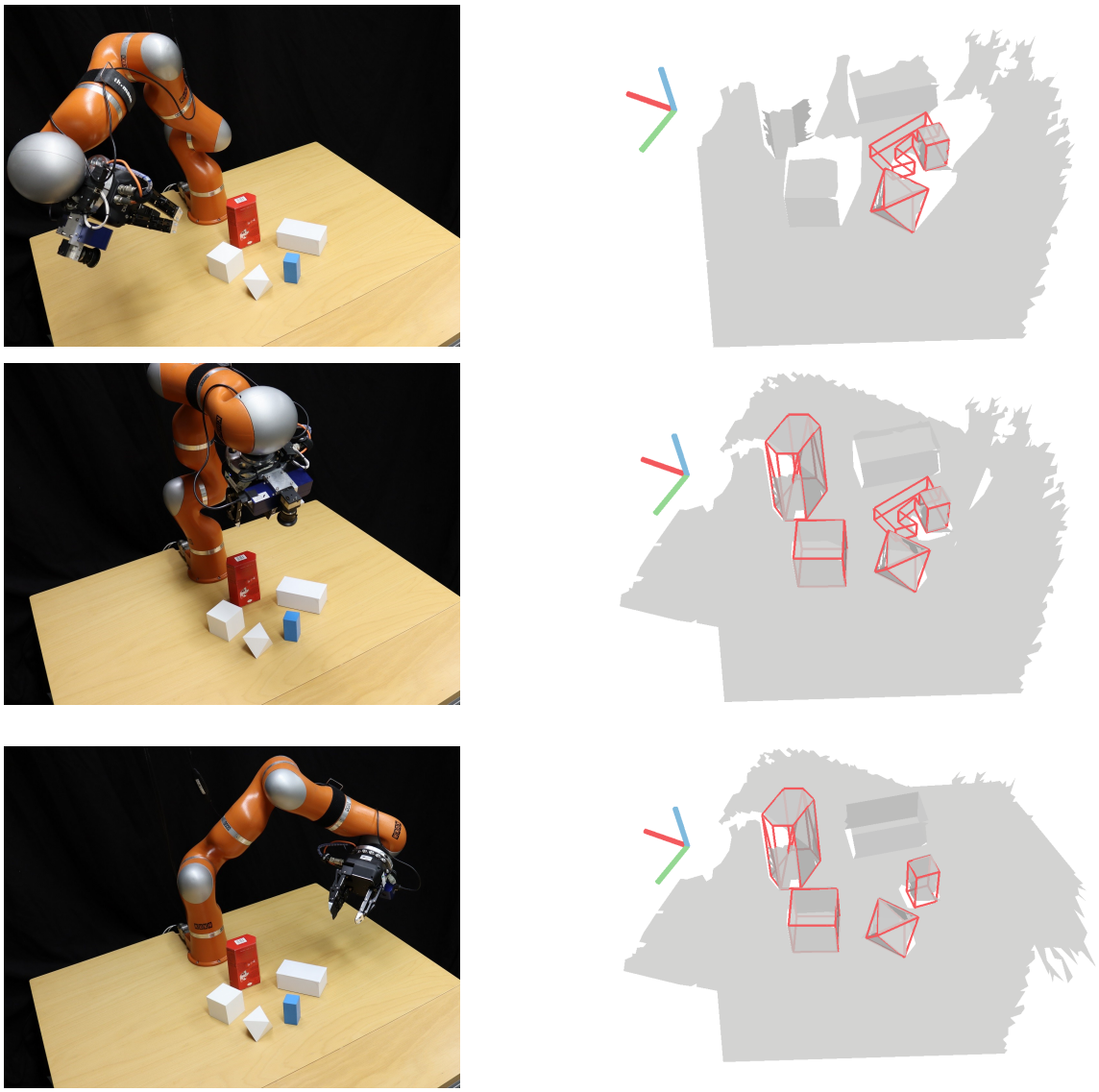


Abbildung 3.7: Rekonstruktion einer Szene aus drei festen Sichten. Nach der letzten Sicht wird ein Objekt aufgrund fehlender Flächen nicht erkannt, da die festen Roboterposen ungeeignet für diese Orientierung sind.

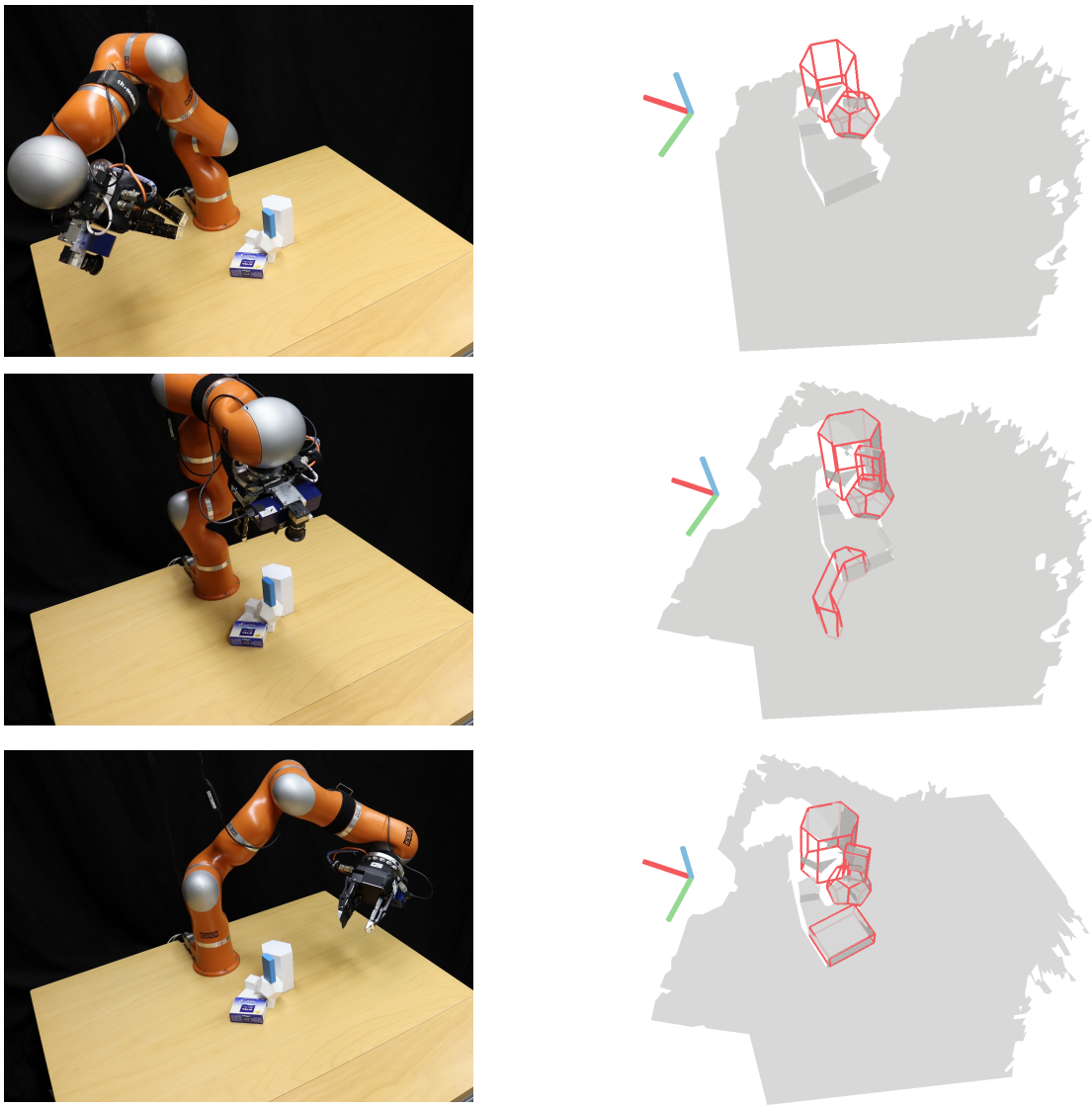


Abbildung 3.8: Rekonstruktion einer weiteren Szene aus drei festen Sichten. Nach der letzten Sicht wird ein Objekt aufgrund fehlender Flächen nicht erkannt, da die aufgrund von Verdeckungen nicht sichtbar sind.

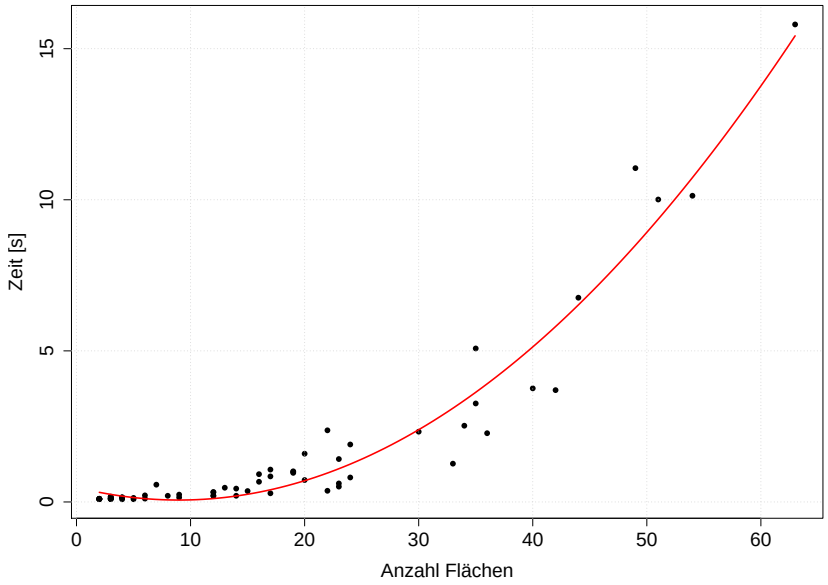


Abbildung 3.9: Zeitmessung der Objektwiedererkennung: Dauer in Sekunden gegen die Anzahl an Flächen in der Szene (schwarze Punkte) sowie eine quadratische Anpassung (rot)

Zeitmessung

Über die Experimente in dieser Arbeit hinweg wurde die Zeit für den kompletten Wiedererkennungsschritt gemessen und auch die Anzahl an Flächen in der Szene. Das Ergebnis ist in Abbildung 3.9 dargestellt und mit der Dauer zur Hypothesenerezugung und Auswahl gegen die Anzahl der Flächen in der Szene aufgetragen. Die Anzahl an Flächen in der Objektdatenbank war unter Verwendung aller Einträge während der Experimente mit insgesamt 197 Flächen konstant. In Rot ist eine quadratische Anpassung an die Daten eingezeichnet. Gemäß dem Bestimmtheitsmaß $R^2 = 0,9475$ beziehungsweise dem adjustierten Bestimmtheitsmaß $\bar{R}^2 = 0,9458$ ist die Qualität der Anpassung hoch. Die Hypothesenerzeugung setzt sich dabei primär aus zwei Teilschritten zusammen: Dem berechnen aller Deskriptoren und dem Vergleich der Deskriptoren zwischen Szene und Objektdatenbank. Der asymptotische Aufwand für das Erzeugen der Deskriptoren ist dabei kubisch in der Anzahl der Flächen, allerdings müssen diese nur für die Szene bestimmt werden. Die Deskriptoren für die Objekte verändern sich nicht und können vorab bestimmt werden. Der Vergleich der Deskriptoren selbst ist quadratisch in der Anzahl ebendieser. Solange die Anzahl an Flächen in der Welt kleiner ist als die in der Objektdatenbank überwiegt bei der Gesamtlaufzeit somit der quadratische Anteil des Deskriptorvergleichs.
Dieses Ergebnis deckt sich mit der Erwartung, dass die Laufzeit sich mit höherer Flächenzahl steigert, da sich die Anzahl möglicher Hypothesen in der Szene erhöht, wenn weitere Flächen hinzugefügt werden. Durch Verwendung höher getakteter Hardware sowie einem größeren Parallelisierungsgrad lässt sich die Laufzeit reduzieren. Da die Hypothesenerzeugung für jedes Objekt individuell berechnet werden muss, ist dies möglich. Auch für die Hypothesenauswahl ist eine Parallelisierung möglich, indem der verwendete *Jackknife*-Ansatz mit unterschiedlicher Auslassung der ersten Objekte gleichzeitig bestimmt werden kann.

3.4 Fazit

In diesem Kapitel wurde ein vollständiger Ansatz zur Wiedererkennung von planar rekonstruierbaren Objekten vorgestellt, welcher ausschließlich die Oberflächeninformation dieser Objekte verwendet. Alle Objekte liegen dabei in einer Objektdatenbank vor. Dazu werden zunächst basierend auf den Oberflächennormalen dieser Objekte Hypothesen berechnet. Diese werden mit einem Qualitätsmaß basierend auf der Überschneidung der Rekonstruktion und dem Eintrag in der Objektdatenbank bewertet. Somit fällt dieser Ansatz in die Gruppe der modellbasierten Methoden. Dies ermöglicht explizit das einfache Hinzufügen weiterer, bisher unbekannter Objekte. Aus den berechneten, möglichen Hypothesen wird mit einem Jackknife-Ansatz eine Teilmenge der am besten passenden mit dem Ziel ausgewählt, das globale Qualitätsmaß zu maximieren.
Im Rahmen der Evaluation wurde gezeigt, dass mit diesem Ansatz untexturierte, geometrisch primitive Objekte wiedererkannt werden können. Vor allem durch die geringe Anzahl an Verwechslungen trotz ähnlicher Objekte ist ein hoher Wiedererkennungsgrad gegeben. Aufgrund des hohen Abstraktionsniveaus ist es möglich, eine Vielzahl an Hypothesen zu bestimmen und zu vergleichen. Allerdings steigt die Laufzeit quadratisch mit der Anzahl an Flächen. Ein Nachteil dieses hohen Abstraktionsniveaus ist der Bedarf nach drei linear unabhängigen Flächen. Gerade bei einer geringen Anzahl oder ungeeigneten Sichten auf die Szene ist die Anforderung nicht immer zu erfüllen. Ebenso erschweren Verdeckungen die korrekte Wiedererkennung. Eine Möglichkeit, dieses Problem zu lösen, ist das Hinzufügen weiterer Informationen, allem voran Textur bzw. Farbinformation. Alternativ können gezielte Sichten auf die Szene die Wiedererkennungsrate verbessern, indem nach neuen Flächen gesucht wird. Somit lässt sich die erste Frage

F1 Inwieweit können geometrisch primitive, untexturierte Objekte wiedererkannt werden?

zunächst dahingehend beantworten, dass dies aufgrund der gewählten Repräsentationsform vor allem durch den Brute-Force Ansatz ermöglicht wird, aber durch Verdeckungen und eine quadratische Komplexität limitiert ist. Bei einer expliziteren Repräsentation wie beispielsweise Punktwolken gäbe es aufgrund der hohen Anzahl individueller Deskriptoren eine zu hohe Anzahl an Möglichkeiten, die nicht mehr verarbeitet werden kann. Durch die Reduktion auf Flächen sinkt die absolute Anzahl an möglichen Deskriptoren stark ab. Nichtsdestoweniger leidet auch der hier vorgestellte Ansatz an der hohen kombinatorischen Komplexität. Um dem entgegenzuwirken, kann in weiteren Arbeiten die Anzahl zu untersuchender Flächen stärker reduziert werden. Zum einen, indem nicht alle Objekte aus der Objektdatenbank geprüft werden, beispielsweise durch eine Hierarchisierung. So lassen sich mehrere Objekte mit ähnlicher Geometrie (und somit ähnlichem bis identischem Deskriptor zum Beispiel alle Quader) in eine Gruppe zusammenfassen. In einem ersten Schritt kann zunächst geprüft werden, welche Gruppe an Objekten vorliegt, bevor dann die tatsächliche Objektinstanz geprüft wird. Die Anzahl zu untersuchender Flächen in der Umweltrepräsentation kann verringert werden, indem unmögliche bis unwahrscheinliche Flächenkombinationen, beispielsweise anhand des Abstands oder von Nachbarschaftsbeziehungen, erkannt und vorgefiltert werden.

Kapitel 4

Active Vision für Objektwiedererkennung

Inhalt

Als Problem der Objektwiedererkennung bleibt die Frage offen, wie eine vollständige Wiedererkennung aller Objekte möglich ist, wenn sie zu verteilt für eine Aufnahme sind. Weiterhin kann auch nicht jedes Objekt aus einer Sicht identifiziert werden. Dieses Problem wird durch Verdeckungen noch verstärkt.
In diesem Kapitel wird ein *Active Vision*-Ansatz vorgestellt, wie neue Sichten für den Roboter bestimmt werden können. Dabei wird sowohl die *Exploration* der Szene betrachtet als auch die *Validierung* bestehender Hypothesen. Dazu werden Untersuchungsregionen identifiziert und daraus mögliche Sichten bestimmt (Abschnitt 4.2). Grundzüge dieser Methode und der Ergebnisse wurden in [Rohner20a] bereits veröffentlicht.

4.1 Stand der Forschung

Der Ansatz *Active Vision* in Form von beweglichen Kameras zur Szenenrekonstruktion existiert bereits seit längerem in unterschiedlichen Anwendungen, beispielsweise in den Bereichen *Shape from Contour* oder *Structure from Motion* [Aloimonos88]. In der Robotik und Objektrekonstruktion sind Methoden zum Bestimmen zusätzlicher, informationsmaximierender Sichten auf die Szene unter *Next Best View* [Connolly85] bekannt. Dabei ist zu unterscheiden, welches Ziel verfolgt wird. Auf der einen Seite kann die vollständige Erfassung der Umwelt oder eines Werkstückes von Interesse sein. Andererseits ist die vollständige Wiedererkennung aller Objekte ein notwendiges Ziel. Bei einer vollständigen Rekonstruktion werden zwar im Allgemeinen auch alle Objekte richtig erkannt, dafür ist die Anzahl an aufgenommenen Sichten allerdings unnötig groß für eine Objektwiedererkennung. Ein Überblick über Arbeiten zum Thema *Active Vision* und *Next-Best-View* finden sich in [Chen11] und [Grotz21]. Damit verwandt ist *Visual Attention* [Bundesen90], was Auswahl und Fokussierung aktuell relevanter Information beschreibt. Siehe [Potapova17] für eine Übersicht im Hinblick auf Robotik.
Weitere Ansätze im Bereich der Robotik können anhand mehrerer Kriterien unterschieden werden. Das ist zum einen die Anzahl an Kameras und zum anderen die mögliche Bewegung in der Szene. Wie in den einleitenden Kapiteln diskutiert, können Kameras sowohl beweglich (beispielsweise auf mobilen Plattformen oder als *Eye-in-Hand* Kamera) oder statisch fixiert sein (als *Overhead* Kamera). Für Kamerasysteme, die ausschließlich statische Kameras verwenden, ist *Active Vision* keine Option, da die Kamera nicht in der Szene bewegt werden kann; in Kombination mit beweglichen Kameras ist sie aber anwendbar, ebenso in Anwendungen nur mit beweglichen Kameras. Ein letztes Kriterium ist die Veränderung der Szene während des Aufnahmeprozesses. Dabei kann die Szene unverändert belassen oder durch einen gegebenenfalls vorhandenen Manipulator modifiziert werden, um Verdeckungen zu reduzieren oder Objekte gezielt zu betrachten.
Für die Kombination aus statischen und beweglichen Kameras existieren mehrere Ansätze. Eine Möglichkeit sind *Master-Slave* Ansätze, in denen eine statische Kamera Regionen in der Welt identifiziert, die anschließend von einer beweglichen Kamera näher betrachtet werden sollen [Al Haj11, Xiong12, Ilie14]. Mehrere bewegliche Kameras, die zusammenarbeiten, können in einem *Active Camera Network* [Al Haj11, Aghajan09, Kyrkou20] organisiert werden.
Mit dem Ziel der vollständigen Rekonstruktion ist vor allem die Reduzierung der Anzahl an notwendigen Sichten von Interesse. In dieser Aufgabe wird meist nur eine einzelne bewegliche Kamera verwendet. Dazu können weitere Sichten ausgehend von der Oberfläche der bisherigen Rekonstruktion sowie die dazugehörige Unsicherheit erfolgen [Dunn09]. Ein anderer Ansatz diskretisiert die möglichen Sichten auf einem Gitter und berechnet einen effizienten Pfad, um alle anzufahren [Elzaiady17]. Berücksichtigt werden müssen dabei Ungenauigkeiten sowohl des Sensors als auch eine mögliche unpräzise Positionierung [Vasquez-Gomez17, Yu04]. Die Planung von nächsten Sichten kann dabei direkt im Konfigurationsraum erfolgen [Suppa04]. Ansätze für das *Next Best View Problem* finden weiterhin Anwendung in der mobilen (und humanoiden) Robotik [Isler16, Monica16, Grotz21] oder in der Vervollständigung von Früchten [Menon23].
Für das Ziel, die Objektwiedererkennung zu verbessern, sind [Wilkes92] und [Tistarelli94]

erste Beiträge. Für mobile Plattformen [Holz13] sowie eine bewegliche Kamera [Gratal10] ist diese Aufgabe untersucht. Weitere Ansätze verzichten auf die Verfügbarkeit von Objektmodellen [Hoseini22], fokussieren sich gezielt auf ein Objekt [McGreavy16] oder versuchen bisher unentdeckte Objekte in der Umgebung wahrzunehmen [Langer17].
Wie eingangs beschrieben, kann auch die Szene modifiziert werden, um zusätzliche Information erfassen zu können, was als *Interactive Perception* bekannt ist, worüber [Bohg17] einen Überblick gibt und unterschiedliche Anwendungsgebiete wie Objektwiedererkennung, Objektsegmentierung oder Greifplanung identifiziert. Um die Objektwiedererkennung zu erleichtern, kann der Roboter die Objekte in der Szene vereinzeln [Sinapov13] oder in eine spezifische Konfiguration überführen [Cusumano-Towner11]. Alternativ kann es genügen, Objekte, die Verdeckungen verursachen, in der Szene zu bewegen [Marques25].
Eine weitere Anwendung von Active Vision ist die Greifplanung durch Fokussieren auf ein spezifisches Objekt [Bohg12], um die Unsicherheit der Pose für den Griff zu reduzieren [Welke13] oder um den Zustand des Griffs zu überwachen [Arruda16].
Für die Objektwiederekrennung existieren zwei unterschiedliche Herausforderungen: Eine davon ist die Notwendigkeit, alle vorhanden Objekte korrekt wiederzuerkennen. Dabei ist es aber möglich, dass Objekte falsch wiedererkannt werden. Somit müssen mit weiteren Sichten sowohl neue Hypothesen erzeugt als auch bestehende Hypothesen sichergestellt werden. Zum anderen muss die Erzeugung neuer Sichten online erfolgen können und ermöglicht keine aufwändige Optimierung. Weiterhin ändert sich die Umweltrepräsentation mit jeder neuen Sicht, weshalb nach jedem Schritt diese neu evaluiert werden muss.

4.2 Beschreibung des Verfahrens

Im Weiteren wird der Ansatz erläutert, wie Forschungsfrage F2 im Hinblick auf B-Reps als Umweltrepräsentation beantwortet werden kann.

F2 Inwieweit unterstützen zusätzliche lokale Sichten die Objektwiedererkennung? Inwieweit können neue Sichten szenenspezifisch bestimmt werden?

Die Notwendigkeit mehrerer Sichten wurde bereits in der Evaluation des vorangegangenen Kapitels diskutiert. Darauf aufbauend werden zunächst weitere Probleme erläutert sowie weitere Grundlagen für das Kapitel dargelegt. Davon ausgehend wird der allgemeine Ansatz erläutert, wie szenenspezifisch neue Sichten erzeugt werden können. Dazu müssen zunächst Untersuchungsregionen (oder auch *Region of Interest*) identifiziert werden. Für die einzelnen Untersuchungsregionen werden anschließend mögliche Sichten generiert. Diese werden gefiltert, um ungeeignete Sichten zu entfernen. Abschließend wird mit Hilfe einer Heuristik eine neue Sicht ausgewählt. In der Evaluation wird anschließend sowohl der Nutzen dieses Ansatzes untersucht als auch die Frage beantwortet, inwieweit zusätzliche lokale Sichten die Objektwiedererkennung unterstützen.

4.2.1 Grundlagen

Die Notwendigkeit für Active Vision in Form von zu wenig Information über ein Objekt hat mehrere Ursachen, vor allem dann, wenn nur ein einzelner Sensor verwendet wird. Dabei haben Verdeckungen den größten Einfluss. Das bedeutetet zum einen, dass manche Objekte aus einer Sicht nicht eindeutig korrekt identifizierbar sind (aufgrund von Ähnlichkeit zwischen Objekten), und zum anderen, dass zu wenig Information von einem Objekt erkannt wird, sodass kein Deskriptor bestimmt werden kann. In dieser Arbeit sind das die Winkel zwischen drei linear unabhängigen Flächen. Ein weiteres Problem sind ähnliche Objekte in der Objektdatenbank. Zwar kann ein Deskriptor bestimmt werden, allerdings existiert dieser beziehungsweise ist deckungsgleich für mehrere Objekte (beispielsweise alle quaderförmigen Objekte in der Datenbank). Das richtige Objekt kann somit im allgemeinen nur dann richtig identifiziert werden, wenn entweder einzelne Flächen zu großen Teilen rekonstruiert oder andere signifikante Merkmale aufgenommen wurden. In dieser Arbeit müssen drei linear unabhängige Flächen pro Objekt rekonstruiert werden. Abschließend können einzelne Szenen zu groß sein, sodass diese nicht von einem einzelnen Sensor vollständig erfasst werden können.
Für das weitere Vorgehen muss die B-Rep Definition um eine Eigenschaft ergänzt werden. Bei der Rekonstruktion von Kanten wird für diese zusätzlich hinterlegt, wie sie entstanden sind. Dafür gibt es drei unterschiedliche Möglichkeiten. Zum einen kann ein *Normalensprung* vorliegen. Das bedeutet, dass an einer Kante zwei Flächen existieren, und somit sich die Normale zwischen den beiden Flächen ändern muss (sollte sich die Normale nicht ändern, wäre es eine Fläche). Die zweite Möglichkeit ist ein *Tiefensprung*, der entsteht, wenn eine Fläche aus Sicht der Kamera durchgängig wäre, sie es allerdings aufgrund des Tiefenunterschiedes in 3D nicht ist. Abschließend ist ein *Beobachtungssprung* möglich. Dieser tritt dann auf, wenn die zugrundeliegende Punktwolke an einer Stelle endet. Diese drei Typen sind in Abbildung 4.1 grafisch dargestellt.
Im Weiteren wird auf die Formalisierung aus dem vorangegangenen Kapitel (Abschnitt 3.2.2) zurückgegriffen, die hier knapp zusammengefasst wird: Ein einzelnes B-Rep A ist ein 4-Tupel F, E, V, B. Mit dem Index X_A wird im Weiteren der Bestandteil X des Tupels identifiziert, beispielsweise F_A für alle Flächen F vom B-Rep A. Aus mindestens einer Sicht wird ein B-Rep W generiert, welches die Welt repräsentiert. Dazu werden Objekte wiedererkannt und die entsprechenden Hypothesen in der Menge $H = \{h_1, ..., h_n\}$ gesammelt. Jede Hypothese besteht dabei unter anderem aus einem B-Rep-Modell des entsprechenden Objekts. Sowohl die Umweltrepräsentation als auch die Hypothesen sind die Eingabe für den Active Vision Ansatz.

4.2.2 Ansatz

Wie in der Einleitung beschrieben, existieren unterschiedliche Probleme, die den Einsatz von Active Vision verlangen. Verdeckungen und große Szenen zeigen dabei vor allem auf, dass eine gegebene Szene weiter *exploriert* werden muss, damit diese vollständig erfasst werden kann. Ähnliche Objekte und der Wiedererkennungsansatz im Allgemeinen hingegen verlangen, dass bestehende Hypothesen *validiert* werden müssen, da nicht anhand jeder berechneten Hypothese das richtige Objekt wiedererkennt wird. Da der Wiedererkennungsansatz auf den einzelnen Flächen in der Szene beruht, ist es für den Active Vision Ansatz notwendig, mög-

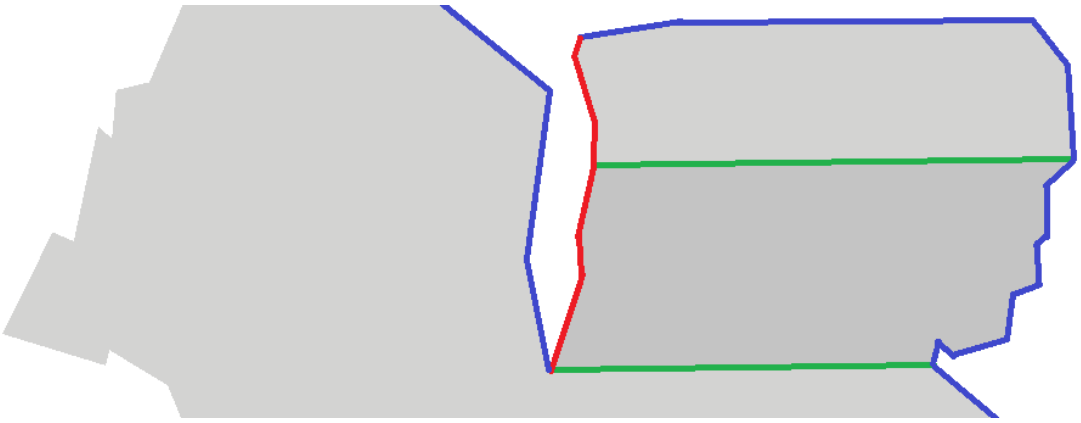


Abbildung 4.1: Unterschiedliche Kantentypen eines B-Reps: Tiefensprung (rot), Normalensprung (grün) und Beobachtungssprung (blau), aus [Rohner20a]

lichst viel Information über neue Flächen zu sammeln. Wenn ähnliche Objekte vorliegen, ist die Vollständigkeit einzelner Flächen hilfreich, aber nicht zwingend erforderlich. Daher konzentriert sich der Ansatz darauf, neue Flächen zu entdecken. Diese führen entweder dazu, dass neue Objektinstanzen generiert werden können (Exploration) oder dass bestehende Hypothesen bestätigt oder verworfen werden können (Validierung). Darüber hinaus muss festgelegt werden, wie eine *Initialiserung* der Welt stattfindet, sodass neue Sichten berechnet werden können. Dabei wird von einer zunächst leeren Welt ausgegangen. Da noch keinerlei Information vorliegt, ist eine Exploration notwendig. Im einfachsten Fall ist das eine beliebige oder vorher festgelegte Pose. Alternativ kann eine Explorationsfahrt stattfinden, bis eine Fläche in der Welt rekonstruiert wird. Sobald mindestens eine Fläche erkannt wurde, wird der bereits beschriebene Wiedererkennungsansatz angewendet. Diese berechneten Hypothesen sind zusammen mit der Umweltrepräsentation die Eingabe für die nächsten Schritte. Dazu werden für den Schritt der Exploration alle Flächen in der Szene gesammelt, die keine Korrespondenz mit einer der Flächen der Hypothesen haben. Diese Flächen sind deshalb interessant, weil jede Fläche in der Szene durch ein Objekt in der Welt entstanden sein muss (die Arbeitsoberfläche sowie der Roboter können anhand der bekannten Pose vorab entfernt werden). Daher muss es für ein Objekt, welches keine Hypothese hat, noch weitere Flächen geben. Für die Validierung werden die Objektflächen der Hypothesen betrachtet, die keine Korrespondenz mit der Welt haben. Da diese Stelle der Welt noch nicht rekonstruiert wurde, kann nicht sichergestellt werden, dass die Hypothese korrekt ist. Die so gesammelten Flächen für die Exploration und Validierung sind im Weiteren als *Untersuchungsregionen* (*Regions of Interest*) zusammengefasst, bei Bedarf aufgeteilt in *Explorationsflächen* und *Validierungsflächen*. Für die bestimmten Flächen werden schließlich mögliche Sichten generiert und diese mit dem Ziel bewertet, dass diese Sicht die neue Information maximiert. Einzelne Sichten werden aufgrund von Kollisionen oder einem geringen Nutzen vorab entfernt. Für die ausgewählte Sicht wird abschließend die inverse Kinematik gelöst. Falls das nicht erfolgreich ist, wird eine andere Sicht ausgewählt. Bei Erfolg nimmt der Roboter die neue Pose ein, eine Punktwolke wird aufgenommen und als B-Rep rekonstruiert. Das B-Rep wird in die Umweltrepräsentation eingefügt und die neuen Hypothesen werden bestimmt. Der Prozess wird nun wiederholt. Sobald keine möglichen Sichten mehr existieren, terminiert der Prozess.

4.2.3 Identifikation von Untersuchungsregionen

Die Identifikation der Untersuchungsregionen richtet sich nach unterschiedlichen Kriterien. Das Ziel von Active Vision ist in dieser Arbeit keine vollständige Rekonstruktion der Szene. Vielmehr sind korrekt wiedererkannte Objekte von Interesse, da diese genauer beschreiben, was für Gegenstände in der Szene vorliegen. Damit Objekte wiedererkannt werden können, müssen drei linear unabhängige Flächen rekonstruiert worden sein. Gleichzeitig dienen mehrere gesehene Flächen der Validierung von bestehenden Hypothesen.
Für die Exploration ist es das Ziel, neue Flächen zu identifizieren. Das bedeutet, dass Flächen von Interesse sind, die bisher nicht von einer Hypothese erklärt werden. Stattdessen muss in der Nähe bereits bestehender, unerklärter Flächen exploriert werden. Dem liegt die bisherige Annahme zugrunde, dass in der Szene keine Objekte existieren, die nicht wiedererkannt werden. Somit muss jede rekonstruierte Fläche durch ein Objekt erklärbar sein. Da weiterhin alle Objekte zumindest zusammenhängend sind, muss jede Fläche zumindest einen Nachbarn haben. Diese benachbarten Flächen zu entdecken und somit neue Hypothesen zu erzeugen, ist das Ziel der Exploration. Um zu bestimmen, ob in der Umweltrepräsentation W eine Fläche zu einer anderen Objektfläche korrespondiert, wird die Funktion `explainedby` genutzt [Sand19].

$$\texttt{explainedby}(f,g) \iff$$
$$\alpha(f,g) < \delta_{\text{corr}} \wedge A(f,g) > A_{\text{corr}} \wedge d(f,g) < d_{\text{corr}}$$

Dabei beschreibt $\alpha(f,g)$ den Winkel zwischen zwei Flächen f und g, mit $A(f,g,)$ wird die Überscheidungsfläche bestimmt und mit $d(f,g)$ der Abstand zwischen den beiden Flächen. Falls alle diese Maße die anwendungsspezifischen Grenzwerte für eine Korrespondenz δ_{corr}, A_{corr} und d_{corr} erfüllen, erklären sich die Flächen f und g gegenseitig. Somit lassen sich die Untersuchungsregionen für die Flächen bestimmen. Dazu wird zunächst die Menge aller Hypothesenflächen bestimmt

$$F_H = \{F | ((F,E,V,B),T,q) \in H\}$$

indem für jede Hypothese (bestehend aus einem B-Rep mit den üblichen geometrischen Elementen (F,E,V,B), einer Starrkörpertransformation T und dem Qualitätsmaß q) die Objektflächen des dazugehörigen B-Rep Modells gesammelt werden. Dazu werden alle Flächen aggregiert, die Teil des B-Reps einer Hypothese sind.
Für jede Fläche der Umweltrepräsentation F_W kann geprüft werden, ob diese durch eine Fläche der Hypothesen erklärt wird. Alle Flächen, für die das nicht zutrifft, sind Regionen von Interesse für die Exploration:

$$R_W = \{f \in F_W | \nexists g \in F_H : \texttt{explainedby}(f,g)\}$$

Dieser Prozess kann analog für die Menge der Validierungsflächen durchgeführt werden. Diese sind dann von Interesse, wenn eine Hypothese existiert, einzelne Flächen der Hypothese aber nicht durch die Szene erklärt werden. Die Hypothese spiegelt nur die aktuelle bestmögliche Erklärung für die Szene wieder. Mit mehreren rekonstruierten Flächen können bestehende

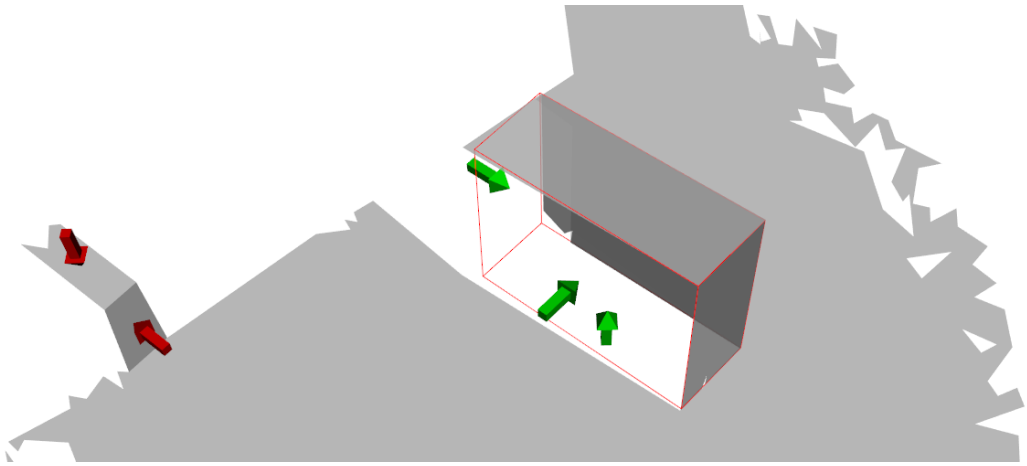


Abbildung 4.2: Unterschiedliche Flächentypen für Active Vision: Explorationsflächen (Flächen in der Umweltrepräsentation ohne korrespondierende Hypothese, rote Pfeile) und Validierungsflächen (Hypothesenflächen ohne korrespondierende Flächen in der Umweltrepräsentation, grüne Pfeile) sowie erkanntes Objekt (roter Rahmen), aus [Rohner20a]

Hypothesen falsifiziert werden. Die Flächen für die Validierung R_H sind somit alle Flächen von Hypothesen (F_H), die keine Korrespondenz in der Szene haben:

$$R_H = \{g \in F_H | \nexists f \in F_W : \texttt{explainedby}(g, f)\}$$

4.2.4 Bestimmung von möglichen Sichten

Ausgehend von den Flächen als Untersuchungsregionen müssen mögliche Sichten berechnet werden, die entweder zur Validierung oder Exploration beitragen. Die Unterteilung ist auch bei der Berechnung möglicher Sichten notwendig, da für den Fall der Exploration nicht gezielt die Fläche betrachtet werden soll, sondern die nähere Umgebung. Bei den Flächen der Validierung ist dagegen die konkrete Fläche von Interesse.
Für die Exploration können die Flächen dadurch genutzt werden, dass über die dazugehörigen Kanten auf die Umgebung geschlossen werden kann. Dazu wird für jede Kante e_f einer unerklärten Fläche $f \in R_W$ der Mittelpunkt c_f mit Hilfe der Vertizes der Kante bestimmt. Darüber hinaus ist die Normale $\vec{n}_f \in \mathbb{R}^3$ der Fläche f bekannt. Über diese Normale lässt sich eine Blickrichtung bestimmen, in dem die Richtung der Normale invertiert wird. Zusätzlich kann diese entlang der Kante e_f von der Fläche f weg rotiert werden, sodass nicht direkt die Fläche f betrachtet wird. Die Drehrichtung ist durch die zugrundeliegende Datenstruktur bekannt. Aus der Halb-Kanten Darstellung von B-Reps und der standardisierten Drehrichtung von Kanten entlang von Flächen lässt sich diese Richtung ableiten. Der Winkel, um den die Kante gedreht wird, hängt dabei vom verwendeten Tiefensensor ab. Ein intuitiver Vorschlag sind 45 Grad, da so mit hoher Wahrscheinlichkeit angrenzende Flächen unabhängig von der vorliegenden Form gesehen werden. Durch die Rotation ist die Blickrichtung auf die Fläche gegeben; um eine vollständige Pose für den Tiefensensor zu bestimmen, ist auch noch eine Translation notwendig.
Ausgehend von dem Schwerpunkt der Kante c_E kann die Translation mit Hilfe des Abstandes

$d_E \in \mathbb{R}^+$ bestimmt werden. Da die Richtung bereits bekannt ist, kann so der Punkt im Dreidimensionalen eindeutig bestimmt werden. Eine mögliche Sicht setzt sich aus dem Aufpunkt $\vec{p}_E \in \mathbb{R}^3$ und der Sichtrichtung zusammen, welche später für die Berechnung der Pose genutzt werden. Der Wert für den Parameter d_E hängt dabei von dem verwendeten Sensor und der aktuellen Szene ab. So sind für manche Tiefensensoren Mindestabstände zum Objekt einzuhalten, damit das Messprinzip angewendet werden kann. Die Größe der Objekte ist ebenfalls relevant, da bei kleineren Objekten auch ein kleinerer Abstand zu diesen sinnvoll ist.
Für die Validierung wird ein ähnlicher Prozess vorgeschlagen. Hierbei soll die Untersuchungsregion (also eine konkrete Fläche) $f_V \in R_H$ direkt betrachtet werden, um diese zu validieren. Als Aufpunkt $\vec{p}_V \in \mathbb{R}^3$ wird daher der Mittelpunkt c_V der Fläche genutzt. Als Sichtrichtung $\vec{n}_V \in \mathbb{R}^3$ kann direkt die invertierte Flächennormale genutzt werden, da zum einen eben diese Fläche betrachtet werden soll und zum anderen die Qualität der Tiefendaten höher ist, wenn der Sichtstrahl möglichst senkrecht zur Fläche steht. Eine Anpassung des Winkels ist dabei trotzdem möglich, falls mehr Abstand zur Arbeitsoberfläche erzeugt werden soll.
Mögliche Sichten (unabhängig davon, ob für die Exploration oder die Validierung) lassen sich als $v = (\vec{p}, \vec{n}, q, t)$ definieren. Eine Sicht setzt sich aus dem Aufpunkt $\vec{p} \in \mathbb{R}^3$, der Sichtrichtung $\vec{n} \in \mathbb{R}^3$, dem Typ t der Sicht basierend auf dem Kantentyp und einem Qualitätsmaß $q \in \mathbb{R}^+$ zusammen. Für den Typ t der Sicht gibt es insgesamt drei Möglichkeiten: Zum einen kann es eine Sicht für die *Validierung* sein, zum anderen für die Exploration ausgehend von einem *Tiefensprung*. Schließlich kann es noch eine Sicht für die Exploration sein, ausgehend von einem *Beobachtungssprung*. Für Kanten mit dem Typ *Normalensprung* werden nach dieser Definition keine möglichen Sichten erzeugt. Dem liegt die Tatsache zu Grunde, dass für diese Kanten alle benachbarten Flächen bereits rekonstruiert wurden. Entlang dieser Kanten weiter zu explorieren, würde daher kein zusätzliches Wissen für die Objektwiedererkennung generieren.
Als Qualitätsmaß q kann die Länge der Kanten e_f verwendet werden (für die Exploration) oder der Flächeninhalt der Fläche f_V (für die Validierung). Somit werden Kanten bevorzugt, die mit höherer Wahrscheinlichkeit Einfluss auf die Szene nehmen und benachbarte Flächen mit rekonstruieren. Die Länge der Kanten gibt zudem einen Hinweis, ob eine Kante tatsächlich rekonstruiert werden soll und von Interesse ist oder nur aufgrund von Rauschen entstanden ist. Für die Flächen ist deren Flächeninhalt ein geeignetes Kriterium, da größere Hypothesenflächen einen größeren Einfluss bei der Hypothesenauswahl haben als kleinere, und sich somit ein Fehler mehr auswirken würde. Ein anderes Kriterium wäre beispielsweise der Anteil der Flächen der korrespondierenden Hypothese, die bereits durch die Szene erklärt sind. So würde sich der Fokus mehr auf wenig erklärte Hypothesen verschieben.

4.2.5 Auswahl möglicher Sichten und Posenbestimmung

Alle berechneten möglichen Sichten werden in einer Menge Φ gesammelt, die sowohl die Sichten für die Exploration basierend auf $f_E \in R_W$ und Validierungssichten für $f_V \in R_H$ beinhaltet. Diese Sichten wurden nur lokal basierend auf der Geometrie einzelner Flächen bestimmt. Daher ist es möglich, dass einzelne Sichten aus unterschiedlichen Gründen verworfen werden müssen: Die zugrundeliegende Kante ist zu klein und wahrscheinlich aufgrund von Rauschen entstanden. Da eine solche Kante zudem zur restlichen Fläche oft verdreht ist, würde diese

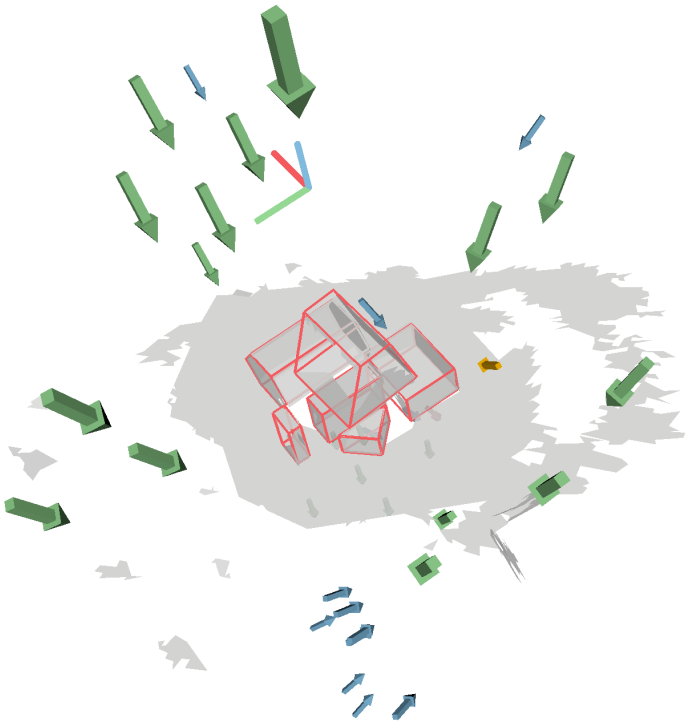


Abbildung 4.3: Alle möglichen Explorations- und Validierungssichten (Exploration in blauen und gelben Pfeilen, Validierung in grün), sowie eingezeichnetes Weltkoordinatensystem in der Roboterbasis, aus [Rohner20a]

Sicht aus einer ungeeigneten Richtung auf die Fläche schauen. Darüber hinaus kann sich die Sicht zu nah an der Szene oder am Roboter befinden (in Abhängigkeit der gewählten Parameter), was zu Kollisionen entweder mit der Szene oder dem Roboter selbst führen kann. Durch die Drehrichtung der Kanten kann es weiterhin passieren, dass Sichten erzeugt werden, die Objekte von unterhalb der Arbeitsfläche betrachten würden. Abschließend können Sichten erzeugt werden, für die es unmöglich ist, die zu untersuchende Fläche wahrzunehmen. Grund dafür sind Verdeckungen durch andere Rekonstruktionen in der Szene. Die erzeugte Sicht müsste somit durch eine bereits bestehende Fläche hindurchschauen, um die Untersuchungsregionen wahrzunehmen.
Alle Sichten in Φ werden mit den folgenden Methoden gefiltert (prototypisch für die bereits verwendeten Arbeitsflächen): Jede Sicht mit einem Qualitätsmaß $q < q_{\text{exp}} \in \mathbb{R}$ und dem Typ t der *Exploration* wird verworfen. Für die Kollisionserkennung wird zunächst eine heuristische Filterung durchgeführt: Alle Sichten, die einen Abstand kleiner als $d_{\text{ws}} \in \mathbb{R}^+$ zur Arbeitsoberfläche haben, werden aus S entfernt. Analog wird dieser Schritt durchgeführt, wenn der Abstand zum Roboter kleiner als $d_{\text{bot}} \in \mathbb{R}^+$ oder als $d_{\text{sce}} \in \mathbb{R}^+$ für den Abstand zur rekonstruierten Szene ist. Weiterführende Kollisionserkennung kann später berechnet werden, um auch Selbstkollision während Transferfahrten sicher auszuschließen (siehe dazu den Ausblick). Weiterhin müssen alle Sichten entfernt werden, die lediglich die Arbeitsfläche wahrnehmen würden. Je nach Art und Größe der Arbeitsfläche lässt sich dieser Filter direkt umsetzen. Bei einer einfach Arbeitsplatte muss nur die z-Richtung des Richtungsvektors betrachtet werden. Dazu wird für eine Normale $\vec{n} = (x, y, z)$ geprüft, ob $z < 0$ erfüllt ist. Bei komplexen Arbeitsoberflächen, die beispielsweise voneinander getrennt im Raum verteilt sind, ist diese Berechnung

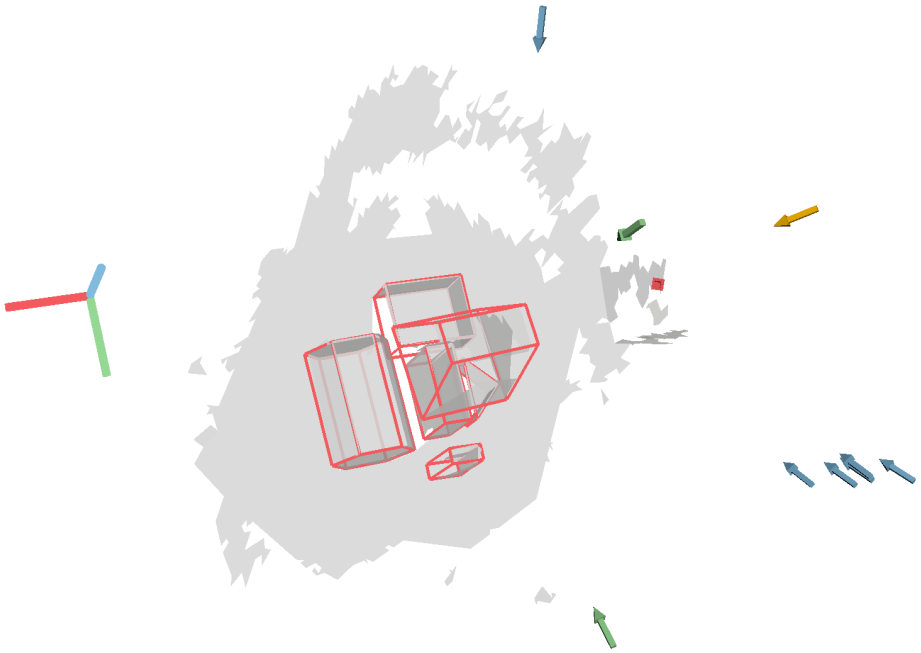


Abbildung 4.4: Alle verbleibenden Sichten nach der Filterung. Explorationssichten durch blaue und gelbe Pfeile dargestellt, Validierungssichten in grün, sowie die nächste Sicht in rot, aus [Rohner20a]

aufwändiger. Die Verdeckungsüberprüfung lässt sich darüber bestimmen, ob der erzeugte Sichtstrahl bis zur Untersuchungsregion auf Schnitt sowohl mit der Szene als auch mit den berechneten Hypothesen getestet wird. Falls ein Schnitt vorliegt, wird die Sicht verworfen.
Für alle gefilterten Sichten ist davon auszugehen, dass diese einen Beitrag für die Wiedererkennung der vollständigen Szene leisten. Aus diesen Sichten muss eine ausgewählt und die dazugehörige Pose bestimmt werden. Falls für diese Pose die inverse Kinematik nicht gelöst werden kann, wird die zugrundeliegende Sicht verworfen, die nächstbeste ausgewählt und der Prozess wiederholt. Falls keine Sicht mehr vorliegt, terminiert der Prozess.
Für die Auswahl der Sicht werden der Typ t und das Qualitätsmaß q verwendet. Bei der Auswahl der Sichten hat die *Validerung* die geringste Priorität, da für korrespondierende Flächen zumindest schon eine Hypothese existiert. Explorationsflächen sind daher mehr von Interesse, da durch diese die Szene schneller vollständig erklärt wird. Bei der *Exploration* haben schließlich Sichten mit einem *Tiefensprung* Vorrang vor Sichten mit einem *Beobachtungssprung*, da bei *Tiefensprung*-Kanten sichergestellt ist, dass zwischen den zwei rekonstruierten Flächen eine bisher nicht rekonstruierbare Kante existieren muss (sonst wäre es kein Tiefensprung). Durch diese Kante wird eine neue Fläche wahrgenommen, wodurch neue Deskriptoren bestimmt werden. Für *Beobachtungssprünge* ist diese Aussage nicht möglich, da keine gesicherte Information vorliegt, wie diese Fläche im weiteren Verlauf rekonstruiert wird. Beispielsweise kann eine Fläche aufgrund der Rekonstruktion unterbrochen sein und wird durch eine zusätzliche Sicht verknüpft, wodurch keine neue Information für die Objektwiederkennung vorliegt. Da eine klare Hierarchie bezüglich des Typs vorliegt, können die Qualitätsmaße innerhalb der Typen verglichen werden (wodurch keine Gewichtung zwischen Kantenlänge und Flächeninhalt stattfinden muss). Dementsprechend wird die Sicht mit dem größten Qualitätsmaß gewählt.

Insgesamt wird somit die längste Explorationskante basierend auf einem Tiefensprung ausgewählt (sofern vorhanden).
Für eine ausgewählte Sicht wird nun die 3D-Pose $P \in \mathbb{R}^{4\times4}$ (als affine Transformationsmatrix) berechnet: Der translatorische Anteil von P entspricht dem Aufpunkt $\vec{p}$ der aktuellen Sicht. Um die drei Rotationsachsen zu bestimmen, wird die Sichtrichtung $\vec{n}$ als erste Komponente festgelegt. Um Verdeckungen mit dem Roboter zu minimieren, wird weiterhin festgelegt, dass die Kamera sich unterhalb des Greifers befinden soll. Hierüber wird die zweite Achse definiert. Je nach Montage und Aufbau der Kamera und des Robotersystems sind hier Alternativen möglich. Eine feste Entscheidung erleichtert allerdings die Berechnung der Pose, da sonst ein Freiheitsgrad offen bleibt. Gleichzeitig wird dadurch die Lösbarkeit der inversen Kinematik erschwert. Um einen Freiheitsgrad zu fixieren, wird der Vektor $\vec{\text{down}} = \begin{pmatrix}0 & 0 & -1\end{pmatrix}^T$ definiert. Die zweite Komponente berechnet sich somit als $\vec{\text{down}} \times \vec{n}$. Damit ein rechtshändiges Koordinatensystem entsteht, wird die letzte Komponente als $\vec{y} = \vec{z} \times \vec{x}$ berechnet. Die gesamte Pose in Weltkoordinaten bestimmt sich als:

$$P = \begin{pmatrix} \vec{\text{down}} \times \vec{n} & \vec{n} \times (\vec{\text{down}} \times \vec{n}) & \vec{n} & \vec{p} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Bevor die inverse Kinematik für die berechnete Pose bestimmt wird, muss geprüft werden, ob diese Pose bereits schon einmal angefahren wurde. Da trotz der Heuristiken der Fall eintreten kann, dass keine neue Information an einer Pose generiert wird, wäre der Prozess in einem Deadlock. Solange sich an der Szene nichts ändert, wird immer wieder die gleiche Sicht als Kandidat gewählt. Allerdings erzeugt diese keine neuen Informationen für die Szene, und der Prozess wiederholt sich. Für dieses Problem gibt es unterschiedliche Lösungen. Beispielsweise kann man die letzte Pose oder alle bisher angefahrenen speichern und so Wiederholung vermeiden. Alternativ kann die Pose leicht variiert werden, so dass trotz gleicher Sicht neue Information erzeugt wird.
Im Falle einer unlösbaren inversen Kinematik wird die Pose leicht angepasst, um so lokalen Lösungsproblemen der inversen Kinematik (beispielsweise aufgrund geometrischer oder numerischer Einschränkungen) zu entgehen. Da eine Rotation das Ergebnis der Active Vision negativ beeinflussen kann (beispielsweise durch einen daraus resultierenden ungünstigen Blickwinkel, gegebenenfalls verstärkt durch das Messprinzip des Tiefensensors), soll nur die Translation modifiziert werden. Da der Sensor durch die gewählten Parameter für die Sichtenerzeugung bereits vergleichsweise nah an der Szene ist, wird entlang der invertierten Sichtrichtung die Pose neu bestimmt und versucht, die inverse Kinematik zu lösen. Falls dies nicht erfolgreich ist, wird die Pose als unlösbar verworfen (beispielsweise wenn sie außerhalb vom Arbeitsraum des Roboters liegt) und eine neue Pose für die nächste Sicht berechnet. Alternativ kann anstatt der Linie das Volumen des invertierten Sichtkegels abgetastet werden, da auch bei einem Versatz der Pose aufgrund des Messprinzipes die relevanten Flächen erkannt werden.

4.2.6 Initialisierung des Weltmodells

Bevor neue Sichten bestimmt werden können, muss eine initiale Repräsentation der Welt erzeugt werden. Initial liegt eine leere Welt vor, da keine Annahmen zur Anwesenheit von bestimmten Objekten oder Flächen getroffen werden können. Somit muss eine Initialisierung des Weltmodells erfolgen. Je nach Ausdehnung und Beschaffenheit der Arbeitsoberfläche reicht dazu zunächst eine einzelne feste Sicht. Bei mehreren, verteilten Arbeitsbereichen sollte für jeden einzelnen eine initiale Repräsentation erzeugt werden. Wie in der Evaluation im vorangegangenen Kapitel gezeigt wurde, reicht für spezifische Objekte und Szenentypen bereits eine einzelne Sicht, um Objekte korrekt wiederzuerkennen. Alternativ kann das Problem der Initialisierung vollständig gelöst werden, wenn die Arbeitsfläche und die Kamera genau bestimmt sind. Mit den Kameraparametern, vor allem dem Frustum und der bekannten Arbeitsoberfläche, kann eine minimale Anzahl an Sichten berechnet werden, sodass die Arbeitsfläche vollständig erfasst wird.

4.3 Experimentelle Auswertung

Dieser Abschnitt umfasst die experimentelle Auswertung, beginnend mit dem Aufbau der Experimente und der anschließenden Evaluation. Die verwendete Hardware sowie Objekte sind identisch zu Kapitel 3.

4.3.1 Aufbau

Nachdem in der Evaluation von Kapitel 3 die Notwendigkeit von mehreren Sichten evaluiert wurde, ist an dieser Stelle von Interesse, inwieweit diese Herausforderung durch den hier entwickelten Ansatz gelöst ist. Die Evaluation ist an die Experimente aus dem vorherigen Kapitel angelehnt, wobei in diesen Fällen die nächsten Posen mit dem hier vorgestellten Ansatz bestimmt und angefahren werden.

Einzelne Objekte

In diesem Experiment werden alle Objekte untersucht, die im vorangegangenen Kapitel nicht durch eine einzelne Sicht erkannt wurden. Es ist zu untersuchen, wie viele Posen notwendig waren, um das Objekt korrekt zu erkennen (falls das Objekt aufgrund veränderter Ausgangspose oder Unterschieden bei der Rekonstruktion aufgrund von Rauschen nicht direkt erkannt wurde). Weiterhin kann die Anzahl an Flächen in der Szene näher betrachtet werden (ausgenommen der Arbeitsoberfläche). Dabei kann die Anzahl an Flächen (wie im Ansatz, Abschnitt 4.2.2 diskutiert) in Explorations- und Validierungsflächen wie auch in die entsprechende Anzahl an möglichen Sichten aufgeteilt werden. Dabei beschreibt die Zahl der Flächen die absolute Menge in der Umweltrepräsentation, wobei bei der Anzahl Sichten ausschließlich die reduzierte Anzahl nach dem Anwenden der vorgestellten Filter gezählt wird. Somit ist die Anzahl an Flächen stets größer oder gleich der Anzahl entsprechender Sichten.

Objekt	Rekonst. Flächen	Exp.-flächen	Val.-flächen	Exp.-sichten	Val.-sichten	Notwend. Posen
Tetrahed.	5	0	4	0	0	2
Cube1	4	0	3	0	0	2
Cube2	5	0	5	0	0	1
Cuboid	4	0	3	0	0	2
Triangle	9	0	5	0	0	2
Tea1	4	0	4	0	0	1
Sweets1	6	0	10	0	0	1
Filter	5	0	4	0	0	2
Peg	5	0	3	0	0	2
Tea2	4	0	3	0	0	2
Tea3	4	0	5	0	0	1
Sponge	5	0	4	0	0	2

Tabelle 4.1: Überblick Wiederkennung einzelner Objekte mittels Active Vision: Für jedes Objekt ist die Anzahl an rekonstruierten Flächen in der Szene, Explorationsflächen, Validierungsflächen, Explorationssichten, Validierungssichten und notwendigen Posen angegeben.

Mehrere Objekte

Neben den einzelnen Objekten sind vollständige Szenen von Interesse sowie der Verlauf der Anzahl Flächen und Sichten und auch die Erkennungsrate als eigentliches Ziel des Active Visions Prozesses. Analog zu der Evaluation in Kapitel 3 sind vor allem Verdeckungen innerhalb der Szene eine Herausforderung. Neben der Schwierigkeit Objekte wahrzunehmen, müssen diese auch bei der Kollisionserkennung für die Filterung von Sichten berücksichtigt werden. Je mehr Objekte in einer Szene vorhanden sind, desto mehr Sichten werden herausgefiltert, um Kollisionen mit der Szene zu vermeiden.

4.3.2 Ergebnisse

In diesem Abschnitt werden die Ergebnisse der Szenarien vorgestellt und diskutiert.

Einzelne Objekte

Von allen Objekten, die in Kapitel 3 nicht direkt erkannt wurden, konnten mit Hilfe von Active Vision alle Objekte nach maximal zwei Posen korrekt erkannt werden, in vier von zwölf Fällen hat eine einzelne Pose bereits ausgereicht. Die Anzahl der Explorationsflächen war nach korrekter Erkennung in jedem Fall null. Das entspricht den Erwartungen, da alle Objektflächen durch eine Hypothese beschrieben werden. Die Anzahl an Validierungsflächen variiert stark je nach Objekt, aufgenommenen Sichten und Modellierungsgenauigkeit. Da keine Explorationsflächen am Ende vorlagen, gibt es dementsprechend auch keine Explorationssichten. Auch für die Validierungsflächen konnten keine Sichten bestimmt werden, da die zugehörigen Sichten zu den Flächen aufgrund der diversen Filter entfernt wurden. Gerade die Filter, die Kollisionen verhindern sollen, entfernen mögliche Sichten. In Tabelle 4.1 sind die Ergebnisse detailliert dargestellt. Als Beispiel sind die angefahrene Pose, berechnete Sichten und Rekonstruktion für ausgewählte Objekte in Abbildung 4.5 und 4.6 dargestellt.

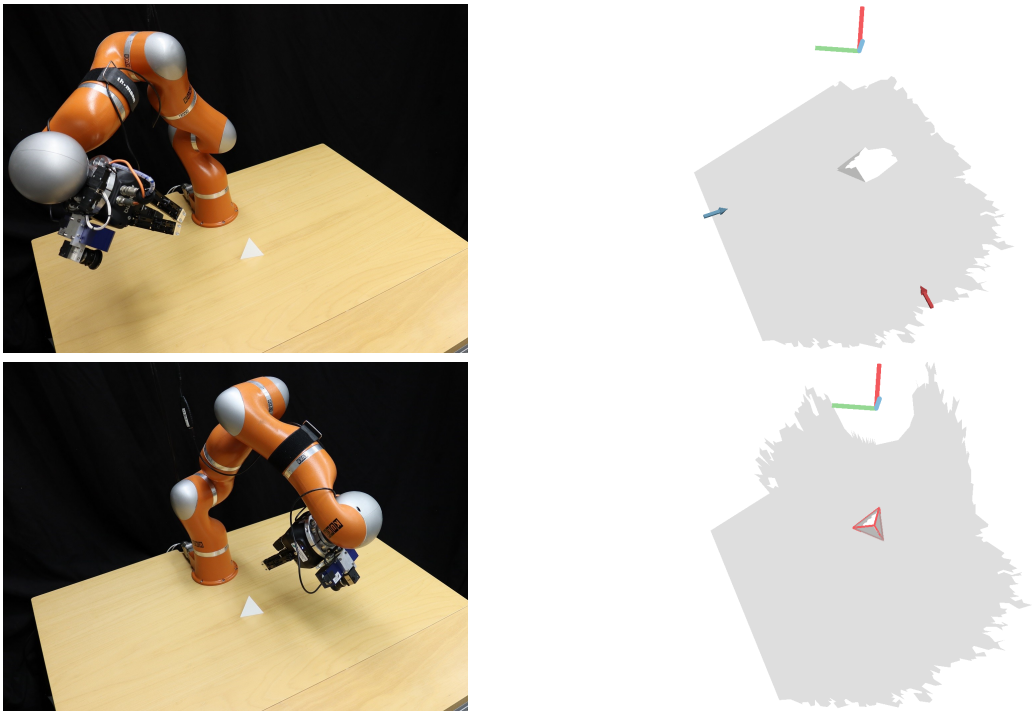


Abbildung 4.5: Wiedererkennung des einzelnen Objekts *Tetrahedron* mit Active Vision aus zwei Sichten

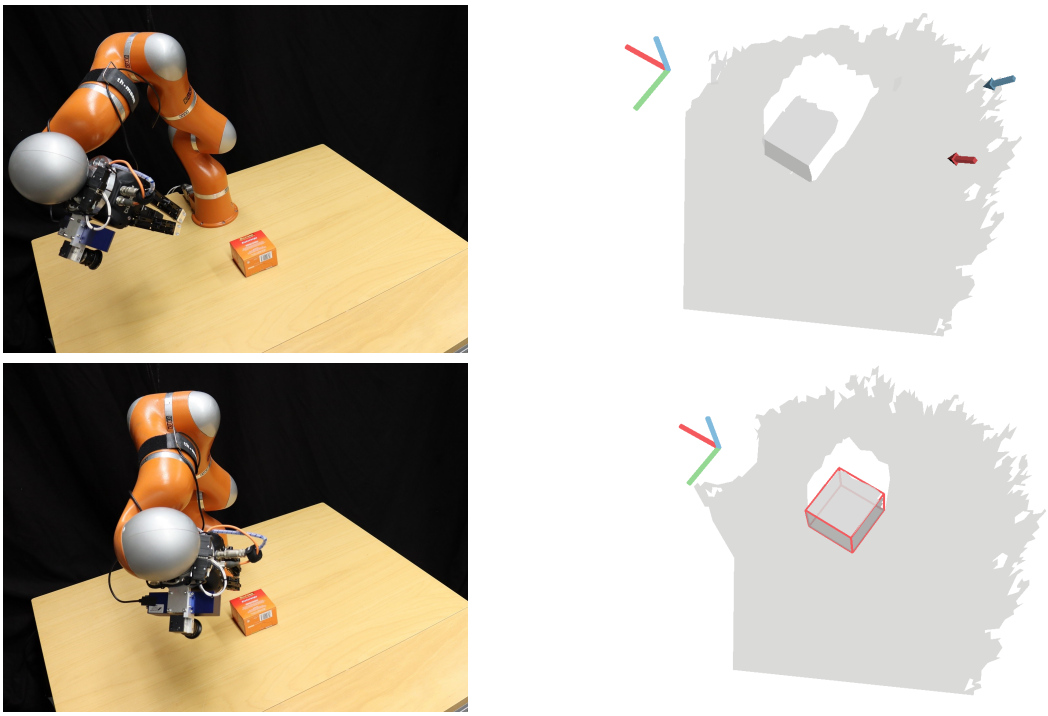


Abbildung 4.6: Wiedererkennung des einzelnen Objekts *Tea2* mit Active Vision aus zwei Sichten

Mehrere Objekte

Zunächst werden die Ergebnisse für eine spezifische Szene vorgestellt (Abbildung 4.7). In jeder Zeile ist links die aktuelle Szene und Pose vom Roboter zu sehen, rechts die entsprechende Rekonstruktion mit erkannten Objekten (rote Rahmen) und mögliche Sichten (farbige Pfeile, wobei der rote Pfeil die Sicht mit dem höchsten Qualitätsmaß anzeigt). In der ersten Zeile wird die Szene aus der bereits bekannten initialen Pose aufgezeichnet. Wie in Kapitel 3 bereits diskutiert, werden manche Objekte bereits erkannt. Aufgrund einer nicht lösbaren inversen Kinematik wird die qualitativ beste Sicht nicht angefahren, nichtsdestoweniger aber eine Fläche, die zum gleichen Objekt gehört. Allerdings ist diese Fläche nicht linear unabhängig zu anderen, bereits rekonstruierten Flächen vom Objekt *hole*, wodurch dieses in diesem Schritt nicht richtig erkannt wird. Die dritte Sicht konzentriert sich gezielt auf das Objekt *tea3*, wodurch dieses korrekt erkannt wird, wie auch das Objekt *peg*. Zudem wird auch ein Großteil der weiteren Arbeitsfläche aufgenommen. In Sicht vier wird schließlich eine weitere Fläche vom Objekt *hole* rekonstruiert, wodurch das Objekt korrekt erkannt wird und die nicht anfahrbare Explorationssicht nicht mehr benötigt wird. Abschließend werden mit einer Sicht von oben einige Flächen validiert. Die abschließende Sicht verbleibt, da die dazugehörige Fläche aufgrund von Verdeckungen nicht validiert werden kann.
Neben den tatsächlichen Posen des Roboters ist die Entwicklung der Anzahl an Sichten und Flächen von Interesse. In Abbildung 4.8 ist links eine Rekonstruktion nach insgesamt acht Sichten abgebildet und rechts der Verlauf der Anzahl der Flächen gegenüber der entsprechenden Sicht. Insgesamt sind neun Objekte in der Szene vorhanden. Nach den ersten zwei Sichten werden drei Objekte korrekt erkannt. Durch die nächsten zwei Sichten steigt die Anzahl an korrekt erkannten Objekten an, gleichzeitig sinkt die Anzahl an Explorationsflächen, da die Flächen der eben erkannten Objekte wegfallen. Dafür steigt die Zahl an Validierungsflächen mit steigender Anzahl an erkannten Objekten an. Im weiteren Verlauf steigt die Anzahl an Explorationsflächen an, da andere Objekte teilweise rekonstruiert, aber noch nicht richtig erkannt werden. Gleichermaßen werden neben der tatsächlichen Explorationsfläche auch Validierungsflächen rekonstruiert, wodurch die Anzahl zu Sicht fünf abnimmt. Im weiteren Verlauf werden alle Objekte richtig erkannt, wodurch in Sicht sieben und acht keine Explorationsflächen mehr vorliegen. In Sicht sieben steigt die Anzahl an Validierungsflächen auf das Maximum an. In Sicht acht wird abschließend eine Validierungsfläche aufgezeichnet, wodurch die Anzahl an Validierungsflächen wieder sinkt.
Abschließend nun einige komplexe Szenen, die mit einer unterschiedlichen Anzahl von Sichten rekonstruiert werden. Ein Überblick findet sich in Tabelle 4.2, Bilder einer ausgewählten Sicht auf die entsprechenden Szenen in Abbildung 4.10 sowie alle Sichten und Rekonstruktionen der ersten Szene in Abbildung 4.9. Dabei bestätigen sich die Ergebnisse der vorangegangenen Experimente. Einzelne Beobachtungen sind explizit hervorzuheben: Fehlende Erkennungen sind auf nicht rekonstruierbare Flächen zurückzuführen, die für eine korrekte Wiedererkennung notwendig wären (beispielsweise Objekte *Triangle* in Szene 1). Bei Objekten mit ähnlicher Flächenanordnung und damit ähnlichen Deskriptoren kam es gelegentlich zu Verwechslungen. Teilweise wurden diese im Laufe des Active Vision Prozesses noch behoben. Dies ist zu einem Teil durch Verdeckungen zu erklären, zum anderen bevorzugt das

Abbildung 4.7: Vollständige Rekonstruktion einer Szene in fünf Schritten. Pro Reihe links die aktuelle Pose des Roboters und rechts die Rekonstruktion mit berechneten Sichten (Pfeile) und erkannten Objekten (Rahmen), aus [Rohner20a]

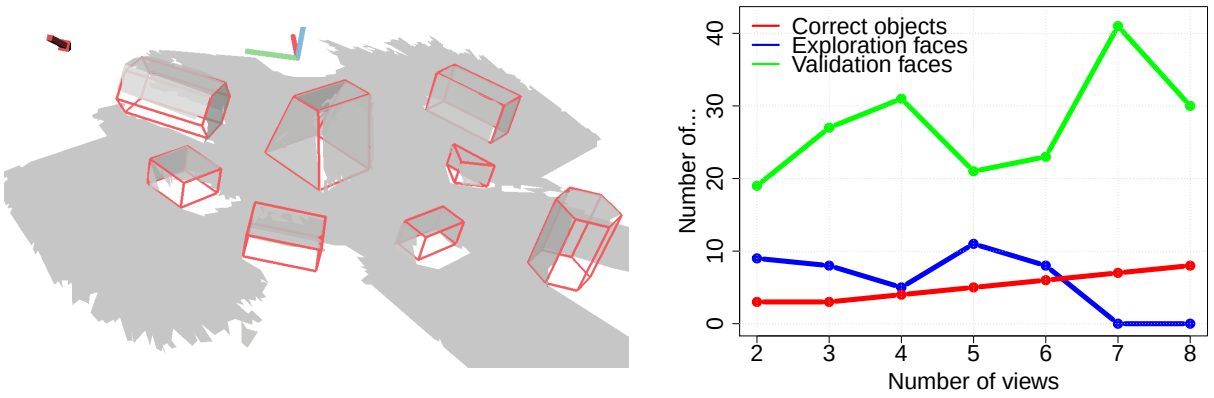


Abbildung 4.8: Vollständige Rekonstruktion einer Szene nach acht Sichten: Links die finale Rekonstruktion mit allen korrekt erkannten Objekten, rechts der Verlauf an korrekt erkannten Objekten, Explorationsflächen und Validierungsflächen nach jeder Sicht, aus [Rohner20a]

Qualitätsmaß für die Hypothesenbewertung kleinere Objekte. Falls ein Objekt nur teilweise rekonstruiert wird, diese Flächen aber sehr gut von einer falschen Hypothese erklärt werden, wird zunächst diese verwendet. Eine Validierung der Hypothesenflächen ist dabei nicht immer möglich. Neben den Verdeckungen können häufig Validierungssichten nicht angefahren werden: entweder aufgrund der Vorfilterung oder einer nicht lösbaren inversen Kinematik (was Sichten außerhalb des Arbeitsbereiches des Roboters mit einschließt). Dieses Verhalten lässt sich am Vergleich der Validierungsflächen mit den Validierungssichten erkennen. In einem Fall (Szene 3) sinkt die Anzahl an rekonstruierten Flächen in der Welt von Sicht 2 zu Sicht 3. Das ist dadurch zu erklären, dass zum einen durch die dritte Sicht nur eine sehr geringe Anzahl bisher ungesehener Flächen hinzugefügt wurde, dafür aber manche Flächen zusammengefügt werden, die zuvor in mehrere Teilflächen zerfallen waren. Dass einzelne Posen nicht angefahren können, ist auch der Grund für verbleibende Explorations- und Validierungssichten, da diese aufgrund der räumlichen Ausdehnung der Szene und geometrischer Einschränkungen des Roboters nicht erreichbar sind.

Szene/ Objekte	Sicht	Rekonst. Flächen	Exp.- flächen	Val.- flächen	Exp.- sichten	Val.- sichten	Korrekt erkannt	Falsch erkannt
1/5	1	12	11	16	3	1	0	2
	2	17	18	6	1	0	1	1
	3	22	6	14	0	0	3	1
2/5	1	14	10	28	3	4	2	1
	2	23	10	40	2	2	2	2
	3	24	3	30	1	0	4	1
3/5	1	17	21	21	4	4	2	1
	2	23	22	18	4	4	2	0
	3	22	11	27	4	5	3	0
	4	33	18	22	5	1	3	0
	5	36	7	21	0	0	4	0
4/5	1	12	13	17	2	1	1	0
	2	16	7	31	2	2	3	0
	3	20	0	32	0	1	4	0
5/5	1	16	10	24	1	1	3	1
	2	19	4	21	0	0	5	0
	3	19	4	21	0	0	5	0
6/12	1	34	11	65	2	6	6	2
	2	49	14	85	6	9	9	1
	3	66	20	78	9	3	9	1
7/10	1	44	15	69	6	14	7	0
	2	54	22	72	8	14	8	0
	3	63	11	77	1	13	9	0

Tabelle 4.2: Zusammenfassung der Evaluationsszenen: Für jede Szene sind die Anzahl an Objekten, sowie die rekonstruierten Flächen, Explorationsflächen, Validierungsflächen, Explorationssichten, Validierungssichten, korrekt und falsch erkannte Objekte nach jeder Rekonstruktion angegeben.

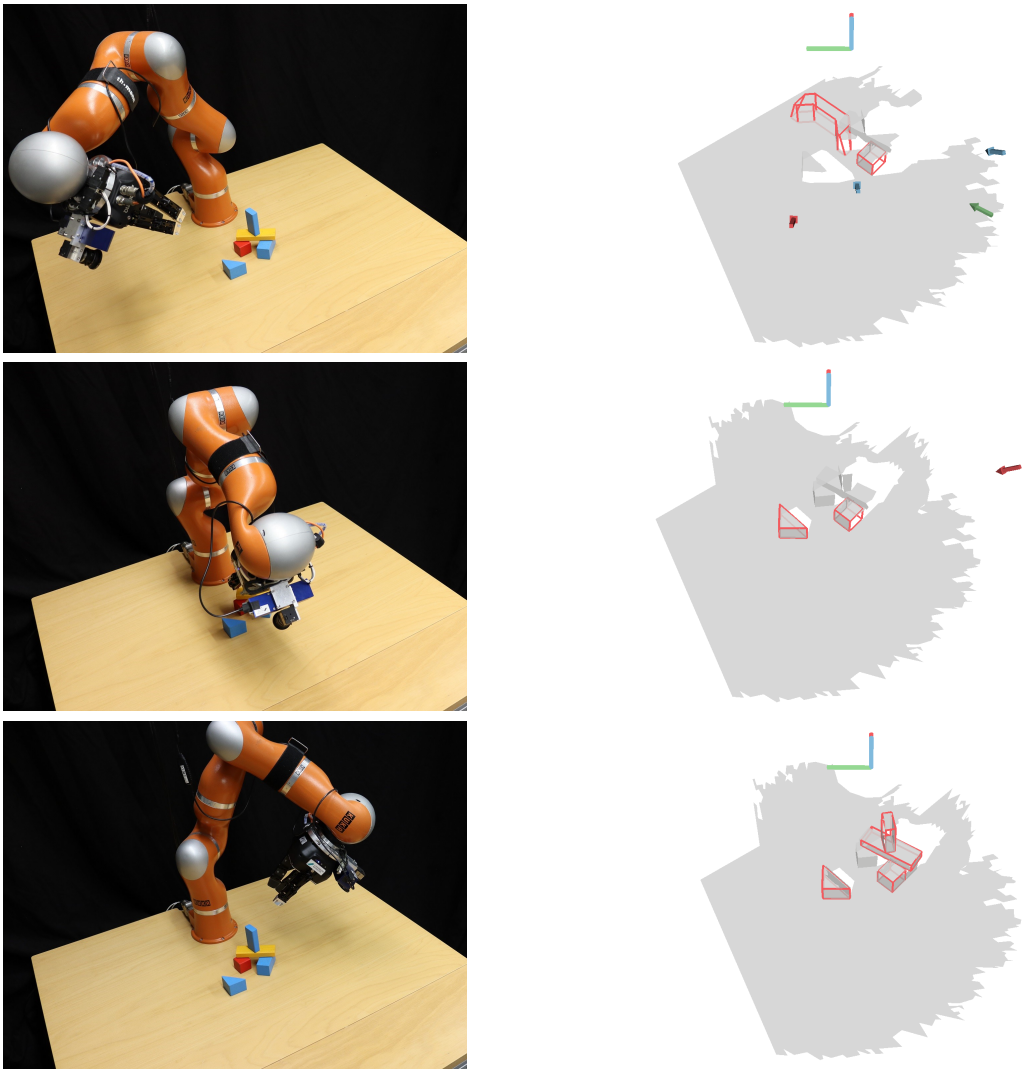
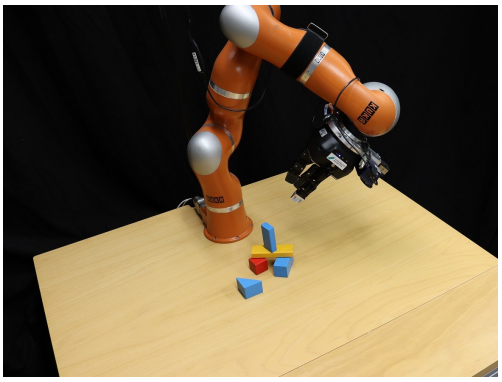
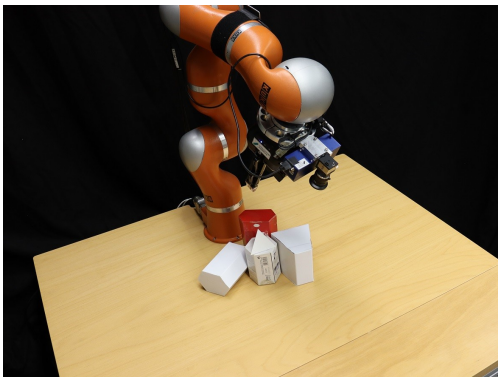


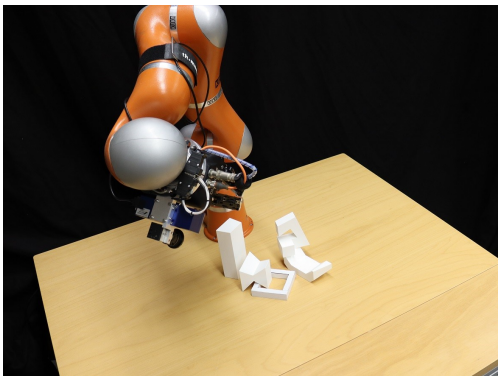
Abbildung 4.9: Einzelne Sichten und Rekonstruktionen der ersten Szene aus Tabelle 4.2



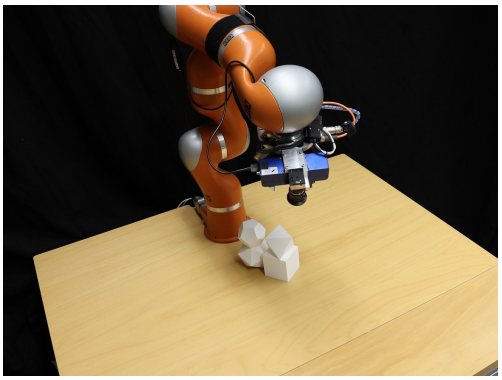
(a) Szene 1, Sicht 3



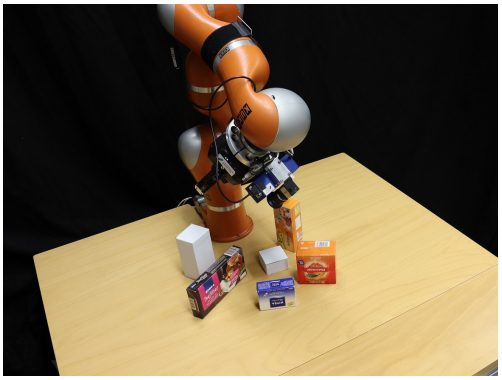
(b) Szene 2, Sicht 2



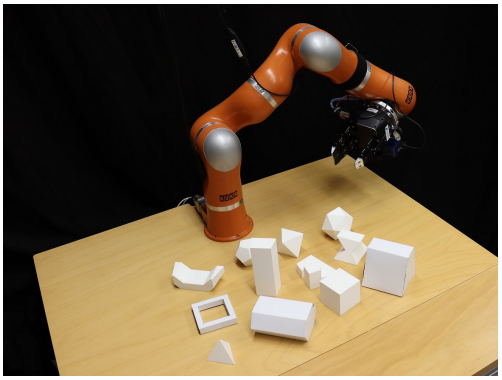
(c) Szene 3, Sicht 5



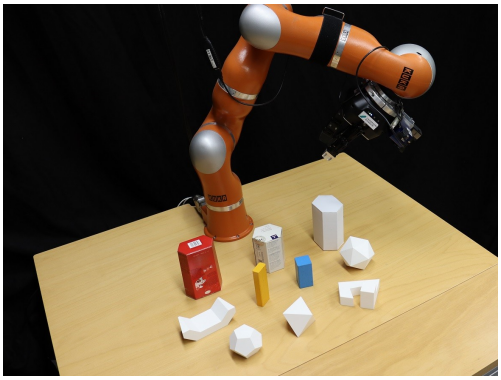
(d) Szene 4, Sicht 2



(e) Szene 5, Sicht 2



(f) Szene 6, Sicht 3



(g) Szene 7, Sicht 3

Abbildung 4.10: Active Vision Prozess für mehrere Szenen mit unterschiedlichen Sichten. Die Szenen unterscheiden sich dabei in der Komplexität, die ausgewählten Sichten illustrieren unterschiedliche mögliche Posen. Siehe Abbildung 4.2 für die Ergebnisse.

4.4 Fazit

Dieses Kapitel beschreibt einen Ansatz zur Erzeugung neuer Sicht für die vollständige, automatische Rekonstruktion einer Szene und die korrekte Erkennung aller vorkommenden Objekte. Dazu wird hinsichtlich des Zwecks einer Sicht unterschieden: Das ist zum einen die Exploration, um neue Hypothesen generieren zu können, zum anderen die Validierung, um bestehende Hypothesen zu überprüfen. Um neue Sichten zu bestimmen, wird dabei die Eigenschaft von der Umweltrekonstruktion verwendet, die verwaltet wie eine Kante entstanden ist. Die möglichen Sichten werden abschließend heuristisch gefiltert, um ungeeignete Sichten aufgrund von Verdeckungen oder Kollisionen des Roboters zu vermeiden. In der Evaluation wurde gezeigt, dass die neuen, aus der Szene abgeleiteten Sichten einen höheren Mehrwert für die Wiedererkennung haben als festgelegte Sichten wie in Kapitel 3. Die größten Einschränkungen waren dabei Verdeckungen durch die Szenen sowie nicht erreichbare Sichten aufgrund der Robotergeometrie beziehungsweise der inversen Kinematik.
Somit lässt sich die Fragestellung

F2 Inwieweit unterstützen zusätzliche lokale Sichten die Objektwiedererkennung? Inwieweit können neue Sichten szenenspezifisch bestimmt werden?

dahingehend beantworten, dass zusätzliche Sichten notwendig sind, um alle Objekte vollständig wiederzuerkennen, vor allem Objekte ohne Textur. Durch Verwendung von B-Reps kann die Unterstützung der Wiedererkennung gezielt erfolgen, indem bisher noch nicht sichtbare Flächen aufgrund eines Tiefensprungs priorisiert werden. Diese Information hilft auch die Sichten spezifisch für die aktuelle Szene beziehungsweise Rekonstruktion zu bestimmen. Eingeschränkt ist der entwickelte Ansatz aufgrund der Filterung möglicher Sichten, da potentiell notwendige Posen entfernt werden. Zudem kann die Einschränkung auf drei linear unabhängige Flächen für die Hypothesenerzeugung zwar durch szenenspezifische Sichten entschärft, aber nicht vollständig gelöst werden.
Als mögliche Erweiterung bietet sich eine Verfeinerung der Kollisionserkennung an, sodass die heuristischen Filter entschärft werden können. Das schließt Selbstkollisionen als auch Kollisionen mit der Arbeitsfläche oder der aktuellen Szene mit ein. Darüber hinaus kann eine Explorationsfahrt für den kompletten Arbeitsbereich umgesetzt werden, indem Sichten auch auf der Arbeitsfläche bestimmt werden. Den Kanten am Ende der Rekonstruktion liegt ein Beobachtungssprung zu Grunde. Allerdings sollte die Blickrichtung auf die Kante gerichtet werden, um von der bereits rekonstruierten Welt weg zu schauen. Weiterhin kann untersucht werden, ob das Betrachten der Szene während einer Transferbewegung des Roboters möglich ist. Zu beachten sind dabei die Synchronisierung zwischen Robotersteuerung, um die aktuelle Pose des Sensors zu bestimmen, und mögliche Anfälligkeiten des Sensors aufgrund der Bewegung.

Kapitel 5

Dynamische Szenen während Active Vision

Inhalt

Während Änderungen in der Kamerapose im Zuge von *Active Vision* auftreten, sind zusätzlich Änderungen an der Szene beispielsweise durch einen Ko-Arbeiter möglich. Diese Änderungen müssen erkannt und in die globale Umweltrepräsentation eingearbeitet werden, damit diese die Umwelt stets korrekt wiedergibt. Ausgehend von der Klassifikation aller eventuell auftretenden Änderungen in der Szene können diese durch den Vergleich vom aktuellen Sensoreindruck mit der bisherigen Umweltrepräsentation erkannt werden. Für jede erkannte Änderung wird eine Methode beschrieben, wie diese eingearbeitet werden kann. Dabei wird berücksichtigt, ob dadurch bereits erkannte Hypothesen betroffen sind (Abschnitt 5.2). Die Korrektheit der Einarbeitung und der Nutzen werden in der Evaluation diskutiert (Abschnitt 5.3). Teile dieses Kapitel wurden in [Rohner22] bereits veröffentlicht.

5.1 Stand der Forschung

Sich verändernde beziehungsweise dynamische Szenen sind eine Herausforderung in unterschiedlichen Disziplinen wie beispielsweise dem autonomen Fahren, dem eigentlichen Computersehen oder der Robotik [Radke05]. Wie bisher in dieser Arbeit aufgezeigt, ist ein Aspekt der Umweltwahrnehmung dabei die interne Repräsentation der Daten, beispielsweise Punktwolken aus LIDAR-Sensoren [Postica16] und Tiefenkameras [Litomisky13, Newcombe15], Ergebnisse einer semantischen Segmentierung [Langer20] oder Bounding Boxen [Mei20]. Analog zur Active Vision bringt die Repräsentationsform unterschiedliche Anforderungen mit. Im Fall von einzelnen Punktdaten muss jeder Punkt betrachtet oder vorab gruppiert werden, zum Beispiel durch ein Clustering, eine Segmentierung oder lokale Deskriptoren [Drews13a, Drews13b]. Falls bereits Bounding Boxen vorliegen, kann die Veränderung anhand der Erkennungen verarbeitet werden.
Die Erkennung von Änderungen entspringt dabei unterschiedlichen Anwendungen. Eine Möglichkeit sind *Drohnen* (*Unmanned Aerial Vehicles*), die Punktwolken von Bauplätzen nach Änderungen untersuchen [Huang22], oder andere Flugobjekte mit einem Laserscanner [Hebel13, Xiao15]. Ähnlich kann die Änderung von Landnutzung und Bedeckung in der Fernerkundung untersucht werden (sowohl auf 2D- als auch 3D-Daten) [Cheng24]. Eine ähnliche Anwendung ist die Untersuchung von Punktwolken im Tagebau [Yang19]. Eine weitere Anwendung sind autonome Roboter, die im Rahmen des *Simultaneous Location and Mapping* (siehe dazu auch Kapitel 2) eine Karte von der Umwelt aufbauen und Änderungen an der Welt erkennen und einarbeiten müssen [Derner21]. Dabei können erkannten Objekten Eigenschaften zugewiesen werden, inwieweit sich diese bewegen [Bore18, Kunze18].
Eine Möglichkeit, Änderungen zu verarbeiten, ist das Verwerfen von Information, welche seit einem gewissen Zeitraum nicht mehr validiert wurde [Izadi11, Monica16, Riedelbauch17]. Dabei werden Elemente der Umweltrepräsentation gemäß der Wahrscheinlichkeit bewertet, ob diese sich noch unverändert an ihrer Position befinden. Im Rahmen der Mensch-Roboter-Kollaboration kann dies durch die Anwesenheit und Nähe eines Menschen gesteuert werden. Alternativ kann ein Vergleich zwischen der bisherigen Umweltrepräsentation und der aktuellen Sicht berechnet werden, um daraus Veränderungen zu erkennen (beispielsweise durch eine Vorder-Hintergrundtrennung, in der bewegte Objekte erkannt werden [Sengar19]).
Ein Spezialfall ist das *Tracking* von Objekten, siehe beispielsweise den Überblick in [Yilmaz06, Fang19, Abdelaziz24, Zhang25]. Dabei sollen die Objekte über einen längeren Zeitraum beobachtet und Änderungen direkt verarbeitet werden. Für Tracking-Ansätze ist es im Allgemeinen aber wichtig, dass das entsprechende Objekt regelmäßig identifiziert wird. Eine Abwandlung davon ist *visual servoing*, bei dem der Sensor das zu untersuchende Objekt aktiv verfolgt. Siehe dazu auch den Überblick in [Kragic02, Li21, Wu22].
In den vorgestellten Methoden muss die Erkennung von Änderungen unterschiedlich genau sein. Dabei reicht eine binäre Klassifikation, dass sich etwas geändert hat, teilweise bereits aus. Ein Objekt, das sich um wenige Zentimeter bewegt, ist bei Pfadplanung mit Sicherheitsabstand nicht relevant; für das Greifen von Objekten jedoch verhindert diese Änderung eine erfolgreiche Ausführung.
Analog zu den bisherigen Kapiteln wird hier ein modellbasierter Ansatz verfolgt. Konkret für

B-Reps soll der Prozess untersucht werden, welche Änderungen möglich sind und wie diese erkannt und verarbeitet werden können. Ansätze, die automatisch Information entfernen - aufgrund der Unsicherheit der Objekte beziehungsweise der Repräsentation - könnten dem Ansatz hinzugefügt werden. An dieser Stelle ist aber vor allem von Interesse, inwieweit die gegebene Information von B-Reps ausreichend ist, um die Umweltrepräsentation in einem gültigen Zustand zu halten. Tracking-Ansätze sind im Rahmen dieser Arbeit keine Option, da der Sensor nicht nur für die ausschließliche Wahrnehmung einzelner Objekte genutzt werden kann. Zudem ist in Domänen wie Haushalten oder Unternehmen kein kontinuierlicher Blick auf die relevanten Objekte möglich, da der Roboter andere Aufgaben zu erfüllen hat.

5.2 Beschreibung des Verfahrens

In diesem Kapitel wird die Forschungsfrage F3 diskutiert und beantwortet:

F3 Inwieweit können Änderungen an der Szene zwischen zwei Aufnahmen erkannt und verarbeitet werden?

Beide Teilfragen werden im Hinblick auf B-Reps als die zugrundeliegende Datenstruktur beantwortet. Wie im Stand der Forschung erläutert, spielt die Art der Repräsentation eine entscheidende Rolle. Im Fall von B-Reps ist bekannt, welche Flächen existieren (und damit auch die zugehörigen Strukturen wie Knoten und Kanten). Anhand der Kanten kann weiterhin die Nachbarschaft von Flächen bestimmt werden. Im weiteren Verlauf wird zunächst das allgemeine Konzept beschrieben. Die folgenden Abschnitte erläutern die einzelnen Schritte, die Klassifikation, Erkennung und Verarbeitung von Änderungen umfassen.

5.2.1 Ansatz

Der bisherige entwickelte Prozess ist ein iterativer Prozess, der nach jedem Verarbeitungsschritt sicherstellt, dass ein gültiges Weltmodell vorliegt. Dieser Ansatz muss somit auch für dynamische Szenen übertragen werden, da die Umweltrepräsentation und die erkannten Objekte stets verfügbar sein sollen. Nach jedem Schritt im Active Vision Ansatz liegt ein neues partielles B-Rep vor, welches in das bestehende Weltmodell integriert werden muss. An dieser Stelle können Änderungen erkannt werden, die beispielsweise während der Fahrt des Roboters entstanden sind. Der Fokus liegt dabei auf den Flächen, die in der Szene rekonstruiert werden. Zum einen, da Flächen alle Informationen darstellen, die im B-Rep hinterlegt sind. Solange die gesamte Fläche richtig rekonstruiert ist, sind auch alle dazugehörigen Kanten und Knoten korrekt. Zum anderen sind Flächen eine robuste Repräsentation der dreidimensionalen Daten, da diese sich aus einer großen Anzahl Punkte zusammensetzen und so Rauschen durch Mittelung entfernt wird [Sand19]. Abschließend sind Flächen für den vorgestellten Objektwiedererkennungsansatz entscheidend: sowohl für die Berechnung der Hypothesen als auch für den Active Vision Prozess. Aus diesen Gründen liegt der Fokus beim Umgang mit dynamischen Szenen auf den Flächen.
Als erster Schritt für den Umgang mit veränderlichen Szenen ist eine Klassifikation aller möglichen Veränderungen notwendig. Dies erfolgt über einen Vergleich der bisherigen Umwelt-

repräsentation mit dem B-Rep aus der aktuellen Sicht, ob eine Fläche in der bisherigen Rekonstruktion vorhanden ist oder nicht. Für jeden möglichen Fall wird anschließend eine Möglichkeit diskutiert, wie diese Änderung entdeckt werden kann. Dazu wird berechnet, welche Flächen aus der aktuellen Sicht sichtbar sein sollten und welche in der Umweltrepräsentation bereits vorhanden sind. Sobald alle möglichen Änderungen bestimmt und erkannt wurden, erfolgt die Behandlung und das Berechnen der aktualisierten Umweltrepräsentation. Da Objekthypothesen bekannt sind (diese sind auch für den Active Vision Prozess notwendig), erfolgt die Behandlung der Änderungen auf zwei Arten: Wenn Objekthypothesen vorliegen, verfügt man über Informationen, welche Flächen in der Szene zusammenhängen. Diese Information wird genutzt, um die Umweltrepräsentation konsistent zu halten. Falls partiell keine Hypothesen vorliegen, wird die Änderung an dieser Stelle ausschließlich über die geometrische Information der B-Reps verarbeitet.

5.2.2 Typen von Änderungen

Ausgehend von Flächen als relevante Information kann die Änderung zwischen einer Sicht und dem bestehenden Weltmodell beschrieben werden. Dabei wird untersucht, ob eine Fläche im Weltmodell bereits vorhanden war oder nicht und ob sie in der aktuellen Sicht rekonstruiert wurde. Weiterhin ist von Interesse, ob die Fläche überhaupt gesehen werden konnte. Da zudem keine *Aging* Methoden verwendet werden, ist die Aussage, ob eine Fläche existiert, binär. Weiterhin ist zu beachten, ob eine Fläche aus der aktuellen Sicht wahrgenommen werden kann. Alle Flächen, die außerhalb des Frustums liegen, werden ignoriert, da zu diesen keine Aussage gemacht werden kann. Insgesamt ergeben sich die folgenden Fälle:

1. Die Fläche ist in der bisherigen Umweltrepräsentation nicht vorhanden ...
   (a) aber sichtbar im aktuellen B-Rep: *Hinzugefügte* Fläche.
   (b) und auch nicht sichtbar im aktuellen B-Rep: Da diese Fläche nicht existiert, wird diese Möglichkeit nicht benannt.
2. Die Fläche ist in der bisherigen Umweltrepräsentation vorhanden ...
   (a) und sichtbar im aktuellen B-Rep: *Validierte* Fläche.
   (b) aber nicht sichtbar im aktuellen B-Rep und
      i. sollte aber sichtbar sein: *Entfernte* Fläche.
      ii. ist nicht sichtbar aufgrund von Verdeckungen oder Kameraeinschränkungen: *Verdeckte* Fläche.

Der Fall, dass eine Fläche *hinzugefügt* wurde, ist der einfachste und wird so auch in der Vorgängerarbeit [Sand19] betrachtet. Dieser Fall tritt vor allem auch dann auf, wenn zwischen der alten Umweltrepräsentation und der neuen Sicht keine Überschneidung existiert. Wenn die Fläche weder in der bisherigen Repräsentation noch in der aktuellen Sicht vorliegt, muss dieser Fall nicht gesondert behandelt werden, da diese Fläche insgesamt nicht existiert und keine Information über die vorliegende Szene beinhaltet.

5.2.3 Erkennen von Änderungen

Um alle Änderungen zwischen zwei B-Reps zu erkennen, wird für jeden möglichen Fall ein Ansatz beschrieben. Dazu wird zunächst bestimmt, dass sich etwas geändert hat, indem die bisherige Umweltrepräsentation W mit der Rekonstruktion C aus der aktuellen Sicht verglichen wird. Die entdeckten Änderungen werden anschließend klassifiziert.
Als Hilfestellung wird eine Projektion von B-Reps auf eine 2D-Ebene benötigt. Mit Hilfe dieser Projektion wird untersucht, welche Flächen aus der aktuellen Sicht sichtbar sein können. Für diese Projektion wird die Pose der aktuellen Sicht $T_C \in SE(3)$ als Starrkörpertransformation benötigt. Für die Projektion der bisherigen Umweltrepräsentation ist zu beachten, dass diese Flächen enthält, die von der aktuellen Pose aus nicht rekonstruiert werden können aufgrund des Frustums der Tiefenkamera oder sonstiger physikalischer Einschränkungen. Das Frustum ist dabei über die Kameraparameter bekannt und kann direkt auf das globale B-Rep angewandt werden, indem dieses gemäß dem Frustum zurechtgeschnitten wird. Physikalische Limitierungen sind weiterhin ein ungeeigneter Sichtwinkel, der bei Tiefenkameras verhindert, dass eine Fläche aufgenommen und somit rekonstruiert werden kann. Alle physikalischen Einschränkungen werden dabei im Tupel L gesammelt. Somit wird eine Projektion `projected` zu einer Pose T für ein B-Rep B beschrieben als

$$\texttt{projected}(B, T, L) \rightarrow {F_B}^{n \times m} \tag{5.1}$$

Jeder Pixel im resultierenden $n \times m$ 2D-Bild beinhaltet somit die Information, welche Fläche am weitesten vorne zur Pose T ist. Die Größe des Bildes muss dabei ausreichend groß sein, damit bei der Projektion keine geometrische Information verloren geht, und kann aus der Strukturgröße abgeleitet werden. Alternativ ist die Auflösung des 3D-Sensors ein geeigneter Wert. Somit bestimmt sich die Projektion für die Umweltrepräsentation W als

$$P_W = \texttt{projected}(W, T_C, L) \tag{5.2}$$

Bei der Projektion von 3D-Objekten auf eine 2D-Ebene handelt es sich um ein klassisches Problem der Computergrafik, was von gängiger Grafikkarten Soft- und Hardware unterstützt wird und daher auf dieser ausgeführt wird. Dazu wird das B-Rep zunächst trianguliert, was aufgrund von ausschließlich planaren Flächen direkt möglich ist. Für jedes Dreieck kann die ursprüngliche Fläche beispielsweise anhand einer gesetzten Textur identifiziert werden. Die Triangulierung wird dann auf die 2D-Ebene mit Ausmaßen $n \times m$ gerendert beziehungsweise geshadert. Für jede Fläche im ursprünglichen B-Rep kann dann anhand der gerenderten Pixel entschieden werden, ob diese Fläche sichtbar ist oder nicht.
Dieser Prozess kann analog für das B-Rep $C = (F_C, E_C, V_C, B_C)$ der aktuellen Sicht angewandt werden. Da beide Projektionen miteinander verglichen werden sollen, ist es notwendig, dass die Projektion auf die Bildebene aus derselben aktuellen Sicht erfolgt (mit identischer Pose). Somit wird sowohl für die Projektion des aktuellen lokalen B-Reps als auch der gesamten Umweltrepräsentation die Pose T_C verwendet.

$$P_C = \texttt{projected}(C, T_C, L) \tag{5.3}$$

Somit liegen für den aktuellen Zeitschritt zwei Projektionen vor: zum einen die aktuell rekonstruierte Sicht, zum anderen das bestehende Weltmodell. Aufgrund der physikalischen Einschränkungen L umfassen beide das gleiche Sichtfeld. Daneben ist wichtig, dass auch durch die Projektion weiterhin bekannt ist, welche Fläche im Dreidimensionalen mit welcher Fläche in der Projektion korrespondiert. Somit lassen sich alle Operationen, die anhand der Projektionen bestimmt werden, in der dreidimensionalen B-Rep Repräsentation ausführen. Um die Änderungen zwischen der aktuellen Sicht und der Umweltrepräsentationen zu bestimmen, werden die Projektionen miteinander verglichen: Für jede Fläche in einer Projektion wird betrachtet, ob diese mit einer Fläche in der anderen Projektion korrespondiert. So lässt sich für eine beliebige Fläche $f \in F_W$ bestimmen, ob diese zu einer gegebenen Pose T bei Anwesenheit eines B-Reps C unter den physikalischen Einschränkungen L sichtbar (1) ist oder nicht (0),

$$\texttt{isvisible}(f, C, T, L) \rightarrow \{0, 1\} \tag{5.4}$$

Dazu kann die Korrespondenz zwischen Flächen im Dreidimensionalen über die Funktion $\texttt{explainedby}(f, g) \rightarrow \{0, 1\}$ bestimmt werden, basierend auf der Position, dem Normalenvektor und der Größe der zwei Flächen f und g (identisch zu der Definition aus dem vorherigen Kapitel, Abschnitt 4.2.3). Als Ergebnis erhält man, ob diese zwei Flächen korrespondieren (1) oder nicht (0).
Mit diesen Definitionen wird für jede Kategorie an möglicher Änderung die Menge an Flächen beschrieben, die dieser Kategorie angehören: Menge A (*added*) für hinzugefügt, Menge V (*validated*) für validiert, Menge R (*removed*) für entfernt und Menge O (*occluded*) für verdeckt. Als erster Fall gilt für *hinzugefügt*, dass die Fläche bisher nicht vorhanden war, in der aktuellen Sicht aber existiert. Bezüglich der Flächenkorrespondenzen bedeutetet dies, dass die zu untersuchende Fläche keine Korrespondenz zu einer Fläche in der bisherigen Umweltrepräsentation aufweist.

$$A = \{g \in F_C | \nexists f \in F_W : \texttt{explainedby}(g, f)\} \tag{5.5}$$

Analog gilt für den Fall *validiert*, dass eine Korrespondenz vorliegt.

$$V = \{f \in F_W | \exists g \in F_c : \texttt{explainedby}(f, g)\} \tag{5.6}$$

Abschließend verbleiben die beiden Fälle, wenn eine Fläche in der aktuellen Rekonstruktion nicht sichtbar ist. Diese sind aufgeteilt in *entfernt* und *verdeckt*. Falls eine Fläche in der aktuellen Rekonstruktion nicht sichtbar ist, aber sichtbar sein sollte, ist diese Fläche *entfernt* worden.

$$R = \{f \in F_w | \nexists g \in F_C : \texttt{explainedby}(f, g) \wedge \texttt{isvisible}(f, C, T_C, L)\} \tag{5.7}$$

Wieder gilt hier analog: Wenn die Fläche in der aktuellen Aufnahme nicht sichtbar ist, in der bisherigen Welt aber wahrnehmbar sein sollte, ist diese nur *verdeckt*

$$O = \{f \in F_W | \nexists g \in F_C : \texttt{explainedby}(f, g) \wedge \neg\texttt{isvisible}(f, C, T_C, L)\} \tag{5.8}$$

5.2.4 Behandlung von Änderungen

Ausgehend von diesen Mengen an Flächen ist der nächste Schritt die korrekte Verarbeitung. Das beutetet in diesem Zusammenhang, dass am Ende der Verarbeitung immer noch ein valides B-Rep vorliegt, um den inkrementellen Ansatz der Umweltrepräsentation zu erhalten. Als zweites, nachgestelltes Ziel soll möglichst viel Information im B-Rep erhalten bleiben. Dabei ist zu beachten, dass Flächen in der Rekonstruktion nicht für sich allein existieren können, da diese - gemäß den getroffenen Annahmen - immer an ein Objekt gebunden sind. Somit ist das Ziel, ein Objekt vollständig zu entfernen, auch wenn nur eine einzelne Fläche des Objekts entfernt wird. Falls ein Konflikt vorliegt - beispielsweise wenn eine Fläche validiert, eine andere aber entfernt ist - wird trotzdem das ganze Objekt aus der Umweltrepräsentation gelöscht. Das ist notwendig, da aufgrund der rein geometrischen Betrachtung nicht sichergestellt ist, dass eine validierte Fläche tatsächlich vom identischen Objekt stammt.
Die Menge an validierten Flächen V kann dabei direkt in die bestehende Umweltrepräsentation fusioniert werden (wie in der Vorarbeit [Sand19] beschrieben). Insgesamt haben diese Flächen einen geringen Einfluss auf das gesamte B-Rep (da bereits eine Fläche in der Welt existiert die mit der aktuellen Fläche korrespondiert) und benötigen keine weiteren Schritte, sodass das Modell weiterhin valide ist. Auf gleiche Art kann die Menge der hinzugefügten Flächen A in die Umweltrepräsentation eingefügt werden.
Als weitere Menge sind die Flächen R zu behandeln, die aus der Umweltrepräsentation entfernt werden müssen. Dafür können die bestehenden Objekthypothesen H genutzt werden. Dazu wird unterschieden, ob eine zu löschende Fläche von einer Hypothese erklärt wird oder nicht. Falls eine Hypothese vorliegt (durch Prüfen aller Hypothesenflächen auf Korrespondenz mit der gerade untersuchten Fläche), werden alle weiteren Flächen in der Welt bestimmt, die zu einer Fläche dieser Hypothese korrespondieren. Da nach den gegebenen Annahmen von einem realen Objekt nicht einzelne Flächen aus der Welt entfernt werden können, sondern nur das komplette Objekt entfernt werden kann, müssen auch alle weiteren Flächen gelöscht werden. Dazu werden zunächst alle Flächen der Hypothesen gesucht, die mit einer zu löschenden Fläche korrespondieren

$$D_{H_0} = \cup_{\{h_i \in H \mid \exists r \in R : r \in \texttt{correspondingfaces}(h_i)\}} \texttt{correspondingfaces}(h_i) \tag{5.9}$$

Die Funktion $\texttt{correspondingfaces}(h_i)$ gibt dabei alle Flächen des globalen B-Reps W zurück, die mit einer Fläche vom Model der Hypothese h_i korrespondieren.
Neben allen Flächen der Hypothese müssen darüber hinaus alle benachbarten Flächen entfernt werden. Dieser Schritt ist notwendig, damit die Umweltrepräsentation als B-Rep valide bleibt. Das Problem tritt dann auf, wenn die zu löschende Fläche eine oder mehrere Kanten in der resultierenden Umweltrepräsentation hinterlässt. Durch das Hinzufügen eines neuen Objekts an dieser Stelle müsste für eine valide Repräsentation eine Kante eingefügt werden. Das Hinzufügen dieser neuen Kante wäre dort nicht realisierbar, da eine Topologieänderung durch sich schneidende Kanten nicht möglich ist [Sand19]. Somit bestimmt sich die Menge aller Flächen, die entfernt werden müssen, als

$$D_{H_1} = D_{H_0} \cup \{f \in F_W \mid \exists g \in D_{H_0} : \texttt{neighbor}(f, g)\} \tag{5.10}$$

wobei die Funktion `neighbor` angibt, ob zwei beliebige Flächen f und g benachbart sind. Diese Relation lässt sich mit Hilfe der B-Rep Datenstruktur anhand der Zwillings-Halbkanten bestimmen [Sand19].
Falls keine Hypothese für eine zu löschende Fläche aus R verfügbar ist, ist das Ziel, vollständige Objekte zu löschen. Da allerdings nicht direkt bekannt ist, wieviele Flächen entfernt werden müssen, damit ein Objekt gelöscht wird (und eine Rekonstruktion eines Objekts durch Verdeckungen in mehrere Flächen zerfallen kann), ist es erforderlich, mehr Flächen zu löschen (siehe zum Beispiel Testobjekt *crane* aus Kapitel 3). Dazu werden ausgehend von der zu löschenden Fläche alle benachbarten Flächen ebenfalls entfernt. Dieser Vorgang wird transitiv wiederholt, bis als benachbarte Fläche die bekannte Arbeitsoberfläche erreicht wird. Somit bestimmt sich zunächst die Menge an zu löschenden Flächen, die keine Hypothese besitzen, als

$$D_{\bar{H}_0} = \{r \in R | \nexists h_i \in H : r \in \texttt{correspondingfaces}(h_i)\} = R \backslash D_{H_0} \tag{5.11}$$

Diesen werden nun in jedem Schritt j die weiteren benachbarten Flächen hinzugefügt.

$$D_{\bar{H}_j} = D_{\bar{H}_{j-1}} \cup \{f \in F_W | \exists g \in D_{\bar{H}_{j-1}} : \texttt{neighbor}(f, g)\} \tag{5.12}$$

Da die Position der Arbeitsfläche vorab bekannt ist, wird diese aus der Umweltrepräsentation W entfernt, wodurch die rekursive Berechnung gesichert terminiert, da die Anzahl an benachbarten Flächen endlich ist. Als Rekursionsanker sind die Flächen der Menge R definiert (*entfernt*).
Bezüglich der verdeckten Flächen sind mehrere Ansätze möglich. Zum einen können diese Flächen analog behandelt werden wie Flächen, die entfernt wurden. Dies hat zur Folge, dass Flächen, die in der Welt noch existieren, entfernt werden und zu einem späteren Zeitpunkt wieder erfasst werden müssen. Alternativ können verdeckte Flächen direkt in der Welt verbleiben. In dieser Arbeit werden verdeckte Flächen beibehalten, um die Einschränkungen bezüglich zu löschender Flächen soweit wie möglich zu vermeiden.

5.3 Experimentelle Auswertung

In diesem Kapitel wird das untersuchte Konzept evaluiert. Zunächst werden der experimentelle Aufbau und anschließend die Ergebnisse erläutert. Wie der Rest des Kapitels basieren Teile der Experimente auf [Rohner22].

5.3.1 Aufbau

Die Experimente für dynamische Szenen setzen sich aus zwei Teilen zusammen: zum einen aus der Validierung, indem jeder mögliche Fall getestet und untersucht wird, zum anderen aus dem Vergleich zwischen dynamischer und statischer Rekonstruktion ganzer Szenen.

Validierung

Zunächst werden die vier vorgestellten Fälle von Änderungen untersucht: *Hinzugefügt*, *Validiert*, *Entfernt* und *Verdeckt*. Dazu wird jeweils eine Rekonstruktion gezeigt, die Änderung in der Szene durchgeführt und die anschließende Rekonstruktion untersucht. Zudem wird an einem Beispiel gezeigt, wie nur kurz auftretende Störsignale auf diese Weise aus einer Szene entfernt werden können. Gründe für Störsignale sind beispielsweise menschliche Ko-Arbeiter, die mit der Szene interagieren, während der Roboter diese rekonstruiert.

Auswertung ganzer Szenen

Analog zu den bisherigen Experimenten wird das Konzept anhand ganzer Szenen evaluiert. Dazu wird zunächst eine Szene rekonstruiert und zwischen den Sichten verändert, sodass die unterschiedlichen Änderungstypen alle enthalten sind. Dies wird solange fortgeführt, bis die Rekonstruktion und Wiedererkennung vollständig ist (*dynamische* Rekonstruktion). Die abschließend verbleibende Szene wird anschließend vom existierenden Active Vision Prozess rekonstruiert (*statische* Rekonstruktion). Um sicherzustellen das die identische Szene rekonstruiert wird, ist die statische Rekonstruktion erst dann durchzuführen, sobald alle Aufnahmen der dynamischen Rekonstruktion erfolgt sind. Abschließend kann die dynamische Rekonstruktion mit der statischen vergleichen werden. Als Evaluationskriterium wird die korrekte Anzahl an erkannten Objekten verwendet sowie überprüft, ob zu viele oder zu wenige Flächen in der dynamischen Rekonstruktion vorliegen.

5.3.2 Ergebnisse

In diesem Abschnitt werden die Ergebnisse der Validierung und Evaluation dargestellt und diskutiert.

Validierung

Als einfachster Fall der Veränderungen in der Szene ist das *Hinzufügen*, zu sehen in Abbildung 5.1. Dabei wird auf eine zunächst leere Arbeitsoberfläche ein Objekt platziert und in die Umweltrepräsentation eingefügt.
Als nächste Möglichkeit wird *Entfernt* untersucht, siehe Abbildung 5.2. Im Rahmen der Validierung wurde dabei der Roboter an der gleichen Stelle belassen. Das Objekt in der Szene ist in der zweiten Sicht nicht mehr wahrzunehmen und auch nicht durch eine andere Rekonstruktion verdeckt. Daher werden die korrespondierenden Flächen aus der Umweltrepräsentation entfernt. Abschließend wird der aktuelle Sensoreindruck in die Umweltrepräsentation eingefügt, wodurch die Arbeitsoberfläche vervollständigt und die Lücke geschlossen wird.
Abschließend verbleiben die Fälle *Verdeckt* und *Validiert*. Zunächst stehen zwei Objekte in der Szene, von denen eines direkt erkannt wird. Während sich der Roboter zu der zweiten Sicht bewegt, um das andere Objekt weiter zu rekonstruieren, wird der Dodekaeder aus der Szene entfernt. In der Rekonstruktion verweilt er aber, da er aus der aktuellen Sicht aufgrund der Verdeckung des zweiten Objekts nicht wahrnehmbar ist. Mit der dritten Sicht von oben wäre

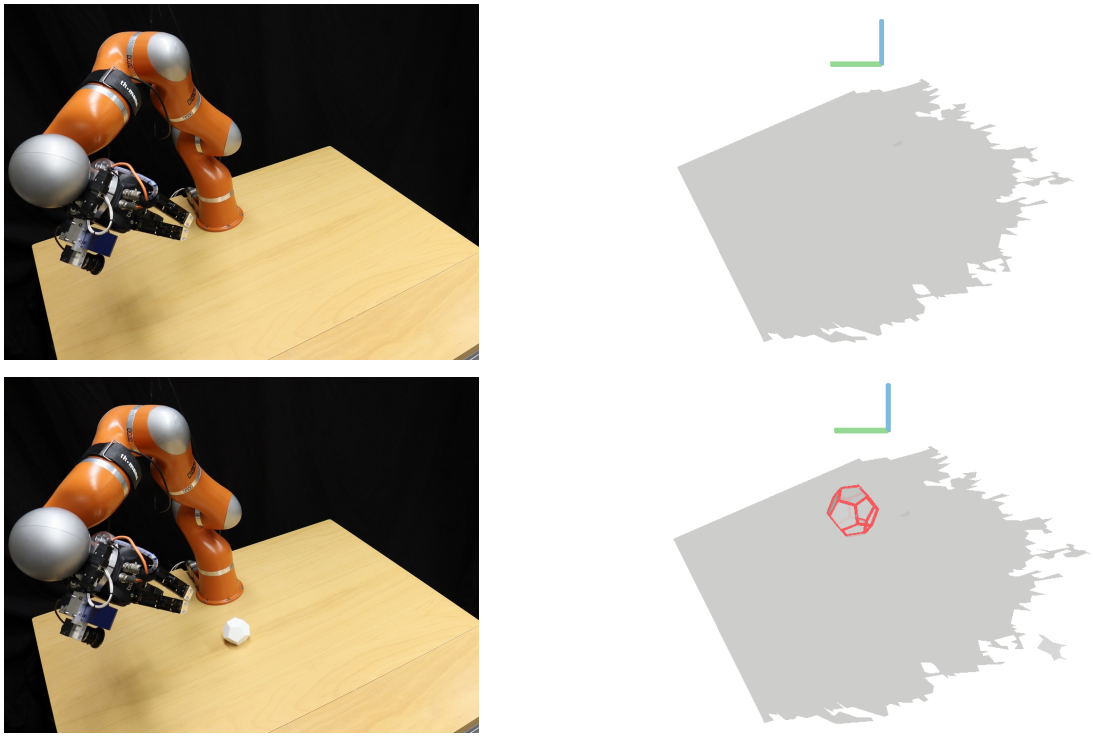


Abbildung 5.1: Validierung der Änderung *Hinzufügen*: In der ersten Reihe die initiale Szene und Rekonstruktion, in der zweiten Reihe nach der Änderung. Zu jeder Rekonstruktion ist zusätzlich das Weltkoordinatensystem in der Roboterbasis dargestellt.

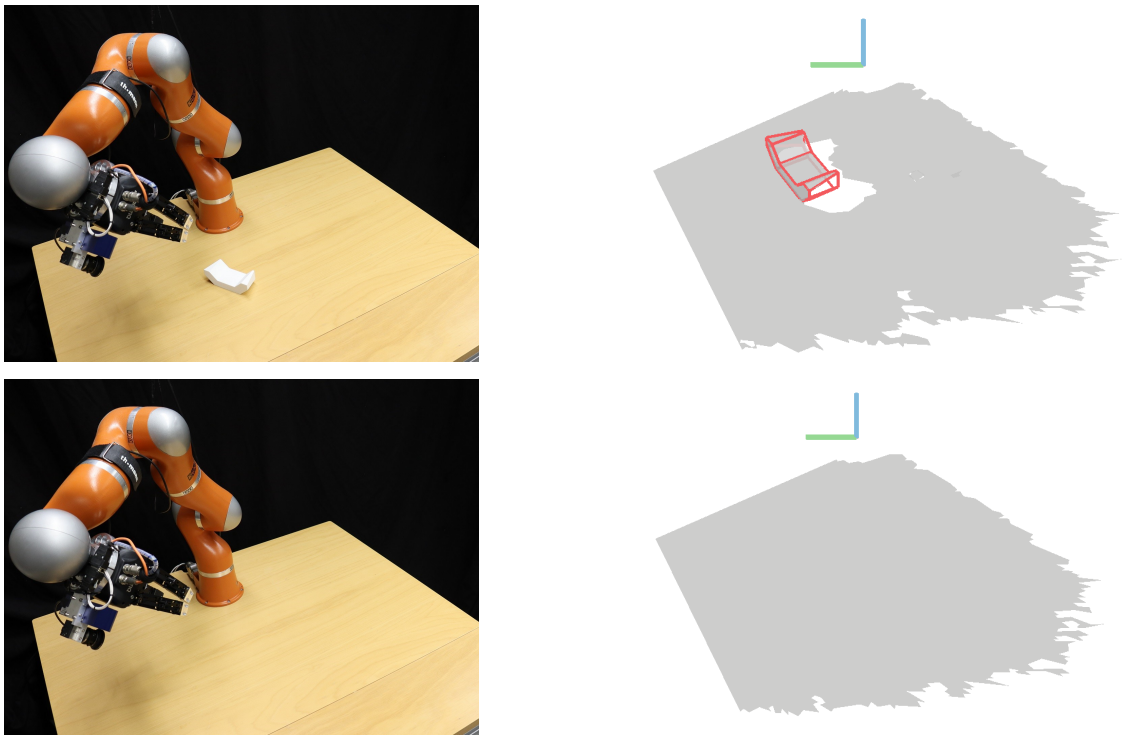


Abbildung 5.2: Validierung der Änderung *Entfernen*: In der ersten Reihe die initiale Szene und Rekonstruktion, in der zweiten Reihe nach der Änderung

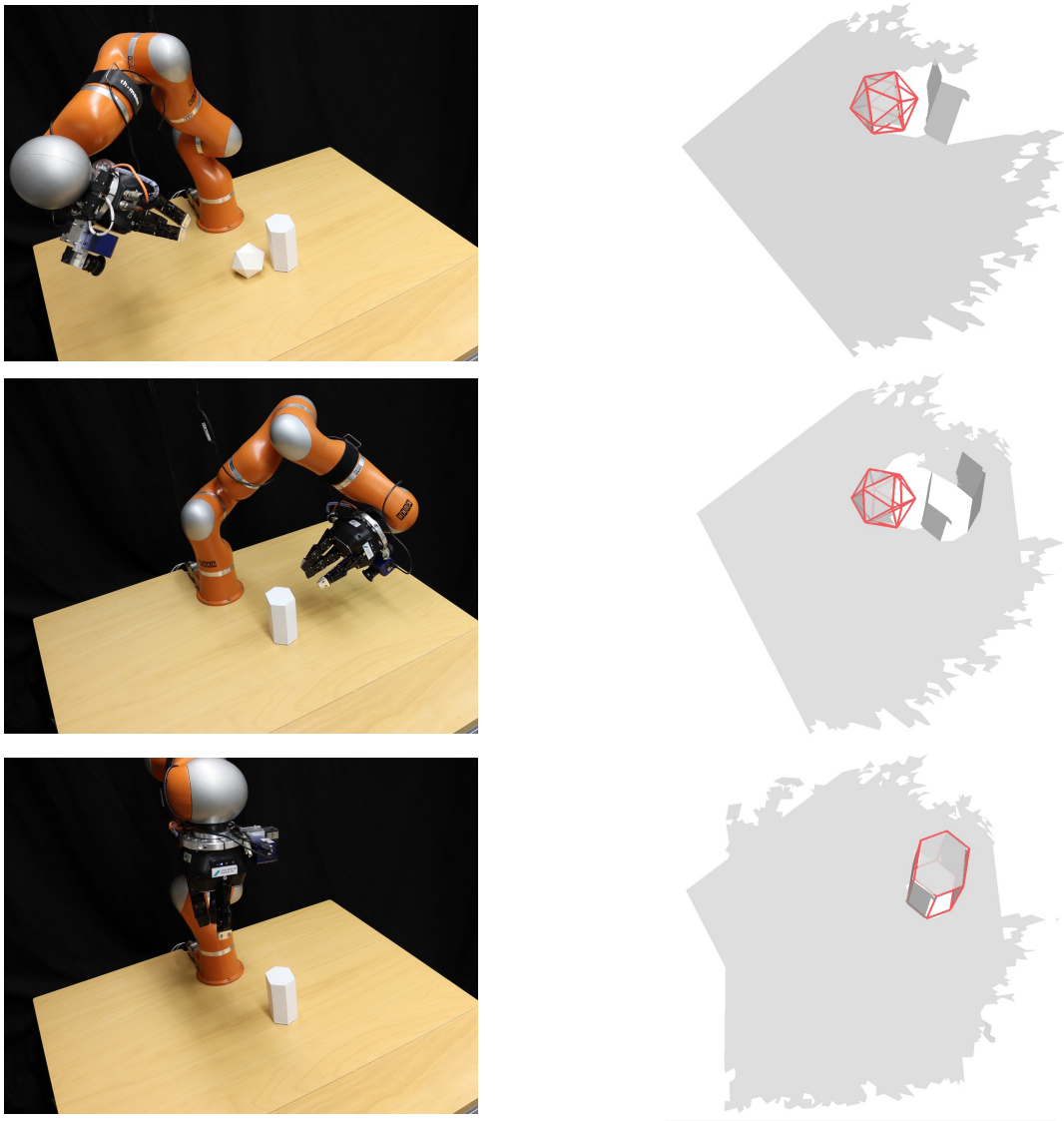


Abbildung 5.3: Validierung der Änderungen *Verdeckt* und *Validiert*: in der ersten Reihe die initiale Szene und Rekonstruktion, in der zweiten Reihe nach der Änderung *Entfernt*. Da das entfernte Objekt aus der aktuellen Pose nicht sichtbar wäre, verbleiben die Flächen in der Rekonstruktion. Aus einer weiteren Perspektive (dritte Reihe) werden die Flächen schließlich entfernt sowie Flächen des verbleibenden Objekts validiert

der Dodekaeder aus der aktuellen Sicht wahrnehmbar. Allerdings korrespondieren keine Flächen in der aktuellen Rekonstruktion mit der Umweltrepräsentation, wodurch diese Flächen gemäß der Hypothese aus der Welt entfernt werden. Mit der dritten Sicht wird zudem das zweite Objekt korrekt erkannt (zu sehen in Abbildung 5.3).
Neben den unterschiedlichen Fällen können Störsignale aus der Umweltrekonstruktion entfernt werden (wie in Abbildung 5.4 zu sehen ist). In der ersten Rekonstruktion ist ein Arm zu sehen, der ebenfalls planar rekonstruiert wird, wodurch eine Vielzahl kleiner Flächen in die Szene eingebaut wird. Im nächsten Schritt wird aus der gleichen Pose erneut ein B-Rep aufgenommen und gemäß der Methodik in die Umweltrepräsentation eingearbeitet. Die Flächen, die zum Arm gehören, sollten aus der aktuellen Perspektive sichtbar sein, sind im aktuellen B-Rep allerdings nicht vorhanden. Diese Flächen werden somit gelöscht, und die aktuelle Re-

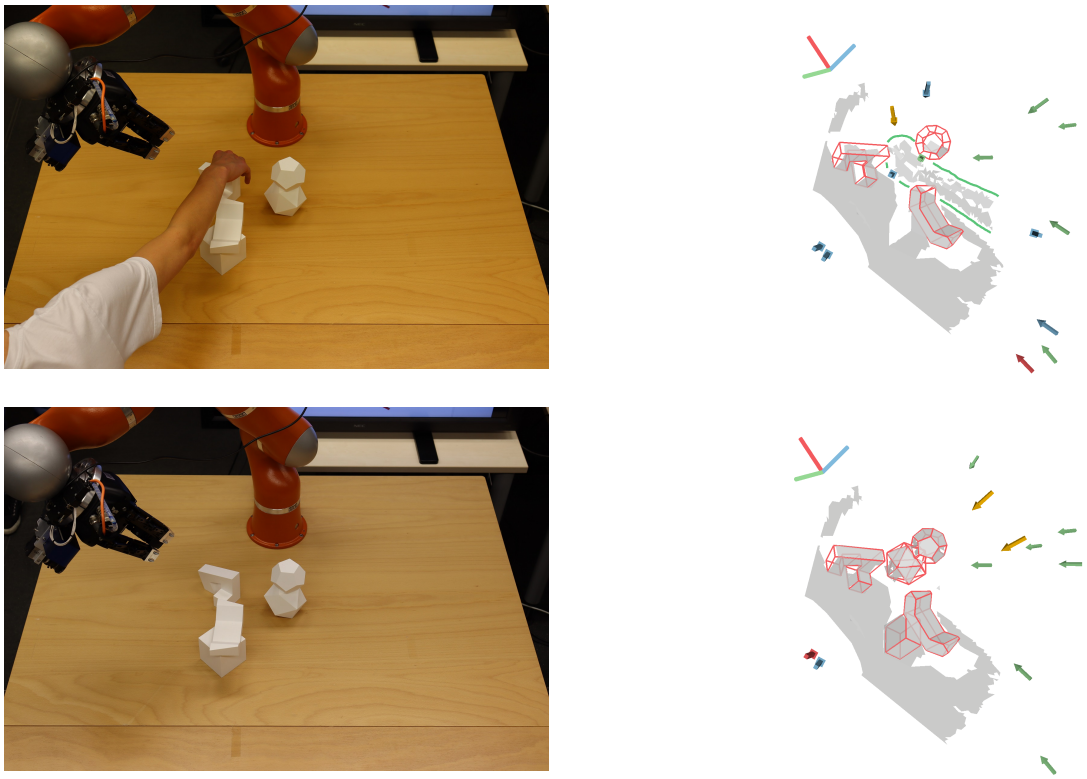


Abbildung 5.4: Validierung des Entfernens von Störsignalen: Ein Arm wird zusätzlich zur Szene rekonstruiert (grüne Umrandung). Nach Entfernen des Arms aus der Szene und einer weiteren Rekonstruktion wurden die überschüssigen Flächen entfernt, aus [Rohner22]

konstruktion wird in die Umwelt eingefügt. Dadurch, dass der Arm einzelne Objekte verdeckt hat, können nun alle Objekte korrekt erkannt werden.

Ganze Szenen

In Abbildung 5.5 ist der Aufbau sowie die Rekonstruktion mit einer kompletten Szene dargestellt. Dabei wurden zwei Objekte hinzugefügt und zwei entfernt. In Abbildung 5.6 finden sich weitere beispielhafte Szenen. Dabei wurden alle Änderungen eingebaut. Falls sich ein Objekt in der Szene bewegt, entspricht dies der Kombination *Entfernen* und *Hinzufügen*, da kein Tracking der Objekte stattfindet. Über alle dynamischen und statischen Rekonstruktionen wurden insgesamt 126 Flächen (dynamisch) und 120 Flächen (statisch) rekonstruiert. Aus den dynamischen Rekonstruktionen hatten 97 Flächen (77%) eine Korrespondenz zur statischen. Für die andere Richtung waren es 98 Flächen (82%). Der Unterschied der eigentlich symmetrischen Beziehung ergibt sich daraus, dass eine Fläche der dynamischen Rekonstruktion durch zwei Flächen der statischen erklärt wurde. Wie in den bisherigen Evaluationen kann dies auftreten, wenn eine Fläche in zwei Bestandteile zerfällt (beispielsweise durch Verdeckung oder Rauschen). Der Grund für die hohe Anzahl an Flächen, die zwischen den Rekonstruktionen keine Korrespondenz haben, liegt an Verdeckungen während der statischen Rekonstruktion. Bei der dynamischen Rekonstruktion sind manche Flächen noch sichtbar, die in der statischen durchgehend verdeckt sind oder aufgrund der Filter der Sichten nicht wahrgenommen

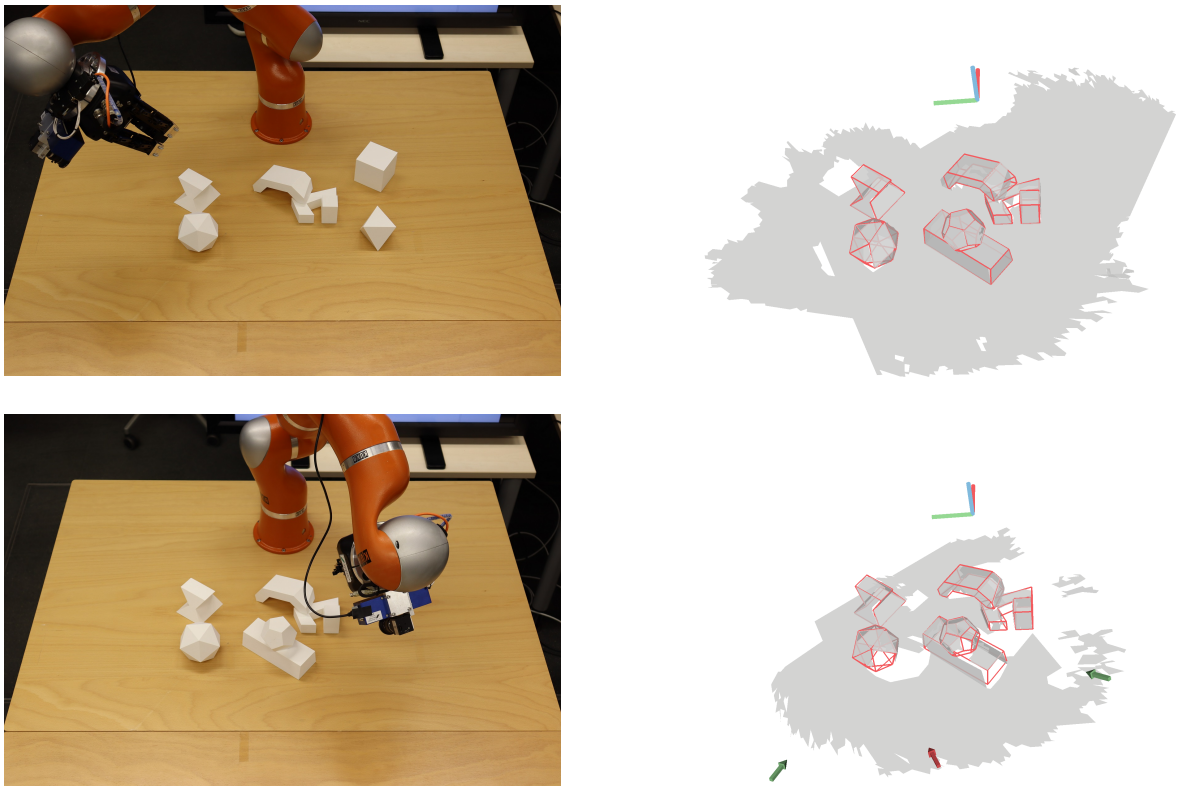
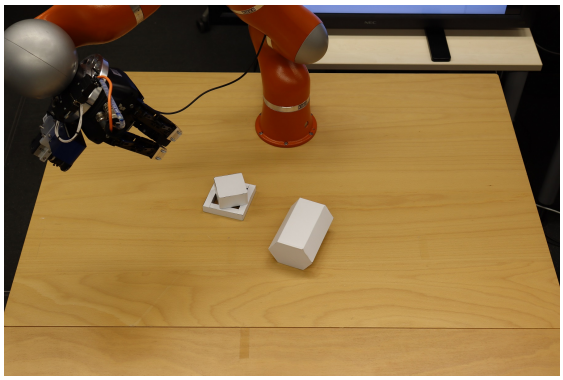
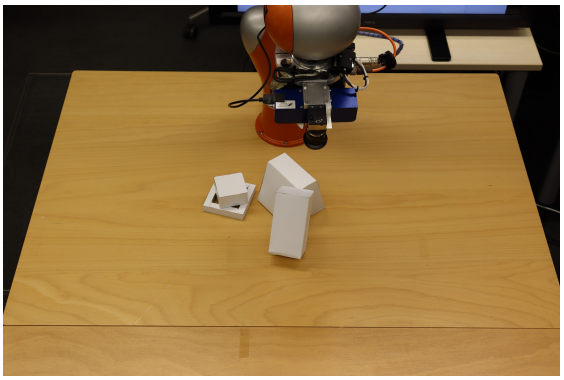


Abbildung 5.5: Vergleich zwischen dynamischer und statischer Rekonstruktion: in der oberen Reihe die ursprüngliche Szene und die finale dynamische Rekonstruktion, in der unteren Reihe die letzte Roboterpose und Rekonstruktion der statischen Szene, aus [Rohner22]

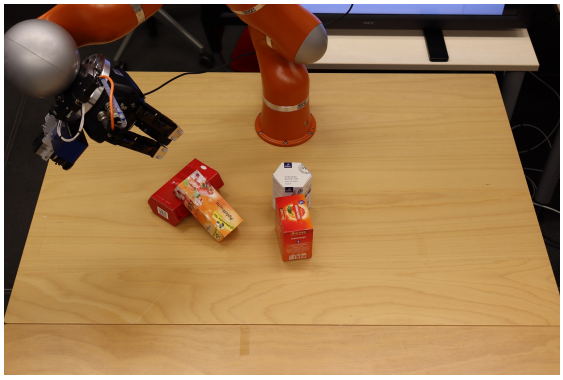
werden. Darüber hinaus sind die angefahrenen Sichten zwischen statischer und dynamischer Rekonstruktion nicht identisch. Dadurch passiert es, dass zusätzliche Flächen nur in einem der beiden Fälle rekonstruiert werden, die für die korrekte Erkennung nicht notwendig sind. Nach dem Entfernen aller Flächen, die zu einer korrekten Hypothese korrespondieren, verbleibt eine Fläche. Das Problem dabei ist, dass zunächst ein Objekt entfernt, an ähnlicher Stelle aber ein anderes Objekt platziert wurde. Dadurch wurde sowohl die Fläche des neuen Objekts eingefügt, die bestehende Fläche aber nicht gelöscht, da zur neuen Fläche eine Korrespondenz bestand. In diesem Fall war sowohl die Normale als auch der überschneidende Flächeninhalt zu groß. Einschränkungen des Sensors beim Betrachten von Flächen mit einem sehr steilen Einfallswinkel erschweren die Rekonstruktion in diesem Fall. Ein direkter Blick kann dieses Problem beheben. Nichtsdestoweniger sind Veränderungen, die sehr ähnliche Flächen erzeugen, als Einschränkungen festzuhalten. In der Rekonstruktion sind diese Flächen mit den hier beschriebenen Methoden nicht unterscheidbar.



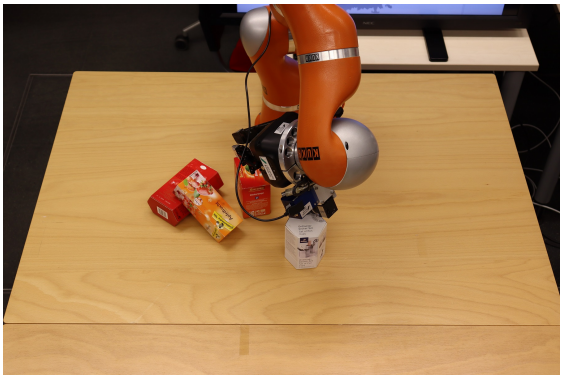
(a) Szene 2, Ausgangslage



(b) Szene 2, Endgültige Szene



(c) Szene 3, Ausgangslage



(d) Szene 3, Endgültige Szene

Abbildung 5.6: Weitere Szenen der Evaluation, jeweils mit der initialen Szene und der endgültigen Objektpositionierung

5.4 Fazit

Dieses Kapitel stellt einen Ansatz vor, wie mit dynamischen Szenen und Änderungen während des Rekonstruktionsprozesses mit Active Vision umgegangen werden kann. Ausgehend von B-Reps werden alle möglichen Fälle von Änderungen abgeleitet. Für jeden dieser Fälle wird anschließend eine Methode beschrieben, den Fall zu erkennen und die Änderungen korrekt einzuarbeiten. Dafür wird die bisherige Umweltrepräsentation mit der aktuellen Sicht verglichen. Je nachdem, ob zu den betroffenen Flächen eine Objekthypothese vorliegt, kann die Änderung präziser behoben werden als bei Flächen ohne dazugehörige Hypothese.
Ausgehend von diesen Ergebnissen kann die Fragestellung

F3 Inwieweit können Änderungen an der Szene zwischen zwei Aufnahmen erkannt und verarbeitet werden?

beantwortet werden. Durch die Verfügbarkeit von vollständigen geometrischen Strukturen in Form der B-Reps sind Analysen der Szene auf diesem Level möglich. Dadurch kann die Information in den meisten Fällen korrekt verarbeitet werden, da es Methoden zum Zusammenführen von mehreren B-Reps gibt. Das Entfernen ist aufgrund der zugrundeliegenden Datenstrukturen ebenfalls möglich. Nichtsdestoweniger schränkt die Wahl der Umweltrepräsentation auf B-Reps die Möglichkeiten ein. Zu kleine Änderungen an der Szene oder zwischen zwei Objekten werden nicht erkannt und damit auch nicht korrekt verarbeitet. Um die

globale Umweltrepräsentation gemäß der Eigenschaften von B-Reps gültig zu halten, muss beim Entfernen von Objekten teilweise mehr entfernt werden, als nach Rekonstruktion der aktuellen Sicht und Einarbeitung in die Umweltrepräsentation notwendig ist. Dieses Problem wird durch die Verwendung von Hypothesen entschärft, allerdings nicht komplett gelöst. Da zudem ausschließlich geometrische Information verwendet wird, können Flächen in der Welt verbleiben, die korrekterweise entfernt sein müssten (beispielsweise wenn ein Objekt in der Szene entfernt und an gleicher Position mit einem ähnlichen Objekt ersetzt wird).
Als Erweiterungen dieser Methoden kann ein zeitbasierter Ansatz hinzugefügt werden (wie im Stand der Forschung dieses Kapitels vorgestellt). Ziel ist dabei, Flächen aus der Umweltrepräsentation zu entfernen, die seit einem bestimmten Zeitraum nicht mehr rekonstruiert wurden, wobei ein Risiko besteht, dass sich an der Szene in diesem Bereich etwas geändert hat. Weiterhin kann näher untersucht werden, welche Flächen gegebenenfalls aus der Umwelt entfernt werden. Aufgrund der vorhandenen Objektdatenbank ist es möglich, zu berechnen, welche Flächen aufgrund eines Objektes zusammenhängen können. Auf diese Weise kann die Anzahl an zu entfernenden Flächen reduziert werden. Abschließend kann die Information, dass es Änderungen an der Szene gab, in den Active Vision Prozess integriert werden. Beispielsweise ist es sinnvoll, bei Flächen, die entfernt wurden, mit einer Validierungssicht zu prüfen, ob die Änderung korrekt verarbeitet wurde.

Kapitel 6

Extraktion semantischer Information

Inhalt

Um die Aussagekraft des zugrundeliegenden Weltmodells über Objektinstanzen hinaus zu erhöhen, wird zusätzliche semantische Information basierend auf den erkannten Objekten und der rekonstruierten Umwelt benötigt.
Dieses Kapitel gibt einen Überblick über verschiedene Arten von Semantik und unterschiedliche Fragestellungen bezüglich der Möglichkeiten und Notwendigkeiten (Abschnitt 6.1). Dies führt zur Betrachtung, inwieweit zusätzliche Semantik hilfreich in der Robotik sein kann. Da der Begriff Semantik in der Literatur vielseitig und unterschiedlich verwendet wird, muss diese Begrifflichkeit zuerst näher betrachtet werden (Abschnitt 6.2). Abschließend wird eine formale Definition erarbeitet, um Semantik basierend auf den wiedererkannten Objekten zu bestimmen und im Weltmodell zu hinterlegen (Abschnitt 6.3).

6.1 Stand der Forschung

Der hier dargestellte Stand der Forschung teilt sich in drei Aspekte: Zunächst wird die Notwendigkeit und der Ursprung identifiziert; anschließend werden unterschiedliche Arten von Semantik sowie Spezifikations- und Zuordnungsprobleme beschrieben. Abschließend werden verschiedene Anwendungen vorgestellt. Als Fazit ist festzuhalten, dass die theoretischen Möglichkeiten umfangreich sind, was alles beschrieben und wie es repräsentiert werden kann. Daher ist als Folge die Überlegung notwendig, welche weiterführende Information überhaupt beschrieben und bestimmt werden kann beziehungsweise soll.

6.1.1 Semantische Lücke

Das Problem der Diskrepanz zwischen menschlichem Verständnis von Szenen und möglicher, technischer Extraktion ist bekannt und in verwandten Disziplinen der Bildverarbeitung als *Semantische Lücke* (*Semantic Gap*) ein Begriff [Enser03, Liu07, Li16]. Beispielsweise im *Content-based image retrieval* ist diese Frage ursprünglich von Interesse [Li16, Sudha15]. Ziel ist es, mit einer abstrakten, semantischen Beschreibung eines Menschen Bilder zu finden, die von dieser Beschreibung erfasst werden. Ein anderes Beispiel ist die Fernerkundung, in der die Unterschiede zwischen einer menschlichen und einer automatisierten Bildanalyse reduziert werden sollen [Bahmanyar15]. Analog existiert diese Herausforderung in der Robotik in unterschiedlichen Anwendungen. Beispielsweise ist bei der natürlichsprachlichen Kommandierung von Robotersystemen eine abstrakte Beschreibung für Menschen natürlicher als eine detaillierte Abfolge von mehreren Roboterposen und Aktionen [Kunze11]. Das Ziel ist somit, den Unterschied in der semantischen Beschreibung zwischen abstrakten, hochstufigen Anweisungen und tatsächlich direkt umsetzbaren Roboteraktionen zu reduzieren.

6.1.2 Arten von Semantik für die Robotik

In diesem Abschnitt wird ein Überblick über Semantik im Rahmen der Robotik gegeben. Dieser Überblick ist angelehnt an [Garg20] und [Liu23] und wird hier ergänzt. Dabei wird sich hier auf verwandte Arbeiten beschränkt, die Informationen über die Objekt(wieder)erkennung hinaus erzeugen, da diese bereits in Kapitel 3 diskutiert wurde. Hierbei lässt sich die Art der Semantik anhand der Verwendung unterteilen. In [Liu23] wird die Semantik den unterschiedlichen Komponenten eines vollständigen Robotiksystems zugeordnet: *Objekt* (welche Eigenschaften und Zusammenhänge liegen vor), *Raum* (in welcher Beziehung stehen Objekte und weitere Räume untereinander), *Aufgaben* (welche Eigenschaften und Voraussetzung haben Aufgaben), *Aktionen* (die tatsächliche Ausführung einer Aktion als Teil der Aufgabe) und *Agenten* (welche Information ist über die unterschiedlichen Agenten gegeben). Jede Komponente wird dabei weiter unterteilt, über die hier ein kurzer Überblick gegeben wird. Semantik bezüglich Objekten wird aufgeteilt in die gewählte Repräsentation [Gao21], Semantik für jede Objektinstanz [Achlioptas20, Thomason22] und Information über eine Klasse von Objekten [Lemaignan17, Tenorth17]. Für Räume wird eine analoge Definition gegeben. Für Aufgaben wird eine Hierarchie eingeführt, womit Aufgaben beschrieben werden können, bestehend aus der Zerteilung von Aufgaben in Teilschritte [Kaelbling11], einer Vorlage für diese

Teilschritte [Garrett21] und der tatsächlichen Spezifikation. Aktionen bestehen zum einen aus einer Repräsentation der Welt und der Aktion als auch aus einer semantischen Repräsentation, wie diese Aktion die Umwelt beeinflusst [Krüger11, Zellers21]. Abschließend können für die Agenten deren Möglichkeiten [Kunze11] und deren jeweilige Umweltrepräsentation definiert werden [Lemaignan17].
Alternativ ordnet [Garg20] Semantik in der Robotik entsprechend ihres Nutzens ein. Ein knapper Überblick wird hier wiedergegeben, für Details siehe ebenda. Eine Möglichkeit vor allem für mobile Roboter ist die Beschreibung der Umgebung, beispielsweise die Zuordnung des aktuellen Raums, in dem sich der Roboter befindet [Rottmann05, Stachniss06, Luperto16]. Neben der Klassifikation des Raums können Objekte in diesem wiedererkannt werden [Nüchter08, Tateno16]. Neben der Beschreibung der Umgebung ist vor allem die Interaktion mit der Umwelt und weiteren Agenten von Interesse. Ein Aspekt ist dabei die Beobachtung von Interaktion, beginnend mit den Aktionen eines menschlichen Agenten (bekannt unter *Action Recognition*). In diesem Rahmen wird die Frage aufgeworfen, welche Aktionen wie genau betrachtet werden müssen [Carreira17]. Einen Überblick für die Robotik gibt [Ramirez-Amaro19]. Dabei kann sich auch vor allem auf Hände und Arme beschränkt werden, da diese primär mit Szenen interagieren [Ramirez-Amaro14], oder auf weitere Extremitäten [Yoon18, Wang18]. Ein anderer Aspekt ist die Beschreibung von Interaktion, beginnend mit Eigenschaften von Objekten, auch als *Affordanzen* bekannt [Gibson79], die von einem Robotersystem gelernt werden können, falls nicht anders gegeben [Sahin07, Aksoy15]. Zur Handhabung von Objekten kann die Greifplanung mit semantischer Information unterstützt werden [Roa15, Tremblay18]. Eine Kombination der bisher erwähnten Semantik ermöglicht das Beschreiben komplexer Aktionen [Blodow11, Zeng18].
Neben den bisher vorgestellten Anwendungsmöglichkeiten kann Semantik auch aus der natürlichen Sprache und alltäglicher Geometrie definiert werden. Eine Möglichkeit sind topologische Zusammenhänge zwischen Objekten [Freeman75, Clementini93, Hernandez94, Li09], wie beispielsweise *berühren* oder *umschlossen*. Als weiterer Ansatz können natürlichsprachliche Richtungen angegeben werden wie *links* oder *oberhalb* [Freeman75,Miyajima94], auch unter Beachtung einer Beobachtungspose in 2D [Moratz02, Kunze14] oder 3D [Chen10]. Abschließend können Distanzen untersucht werden in Form von Größen oder Abständen zwischen Objekten [Kunze14, Thippur15].
Abschließend ist neben der expliziten Formulierung von Semantik auch die implizite Verfügbarkeit von Semantik eine Option. Explizit meint dabei eine direkte Berechnungsvorschrift ausgehend von Objektinstanzen. Implizit bedeutet, dass ein semantisches Verständnis bei der Erfüllung einer Aufgabe genutzt wird, ohne dass es separat berechnet und im Weltmodell hinterlegt wird. Für eine implizite Verfügbarkeit sind vor allem die in Kapitel 2 erwähnten *Foundation Models* und *Large Language Models* relevant [Chen23b], auf die an dieser Stelle nicht noch einmal eingegangen wird.

6.1.3 Spezifikations- und Zuordnungsprobleme

Eine grundlegende Herausforderung bei der Beschreibung und Auswahl von Semantik ist die Frage, welches Wissen für welche Aufgaben notwendig ist. Im Rahmen der künstlichen

Intelligenz sind vor allem das Rahmenproblem (*Frameproblem*) [McCarthy81, Hayes81] und *Ramificationproblem* [Thielscher97] bekannt. Dabei werden Herausforderungen aufgezeigt, wie der Zustand und die Auswirkungen von Aktionen auf eine Welt zu beschreiben sind. Das umschließt sowohl Eigenschaften, die erfüllt sein müssen, um eine Aktion durchzuführen, als auch die Frage, in welchem Zustand sich die Welt nach dem Durchführen einer Aktion befindet. Das Frameproblem bezieht sich dabei vor allem auf Untersuchungen im Rahmen der formalen Logik und ist in diesem Zusammenhang auch gelöst - beispielsweise mittels *Default Reasoning* [Reiter80], *Event Calculus* [Shanahan97], *Situation Calculus* [McCarthy63], *Fluent Calculus* [Thielscher98] oder durch Reduzierung der Komplexität auf eine einzelne Umweltrepräsentation statt vieler möglicher [Ginsberg88a].
Als verwandtes Problem existiert das *Qualifikationsproblem* [McCarthy80], welches die Möglichkeit in Frage stellt, alle notwendigen Vorbedingungen zu formulieren, die für eine erfolgreiche Interaktion notwendig sind. Ein entscheidender Unterschied zum Frameproblem ist dabei die allgemeine Anwendung auf beliebige Aufgaben ohne Einschränkung auf Logik. Ein möglicher Lösungsansatz versucht, die Komplexität zu reduzieren, indem keine Vorbedingungen erfüllt sein müssen, sondern Einschränkungen nicht verletzt sind [Ginsberg88b]. Ein alternativer Ansatz mittels *State constraints* ist das Ausschließen von spezifischen Vorkommnissen, wodurch ebenfalls die Komplexität reduziert wird [Lin94, Baral00].
In diesem Zusammenhang ist auch das *Symbol Grounding Problem* zu erwähnen [Harnad90] (mit Erweiterungen und Lösungsvorschlägen, beispielsweise [Harnad01]). Das Problem entspringt dem *Chinese Room Problem* [Searle80], das die Frage aufwirft, ob eine rein algorithmische Ausgabe auf eine definierte Eingabe als *intelligent* anzusehen ist. Die grundlegende Frage und Konsequenzen wurden vielfach diskutiert (beispielsweise [Taddeo05, Li22]). Dies hat in der Robotik zur Konsequenz, dass semantische Beschreibungen auf roboterspezifische Aktionen abgebildet werden müssen. Für einen Überblick siehe [Coradeschi13,Cohen24] oder im spezifischen [Spangenberg17, Wölfel21].
Abschließend ist das *(Visual) Anchoring Problem* eine offene Fragestellung in diesem Zusammenhang [Coradeschi03]. Dabei wird untersucht, inwieweit Sensorsignale konkreten Objektinstanzen zugeordnet werden können. Diese Fragestellung bezieht sich einerseits auf einen einzelnen Zeitpunkt, sodass einem geometrischen Objekt (Pixel mit Tiefendaten, Voxel, B-Rep Element) eine Objektinstanz zugeordnet wird (in dieser Arbeit beantwortetet durch die Methode der Objektwiedererkennung). Andererseits ist ein zeitlicher Zusammenhang zwischen mehreren Zeitpunkten von Interesse, sodass nachvollziehbar ist, wofür ein spezifisches Objekt verwendet wurde.

6.1.4 Anwendungsmöglichkeiten von Semantik

In [Garg20] werden unterschiedliche Möglichkeiten für die Verwendung von Semantik dargestellt. Diese werden hier gemäß verschiedenen Anwendungen zusammengefasst und ergänzt. Eine Möglichkeit sind *Drohnen* (*Unmanned Aerial Vehicles*), bei denen Semantik beispielsweise für die Pfadplanung verwendet werden kann, um Kollisionen zu vermeiden. Dazu wird die Information verwendet, welche Objekte in der Szene sich wie verhalten können und mög-

licherweise einen Einfluss auf die Flugbahn haben. Falls von Objekten eine Gefahr für die Drohne ausgeht, kann der Pfad angepasst werden [Loquercio18, Benjdira19, Gandhi17].
Eine ähnliche Anwendung in der Robotik ist das autonome Fahren. Auch hier liegt der Fokus auf kollisionsfreien Bewegungen und dem Umgang mit plötzlichen Veränderungen. Der Vorteil im Vergleich zu Drohnen ist aber die größere Menge an verfügbarer Sensorik [Feng19].
Im Bereich der *Service-Robotik* existieren diverse Anwendungsgebiete wie Haushalt [He17, Cheng18], Pflege [Vänni17], Restaurants [Tuomi20] und weitere Industrien [Savela18]. Die grundlegende Frage ist dabei die Akzeptanz von Robotern in den entsprechenden Feldern, die durch ein besseres Verständnis der Umwelt erhöht werden könnte. Darüber hinaus ist Semantik in der Mensch-Roboter-Interaktion relevant, um die Ausdrucksfähigkeit von Robotern und somit das Verständnis zu erhöhen [Breazeal09]. Eine Möglichkeit ist der Versuch, die Interaktion und Bewegung natürlicher zu gestalten [Fang19, Savage19, Lai18, Giambattista16]. Dazu ist auch die Navigation von Robotern in gemeinsamen Arbeitsbereichen relevant [Zender08, Talbot21, Ginting24]. Auch die Interaktion von Roboter zu Mensch in der Spezifikation von Bauteilen [Tellex14] oder in der Mensch-Roboter-Kollaboration [Akkaladevi21] kann mit semantischer Information unterstützt werden.
Weiter gefasst kann Semantik auch in *zivilen* Aufgaben eingesetzt werden. Mögliche Anwendungen in diesem Bereich sind die Analyse von bewegten Kameras in städtischen Szenen [Yazdi18] oder Krisenmanagement [Kostavelis17].
Abschließend findet Semantik auch Anwendung in der Manipulation von Szenen, indem räumliche Relationen genutzt werden [Kartmann21, Liu22, Wang23].

6.1.5 Fazit

Ausgehend vom Stand der Forschung lässt sich für die Aufgabenstellung dieser Arbeit ein vorläufiges Fazit ziehen. Zum einen sind die Verwendungsmöglichkeiten von Semantik im Allgemeinen und in der Robotik im Speziellen sehr vielseitig [Garg20, Liu23]. In der Robotik selbst gibt es diverse Arbeiten, die sich mit aufgabenspezifischer Semantik auseinandersetzt.
Somit verbleiben unterschiedliche Aspekte, die näher betrachtet werden können. Zum einen ist eine allgemeine Definition von Semantik in der Robotik nicht oder nur teilweise geläufig. Für eine eindeutige Basis ist somit von Interesse, was unter Semantik auf welchem Grad der Datenverarbeitung zu verstehen ist. Zum anderen ist eine spezifische Beschreibung von weiterer Semantik im Rahmen dieser Arbeit nicht gegeben. Neben der sehr breiten Untersuchung, wo überall Semantik weiterhilft, sind die jeweiligen Disziplinen in der Tiefe gut untersucht. Darüber hinaus werfen die unterschiedlichen Spezifikationsprobleme die Frage auf, inwieweit eine abschließende Liste von Semantik überhaupt möglich und zielführend ist. Da diese Arbeit nicht auf eine konkrete Domäne beschränkt wird, ist die Einschränkung auf eine Anwendung nicht angebracht. Statt der individuellen Beschreibung von ausgewählter Semantik ist ein allgemeines Konzept, wie Semantik definiert werden kann, von Interesse. Auf diese Weise kann die Menge an bereits gegebenem semantischen Wissen und wie man es aus einer Szene berechnet, verwendet werden, um diese auf eine einheitliche Art zu definieren.
Dabei können aus den bisherigen Ergebnissen und dem Stand der Forschung einige Einschränkungen definiert werden: Da kein kontinuierlicher Blick auf die Szene möglich ist, wird

in der Beschreibung der Definition eine zeitliche Komponente nicht beachtet. Auf diese Weise ist auch der zweite Teil des *Anchoring Problems* in diesem Zusammenhang nicht relevant. Darüber hinaus wird für die Semantik eine Menge an erkannten Objekten als Grundlage angesehen für die semantische Information berechnet werden soll. Weitere Teile der Umweltrepräsentation sollen nicht genutzt werden. Abschließend ist das Ziel ein modellbasierter beziehungsweise regelbasierter Ansatz. Ein Lernverfahren zur Extraktion von Semantik verfügt über die gleichen Nachteile wie Lernverfahren zur Objektwiedererkennung. Zum einen ist der Trainingsaufwand hoch, was die Erweiterung um neues Wissen erschwert. Zum anderen ist eine Nachvollziehbarkeit nur schwer möglich.

6.2 Definition Semantik

Die Grundlage für jede semantische Betrachtung sind *Daten*. Je nach betrachteter Domäne können Daten als Eingabe unterschiedliche Ausprägungen annehmen. Wenn man die natürliche Sprache betrachtet, kann eine Reihe von Buchstaben eine Eingabe sein. Im Rahmen dieser Arbeit ist die Grundlage eine Menge von wiedererkannten Objekten. Um aus dieser Eingabe weiterführende Informationen zu gewinnen, benötigt es zusätzliches *Kontextwissen*. Denn dabei ist zu beachten, dass aus der für sich alleinstehenden Eingabe nicht weiter Information erlangt werden kann. Am Beispiel der Reihe von Buchstaben kann beispielsweise der Wortschatz einer Sprache als Kontextwissen vorhanden sein. Mit Hilfe dieses Wortschatzes kann aus einer Eingabe von Buchstaben dann die Information abgeleitet werden, dass es sich um ein Wort handelt. Ein Beispiel ist als Eingabe die Buchstabensequenz *Bank*. Mit einem englischen Wortschatz als Kontextwissen kann dieser Buchstabenfolge somit ein Wort zugeordnet werden, mit einem französischen aber beispielsweise nicht. Auch mit einem deutschen Wortschatz kann aus dieser Buchstabenfolge abgeleitet werden, dass es sich um ein Wort handelt. Die Bedeutung des Wortes ist damit aber selbst mit Festlegung auf eine Sprache noch nicht definiert, da diese mehrdeutig sein kann. Im Englischen ist es beispielsweise sowohl das Kreditinstitut als auch das Ufer. Um die Bedeutung des Wortes in diesem Zusammenhang festzulegen, ist somit weiteres Kontextwissen notwendig, wie beispielsweise der Rest der Zeicheneingabe.
Explizit formuliert ist Semantik eine Funktion, die als Eingabe Daten und Kontextwissen benötigt. Durch die Anwendung des Kontextwissens auf die Daten erhält man somit als Ausgabe weiterführende Informationen. Diese Definition lässt sich rekursiv mehrmals anwenden. So kann die extrahierte semantische Information zusammen mit anderem Kontextwissen wieder als Eingabe für eine semantische Abbildung dienen.
Diese Definition und rekursive Betrachtung kann nun auf einzelne Schritte dieser Arbeit angewendet werden: Als ursprüngliche Ausgabe des Sensors liegen Rohdaten in Form einer Punktwolke vor. Mit geeignetem Kontextwissen in Form von Parametern für die B-Rep Erzeugung (beispielsweise der Strukturgröße) kann mit Hilfe der entsprechenden *Semantik-Funktion* ein B-Rep aus der gegebenen Punktwolke erzeugt werden (siehe Kapitel 2 beziehungsweise [Sand19]). Der Mehrwert an Information liegt dann darin, dass beschrieben wird, wo sich Flächen, Kanten und Knoten in der Umwelt befinden und dadurch eine Beziehung zwischen den Punkten der ursprünglichen Eingabe hergestellt wird. Diese geometrische Repräsentation

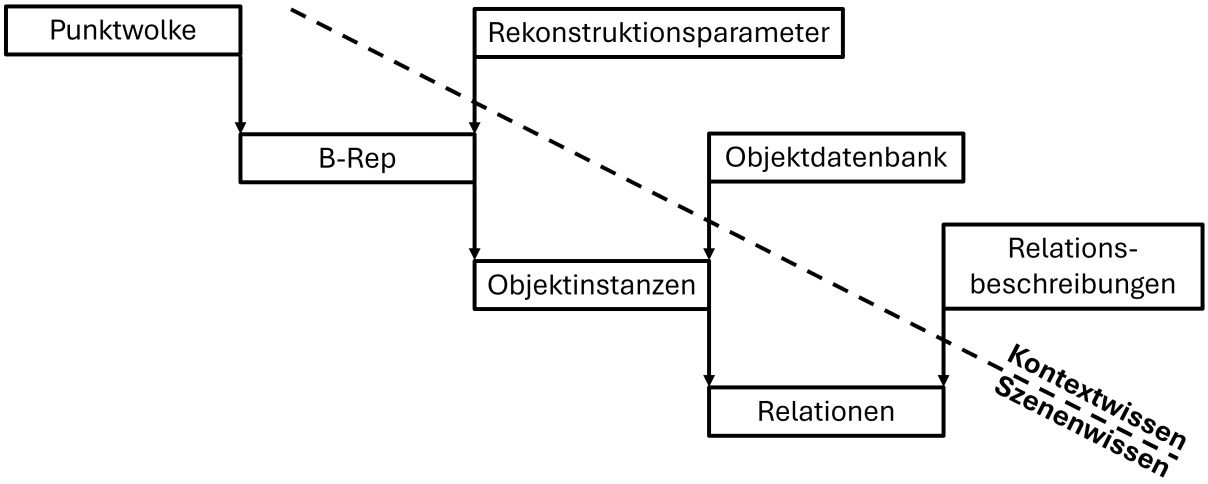


Abbildung 6.1: Schematische Darstellung unterschiedlicher Semantikebenen in dieser Arbeit

(hier B-Reps) wird anschließend mit zusätzlichem Kontextwissen in Form einer Objektdatenbank weiter analysiert. Die Anwendung der Objektdatenbank auf die Umweltrepräsentation gemäß der vorgestellten Schritte (Hypothesenerzeugung und -auswahl, siehe Kapitel 3) führt anschließend zu einer Menge an ausgewählten Hypothesen. Als nächster Schritt ist dann somit das Ziel, aus dieser Menge an Objektinstanzen zusammen mit einer Menge an beschriebenen Relationen, Zusammenhänge zwischen den wiedererkannten Objekten herzustellen, was in anderen Arbeiten häufig als Semantik bezeichnet wird. Hieraus lässt sich die Unterscheidung zwischen allgemeinerem Kontextwissen und den tatsächlichen Daten ableiten: Das Kontextwissen in Form von B-Rep Parametern, Objektdatenbank oder Relationen ist unabhängig von der tatsächlichen Eingabe oder den konkreten Daten in der aktuellen Szene. Trotzdem kann das Kontextwissen abhängig von der betrachteten Domäne sein, beispielsweise ist für die B-Rep Erzeugung je nach Sensor und Objekten eine andere Strukturgröße angebracht. Der ursprüngliche Sensoreindruck zusammen mit allen abgeleiteten Informationen beschreibt aber die vorliegende Szene mit immer größer werdenden Informationsgehalt und lässt sich somit als *Szenenwissen* zusammenfassen. Siehe Abbildung 6.1 für eine Darstellung der unterschiedlichen Semantikebenen in dieser Arbeit.
Mit dieser Definition kann auch die unterschiedliche Auffassung von *Semantik* im Stand der Forschung abgebildet werden. In manchen Arbeiten und Anwendungen wird bereits von Semantik gesprochen, wenn im Sensoreindruck die Objekte der Szene wiedererkannt werden. In anderen Domänen sind die Eigenschaften und Zusammenhänge als Semantik definiert [Garg20]. Insgesamt kann der gesamte Prozess über die Verrechnung der Sensordaten, der Objektwiedererkennung bis hin zu weiteren Informationen als semantischer Zugewinn zusammengefasst werden.

6.3 Beschreibung von Semantik

Ausgehend von der Definition von Semantik wird nun eine Möglichkeit vorgestellt, wie das Kontextwissen beschrieben werden kann, um Zusammenhänge zwischen Objekten darzustellen (somit die letzte Stufe aus dem vorangegangenen Definition). Im Stand der Forschung wurden unterschiedliche Arten von Semantik in Hinblick auf deren Nutzen und Beschreibung vorgestellt. Im Rahmen dieser Arbeit ist die informationsreichste Eingabe die Menge der wiedererkannten Objekte. Neben dem gegebenen CAD-Modell in Form von B-Reps und der dazugehörigen Pose in der aktuellen Szene kann weitere Information in der Objektdatenbank mit hinterlegt sein, die am wiedererkannten Objekt mit gespeichert ist, beispielsweise physikalische Eigenschaften (Material, Masse, Schwerpunkt). Somit sind Eigenschaften, die ein Objekt individuell charakterisieren, bereits abgedeckt. In der Domäne der Robotik sind vor allem Zusammenhänge zwischen den Objekten relevant, da diese für die Handhabung und verwandte Aufgabenstellung entscheidend sind. Diese Zusammenhänge unterscheiden sich dabei je nach Anzahl der betrachteten Objekte, den Einfluss der genauen Positionen oder Vorhandensein von weiteren Objekten. Da das Ziel die explizite Modellierung von Zusammenhängen ist, werden in dieser Arbeit *Relationen* genutzt, um diese Informationen zu modellieren.

6.3.1 Formale Beschreibung

Als Grundlage für die Extraktion von Beziehungen zwischen Objekten als weitere semantische Information werden wiedererkannte Objekte angenommen. Damit sind die Pose der konkreten Objektinstanz in der Szene bekannt sowie weitere Information, die gegebenenfalls in der Objektdatenbank gehalten werden. Analog zur bisherigen Formalisierung ist das die Menge $H = \{h_1, ..., h_n\}$, wobei eine Hypothese h_i mindestens aus dem Objekt-Modell und dessen Transformation in die Szene besteht.
Als mathematisches Grundkonzept zur Beschreibung von Semantik werden Relationen verwendet. Die Motivation dahinter ist - neben dem vorgestellten Stand der Forschung - die Analogie zum Szenenverständnis des Menschen. Für die nähere Beschreibung von Objekten werden Eigenschaften von Objekten relativ zur Szene herangezogen, beispielsweise Größe und Positionierung. Weiterhin soll die Beschreibung der Semantik unabhängig von der Anzahl der Objekte sein. Dadurch ist keine neue Beschreibung notwendig wenn sich die Zahl an beteiligten Objekten ändert. Damit werden Relationen hier stets auf der vollständigen Menge der aktuellen Hypothesen beschrieben.
Um sowohl eine kompakte Beschreibung von Relationen zu ermöglichen, aber gleichzeitig auch eine hohe Mächtigkeit der möglichen Beschreibungen sicherzustellen, wird hier ein hierarchischer Ansatz verfolgt. Dazu werden im Weiteren vier unterschiedliche Definition von Zusammenhängen vorgestellt, wobei die Modellierungsmächtigkeit steigt und höherstufige Definitionen mindestens so mächtig sind wie niedrigere.
Als erste Grundlage werden somit *Nicht reduzierbare Relationen* (`NRR`) definiert:

$$\texttt{NRR} \subseteq \{(h_1, ..., h_k) \in H^k | a((h_1, ..., h_k))\} \tag{6.1}$$

Die Relationen haben einen Grad k, je nachdem wieviele Hypothesen an der Relation beteiligt sind. Ob ein Tupel des kartesischen Produkts tatsächlich in der Relation liegt, wird mit Hilfe der Wahrheitsfunktion $a : H^k \rightarrow \{True, False\}$ bestimmt. An diese Funktion a werden gesonderte Anforderungen gestellt, um die Beschreibung von Relationen zu strukturieren. Dabei sind keinerlei logische Verknüpfungen zugelassen, sondern lediglich direkte Auswertungen einzelner Eigenschaften der entsprechenden Objekte. Eine Kombination mehrerer logischer Funktionen ist somit explizit ausgeschlossen. Die Negation ist erlaubt, da diese alternativ direkt in der Wahrheitsfunktion umgesetzt werden könnte. Als Wahrheitswert darf nicht die Erfüllung höherwertigerer Relationen verwendet werden (die im Weiteren noch vorgestellt werden). Abschließend ist zu beachten, dass nur die Hypothesen selbst, die Element der Relation sind, in der Wahrheitsfunktion zur Verfügung stehen. Ein Bezug auf die gesamte Menge an Hypothesen H ist auf diese Weise nicht möglich und bewusst ausgeschlossen. Nicht reduzierbare Relationen - wie hier definiert - erfüllen somit den Zweck einer Grundlage, um weitere Relationen einfach zu beschreiben.
Als nächster Schritt in der Komplexität der Beschreibung sind *Referenzfreie Relationen* (RFR) anzusehen. Die Definition folgt der vorherigen:

$$\texttt{RFR} \subseteq \{(h_1, ..., h_k) \in H^k | b((h_1, ..., h_k))\} \tag{6.2}$$

Die Wahrheitsfunktion $b : H^k \rightarrow \{True, False\}$ entspricht von der Signatur der vorherigen Wahrheitsfunktion. Zur Verfügung stehen dabei alle logischen Verknüpfungen in beliebiger Anzahl sowie die Kombination von nicht reduzierbaren Relationen beziehungsweise deren Wahrheitsfunktion a. Höherwertige Relationen sind weiterhin keine Option. Da auch hier noch nicht die komplette Szene als Referenz verwendet werden soll, steht H weiterhin nicht zur Verfügung, sondern nur die direkt an der Relation beteiligten Hypothesen.
Die logische Erweiterung der Relationen bezieht nun die vollständige Menge an Hypothesen als mögliche Referenz mit ein, was zur Definition der *Referenzbehafteten Relationen* (RBR) führt.

$$\texttt{RBR} \subseteq \{((h_1, ..., h_k), M) \in H^k \times \mathcal{P}(H) | c((h_1, ..., h_k), M)\} \tag{6.3}$$

Im Vergleich zu den bisherigen Relationen liegt in diesem Fall eine Relation vom Grad $k+1$ vor. Die Erhöhung des Grades spiegelt die Menge an möglichen Referenzobjekten wider. Somit wird die Relation über der Menge $H^k \times \mathcal{P}(H)$ definiert. Dazu wird das kartesische Produkt über alle Hypothesen in der Szene und über die Potenzmenge aller Hypothesen gebildet. Durch die Potenzmenge wird jede mögliche Kombination an Objekten als potentielle Referenzmenge dargestellt. Ein Eintrag der Relation ist somit ein Tupel bestehend aus k Hypothesen sowie der Referenzmenge M. Dies spiegelt sich auch in der Wahrheitsfunktion wieder $c : H^k \times \mathcal{P}(H) \rightarrow \{True, False\}$. Zur Evaluation dieser Funktion sind somit nicht nur die k Hypothesen der Relation notwendig, sondern auch die Referenzmenge. Dadurch, dass die Referenzmenge variabel und gegeben ist, können Relationen spezifisch für bestimmte Gruppierungen bestimmt werden. Durch explizite Angabe der Referenzmenge ist stets die Nachvollziehbarkeit sichergestellt, mit welchen Objekten ein Vergleich gültig ist.
Abschließend ist eine letzte Art von Relationen von Interesse, die die Beschreibungsmächtig-

keit weiter erhöht. Dabei ist das Ziel, nicht nur einzelne Objekte in beliebig-stelligen Relationen in Verbindung zu setzen, sondern die Objekte gruppieren zu können. Auf diese Weise können Mengen von Objekten miteinander in Relation gesetzt werden, definiert durch *Mengenrelationen* (MR)

$$\texttt{MR} \subseteq \{(H_1, ..., H_k) \in \mathcal{P}(H)^k | d((H_1, ..., H_k))\} \tag{6.4}$$

Diese k stellige Relation umfasst dabei nicht wie bisher nur einen Eintrag mit einzelnen Hypothesen, sondern eine beliebige Menge an Hypothesen. Somit ist die Grundmenge dieser Relation das kartesische Produkt der Potenzmenge aller Hypothesen. Im Gegensatz zu referenzbehafteten Relationen muss hier eine Referenzmenge nicht explizit mitgeführt werden, da diese anhand der Potenzmenge aller Hypothesen dargestellt werden kann (dadurch sind die Mengenrelationen auch wieder vom Grad k). In die notwendige Wahrheitsfunktion $d : \mathcal{P}(H)^k \rightarrow \{True, False\}$ gehen daher nur die Gruppen an Objekten ein.
Für diese vier Arten, Beziehungen zu definieren, steigt die Mächtigkeit kontinuierlich an. Beginnend bei Mengenrelationen können diese die gleichen Relationen beschreiben wie referenzbehaftete Relationen. Wie dargestellt kann dafür die Referenzmenge als eine der Stellen in der Relation genutzt werden. Da an die Anzahl an Hypothesen pro Eingabemenge keine Anforderungen gestellt werden, kann eine Menge auch nur aus einem Element bestehen. Somit lassen sich referenzfreie Relationen durch Mengenrelationen beschreiben. Dabei ist zu beachten, dass bei der Mengenrelation eine ein-elementige Menge in die Relation eingeht, bei referenzfreien Relationen das Element direkt (was für die Semantik hier aber gleichbedeutend ist). Der Vergleich zwischen referenzbehaftet und referenzfrei erfolgt auf eine ähnliche Art, da die Menge M an Referenzobjekten leer sein kann. Da referenzbehaftete Relationen keinerlei Einschränkungen an die Wahrheitsfunktion haben, ist die Mächtigkeit mindestens so groß wie die der referenzfreien. Der abschließende Vergleich zwischen refrenzfreien und nicht reduzierbaren Relationen kann direkt erfolgen, da die Formalisierung identisch ist. Nur die Wahrheitsfunktion der nicht reduzierbaren Relationen ist weiter eingeschränkt, wodurch diese gesichert weniger mächtig sind als referenzfreie Relationen. Diese Eigenschaften setzen sich transitiv fort, wodurch mit den Mengenrelationen alle darunterliegenden Gruppen auch beschrieben werden könnten.
Wie im Stand der Forschung dargelegt, ist für manche Zusammenhänge zwischen Objekten eine Beobachtungssicht oder allgemein ein fixiertes Koordinatensystem notwendig (beispielsweise für Richtungen wie *links* und *rechts*). Dies kann im Allgemeinen als eine Transformation zwischen aktuellem und gewünschtem Referenzkoordinatensystem dargestellt und bei der vorgestellten Beschreibung unterschiedlich umgesetzt werden: Zum einen kann die Transformation expliziert mitgeführt werden, wodurch sich alle Beschreibungen der vier Gruppen um die Transformation als zusätzlichen Parameter erweitern würde. Zum anderen können die gegebenen Hypothesen vor der Berechnung der Relationen in das gewünschte Koordinatensystem transformiert werden. Analog zu vorherigen Kapiteln genügt es dabei, die Eckpunkte zu transformieren, da sich durch eine linear affine Transformation die Geometrie der Objekte nicht ändert.

6.3.2 Beispielhafte Beschreibungen

Im Weiteren wird beispielhaft semantische Information mit dem eben vorgestellten Konzept beschrieben. Dazu wird im Folgenden davon ausgegangen, dass alle Objekte bereits im Referenzkoordinatensystem des Beobachters vorliegen und dieses somit nicht weiter beachtet werden muss.

Größenrelationen

Zunächst können Größenrelationen formuliert werden, welche die räumliche Ausdehnung charakterisieren. Repräsentativ für diese Gruppe wird das Adjektiv *groß* in allen sich steigernden Vergleichsformen betrachtet. In der Grundform kann somit die binäre Aussage getroffen werden, dass ein beliebiges Objekt *groß* ist. Dies kann als ein einfacher Vergleich der Objektgröße (beispielsweise aus Objektdatenbank oder Differenz der Ausdehnung) erfolgen. Somit muss die Größe des Objekts lediglich mit einem gegebenen Grenzwert (beispielsweise begründet durch eine entsprechende Anwendung) verglichen werden. Dies lässt sich somit als *nicht reduzierbare Relation* `NRR` darstellen. Die Wahrheitsfunktion umfasst dabei nur einen Vergleich zwischen den zwei gegebenen Größen. Weiterhin werden keine Referenzobjekte benötigt, da sich die Relation nur auf ein Objekt selbst bezieht. Aus diesem Grund ist *groß* als einstellige `NRR` zu definieren.
Der Komparativ *größer* kann ebenfalls als `NRR` definiert werden. Im regulären Sprachgebrauch schließt das genau zwei Objekte ein. Somit liegt eine zweistellige Relation vor mit einer Wahrheitsfunktion, bei der die Größen der zwei Objekte direkt miteinander verglichen werden.
Abschließend muss der Superlativ *am größten* beschrieben werden. Wie bei der Definition der unterschiedlichen Relationen soll die Stelligkeit nicht von der Anzahl der betrachteten Objekte abhängen. Für eine referenzfreie Relation müsste für jede unterschiedliche Anzahl an Objekten diese Relation explizit neu formuliert werden. Aus diesem Grund wird diese Eigenschaft als `RBR` dargestellt. Somit ist *am größten* eine zweistellige Relation wobei alle weiteren Objekte als Referenzmenge M angegeben werden. In der Wahrheitsfunktion kann dann analog zu den bisherigen Fällen die Größe des zu untersuchenden Objekts mit allen Elementen aus M verglichen werden. Falls kein Objekt in M existiert, das größer ist als das Untersuchungsobjekt, ist die Relation *am größten* erfüllt.
Abschließend können Gruppen von Objekten verglichen werden, welche größer sind. Für diese Betrachtung ist eine Mengenrelation *MR* notwendig. Beispielsweise können zwei Mengen von Objekten in einer zweistelligen Relation verglichen werden. Die Wahrheitsfunktion würde dann die beiden Gruppen miteinander vergleichen. So kann zum Beispiel die aufsummierte oder mittlere Größe der jeweiligen Gruppe verglichen werden. An diesem Beispiel wird eine Problematik der Mengenrelationen deutlich: Da die Objekte gruppiert werden müssen, muss jede Stelligkeit der Relation eindeutig definiert werden. Da es sich aber bereits um eine Aggregation von Objekten handelt, ist die Notwendigkeit, unterschiedlich stellige Relationen zu definieren, vergleichsweise gering.
Auf eine ähnliche Weise können auch Richtungsrelationen definiert werden. So kann ein Objekt in einer Szene für sich alleine stehend an einer relevanten Stelle stehen (beispielsweise *links*). Analog kann der Komparativ zwischen zwei Objekten beschrieben werden - genauso

wie relativ zu einer Gruppe oder als Teil einer ebensolchen. Als Kombination dieser Relation kann beispielsweise *links und groß* als *RFR* definiert werden, indem die Relationen *links* und *rechts* mit einem entsprechenden logischen Operator kombiniert werden. Die direkte Kombination wäre per Definition als *NRR* nicht möglich, daher ist eine *RFR* notwendig.

Stabilitätsbeziehungen

Für die fehlerfreie Handhabung von Szenen mit mehreren, sich berührenden Objekten sind Stabilitätsbeziehungen zwischen diesen notwendig. Dazu werden in [Mojtahedzadeh13, Mojtahedzadeh14, Mojtahedzadeh16] die sogenannten *Act* und *Support* Beziehungen zwischen Objekten eingeführt. Das Ziel dabei ist es, Objekte zu identifizieren, die aus der Szene entfernt werden können, ohne dass sich die Pose und Konfiguration der verbleibenden Objekte verändern. Dabei untersucht *Act*, ob ein gegebenes Objekt auf genau ein anderes eine Kraft ausübt. Davon ausgehend beschreibt *Support*, ob ein Objekt ein anderes entgegen der Schwerkraft unter Betrachtung aller weiteren Objekte stützt. Dazu werden mögliche Kontaktzustände berechnet und unter Annahme von Massengleichverteilung und Kenntnis weiterer relevanter physikalischer Größen (beispielsweise Gewicht und Reibung) ein Kräftegleichgewicht aufgestellt. Da die Pose aller Objekte bekannt ist, kann mit den gegebenen Annahmen berechnet werden, wie sich die Szene verhalten würde, wenn ein Objekt nicht mehr vorhanden wäre. Falls sich ein Objekt O in der Szene nach dem Entfernen des untersuchten Objekts A bewegt, folgt daraus, dass O von A gestützt wird.
Diese beiden Relationen werden beispielhaft mit dem vorgestellten Konzept beschrieben. Die Relation *Act* wird ausschließlich zwischen zwei Objekten bestimmt, alle weiteren werden dazu nicht betrachtet. Damit zwei Objekte in dieser Relation stehen, müssen sie sowohl sich berühren als auch Kräfte aufeinander auswirken. Damit umfasst die Wahrheitsfunktion mehrere Objektbeziehungen, die miteinander verknüpft werden. Somit lässt sich *Act* als `RFR` definieren. Eine Definition niedriger in der Hierarchie der Relationen ist aufgrund der Kombination mehrerer Prädikate nicht möglich. Für die genaue Berechnung des zweiten Teils der Wahrheitsfunktion siehe [Mojtahedzadeh13, Mojtahedzadeh16].
Ausgehend von *Act* wird nun *Support* beschrieben. Diese Relation beschreibt ebenfalls die Beziehung zwischen genau zwei Objekten. In diesem Fall sind aber alle weiteren Objekte für die Berechnung ebenfalls notwendig. Somit kann *Support* als `RBR` Relation beschrieben werden. Die beiden Objekte für die Stabilität stellen die Grundmenge der zweistelligen Relation dar, alle weiteren Objekte werden als Referenzmenge eingegeben. Aufgrund der Notwendigkeit der Referenzmenge ist eine `RFR` nicht möglich. Da aber die in Relation zu stellenden Mengen jeweils nur ein Element hätten, ist eine `MR` nicht erforderlich. Für die spezifische Berechnung der Supportrelation in der Wahrheitsfunktion siehe [Mojtahedzadeh14, Mojtahedzadeh16].

Zusammengesetze Baugruppen

In [Duda01] wird das Zusammenführen einzelner Objekte zu einer gemeinsamen Baugruppe als Semantik aufgeführt. Das dort verwendete Beispiel setzt drei Objekte zu einem geometrischen *Bogen* (bestehend aus zwei Pfeilern und einem Balken) zusammen, wenn diese

sich paarweise berühren und gegenseitig entgegen der Schwerkraft halten. Für diese Relation selbst sind nur die Objekte notwendig, die gemeinsam das Konstrukt bilden sollen. Diese Eigenschaft würde für eine RFR Relation sprechen. Für die Definition der zusammengesetzten Baugruppen ist allerdings die vorgestellte Supportrelation erforderlich. Da RFR keine Relation höher in der Hierarchie verwenden dürfen, muss hier auf eine RBR zurückgegriffen werden. Zwar verwendet die eigentliche Relation nicht die Referenzmenge, eine in der Warhheitsfunktion verwendete benötigt diese Information aber. Somit wird per Definition der Relationen sichergestellt, das keine zu niedrigstufige Relation zur Beschreibung verwendet wird. Die eigentliche Relation lässt sich somit als RBR definieren, die die Supportrelation auf mehrere Objektpaare anwendet. Zusätzlich ist die Information notwendig, dass sich die beiden Pfeiler nicht berühren (da es sonst kein Bogen wäre).

6.4 Fazit

In diesem Kapitel wird die Beschreibung und Verwaltung semantischer Information diskutiert. Dazu wurde zunächst die Notwendigkeit semantischer Information dargelegt, um die Lücke zwischen menschlicher Auffassung einer Szene und der Umweltrekonstruktion zu schließen. Der Stand der Forschung umfasst, welche Arten von Semantik in der Robotik existieren und wie diese extrahiert werden können. Für das extrahierte Wissen gibt es diverse Möglichkeiten mit unterschiedlicher Mächtigkeit, diese zu repräsentieren. In den anschließenden Spezifikations- und Zuordnungsproblemen wird die Schwierigkeit der Auswahl von Semantik beschrieben, da eine abgeschlossene Liste nicht definiert werden kann. Abschließend wurden unterschiedliche Anwendungen sowie der Nutzen von Semantik vorgestellt. Als Fazit wurde festgehalten, dass eine spezifische Beschreibung von konkreter Semantik nicht zielführend ist, unter anderem aufgrund der bereits sehr breiten als auch tiefen Betrachtung von möglicher Information. Daraus entstand die Notwendigkeit einer Definition von Semantik, die unterschiedliche Betrachtungsebenen zulässt. Entscheidend ist dabei Kontextwissen, welches zusammen mit zusätzlichen Daten zu weiterführenden Informationen führt. Dazu wurde ein Konzept entwickelt, wie Relationen ohne Betrachtung einer zeitlichen Komponente beschrieben werden können. Die vorgestellten Relationen bilden eine Hierarchie und bieten die Möglichkeit unterschiedlich komplexe Beziehungen zwischen wiedererkannten Objekten einer Szene zu beschreiben. Diese Vorschriften werden anhand von unterschiedlichen Beispielen erläutert. Somit lässt sich die vierte wissenschaftliche Fragestellung

F4 Inwieweit kann semantische Information definiert und extrahiert werden?

beantworten, indem eine mögliche Beschreibung entwickelt wurde, was unter Semantik im Kontext der Robotik zu verstehen ist und wie sie unterschiedliche Auffassungen abbilden kann. Für die Extraktion wurde ein Konzept vorgestellt, das mit einem hierarchischem Ansatz mehrere Möglichkeiten bietet, Beziehungen von einzelnen und mehreren Objekten zu beschreiben. Anhand der zugrundeliegenden Wahrheitsfunktionen kann die Semantik nach individuellen Beschreibungen extrahiert werden. Limitiert ist die Definition von Semantik dadurch, dass per Annahme stets Szenen- und Kontextwissen kombiniert werden muss und

neue Information nicht aus bestehenden Daten heraus erzeugt werden kann. Die formale Beschreibung mittels Relationen arbeitet ausschließlich objektzentriert, womit szenenspezifische Eigenschaften unabhängig von Objektinstanzen nicht modelliert werden können.
Mögliche Erweiterungen für die Betrachtung der Semantik umfassen unterschiedliche Richtungen: Zum einen kann die Definition der Semantik ergänzt werden, um die vollständige Umweltrepräsentation einzuschließen und nicht nur erkannte Objekte. Zum anderen kann durch das Hinzufügen einer zeitlichen Komponente versucht werden, ein Verständnis über den Ablauf innerhalb der Szene abzuleiten. Eine Herausforderung bleibt dabei die zeitlich diskrete Betrachtung des Arbeitsraums sowie die eindeutige Zuordnung von Objektinstanzen über mehrere Zeitpunkte hinweg. Für eine Anwendung (beispielsweise die Mensch-Roboter-Kollaboration) kann von Interesse sein, welche Semantik nach dem Ausführen einer Operation gültig wäre. Dies kann bei der Planung oder Auswahl von möglichen Aufgaben unterstützen, wenn bestimmte Nachbedingungen nach einer Operation in der Szene gelten sollen. Abschließend ist die Komplexität der Berechnung von Interesse. Da für die Extraktion von Relationen teilweise Mengen mit steigender Anzahl an Objekten als Eingabegröße verwendet werden müssen, wird diese Berechnung mit wachsender Objektanzahl aufwendiger.

Kapitel 7

Gesamtsystem

Inhalt

In diesem Kapitel wird das Gesamtsystem vorgestellt, mit dem die vorgestellten Ansätze umgesetzt und validiert wurden (Abschnitt 7.1). Neben der verwendeten Hardware werden drei unterschiedliche Aufgaben und Anwendungen aus dem Bereich der Robotik vorgestellt: Das sind zum einen klassische Pick-and-Place Aufgaben, die durch die wiedererkannten Objekte ermöglicht werden (Abschnitt 7.2.1). Zum anderen ist es die Erzeugung von Präzedenzgraphen ausgehend von einer inkrementellen Rekonstruktion (Abschnitt 7.2.2) und abschließend die automatische Erstellung von Objektmodellen für die Verwendung in der Objektdatenbank (Abschnitt 7.2.3).

7.1 Aufbau

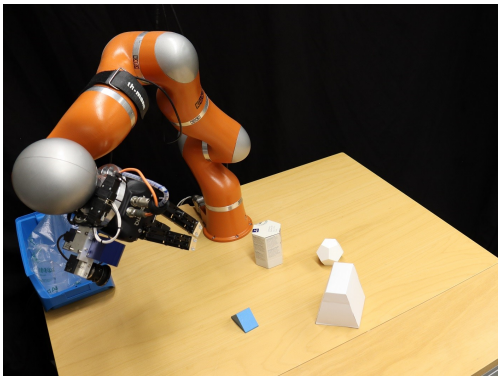
Als Grundlage hinsichtlich der Hardware werden - analog zu den bisherigen Experimenten - ein KUKA LBR IV oder ein Franka Emika Panda verwendet. Beide Roboterarme sind siebenachsige Leichtbauroboter mit Greifern und interner Sensorik. Mittels der in den Gelenken verbauten Sensorik und der bekannten Robotergeometrie kann über die sogenannte Vorwärtskinematik zu jeder Zeit die exakte Pose des Roboters im Raum bestimmt werden. Am LBR IV wurde am Greifer eine ENSENSO N10 Tiefenkamera befestigt, am Franka Emika Panda eine Intel ® Realsense™ D435. Wie in Kapitel 3.3.1 beschrieben, sind die Kameras jeweils auf den entsprechenden *Tool Center Point* (TCP) extrinsisch kalibriert.

7.2 Aufgaben

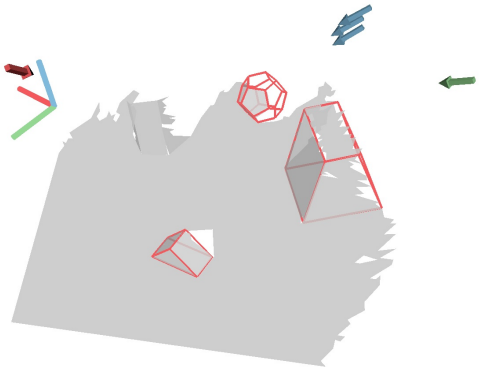
Im Weiteren werden die drei Anwendungen näher erläutert. Das Aufgreifen und Ablegen der Objekte wird beispielhaft an zwei Aufgaben dargelegt. Für das Erzeugen der Präzedenzgraphen und der Objektmodelle wird jeweils ein kurzer Überblick über den Stand der Forschung gegeben und dann der eigentliche Ansatz vorgestellt sowie evaluiert.

7.2.1 Pick-and-Place

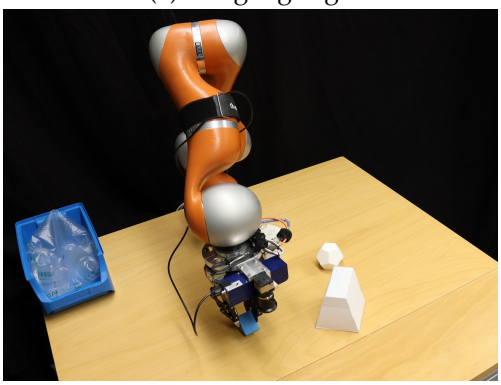
Eine klassische Benchmark-Aufgabe ist die Handhabung von Objekten in der Szene, beispielsweise das Aufnehmen und Ablegen von Objekten mit initial unbekannter Position. Ein konkretes Beispiel für eine *Pick-and-Place* Aufgabe wird in Abbildung 7.1 dargestellt. Ziel ist es dabei, das Objekt *Triangle* in der Szene zu identifizieren, aufzugreifen und an einer festen Position abzulegen. Sobald das Objekt von Interesse erfolgreich rekonstruiert und erkannt wurde, kann dieses vom Roboter gehandhabt werden. Ausgehend von der Wiedererkennung kann eine bereits gegebene (zum Beispiel gespeichert in der Objektdatenbank) oder szenenspezifische Aufgreifpose angefahren werden. Sobald das Objekt gegriffen wurde, kann es mittels einer Transferbewegung zur Zielregion bewegt und schließlich abgelegt werden. Da die Pose sowie die Geometrie des Objekts bekannt ist, kann dieses aus der Umweltrepräsentation anschließend mit den Methoden aus dem Kapitel zu dynamischen Szenen entfernt werden (siehe Abbildung 7.1f). Die verbleibende Lücke im B-Rep kann mit einer der nächsten Sichten geschlossen oder automatisiert vervollständigt werden [Rohner20b].
Eine weitere Beispielaufgabe ist das in Abbildung 7.2 dargestellte Platzieren eines Objektes auf einem anderen Objekt. Dazu müssen zunächst beide Objekte erkannt werden. Mit bekannten Greif- und Ablageposen wird schließlich das Objekt *Peg* auf das *Hole* platziert. Dabei kann das gegriffene Objekt sowohl aus der Szene entfernt als auch an der neuen Position in die Szene eingefügt werden, ohne dass das B-Rep der Umweltrepräsentation ungültig wird (Abbildung 7.2f).



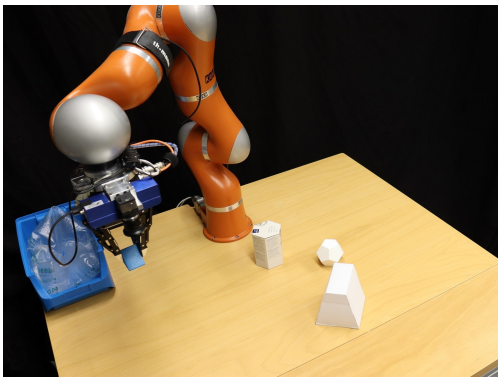
(a) Ausgangslage



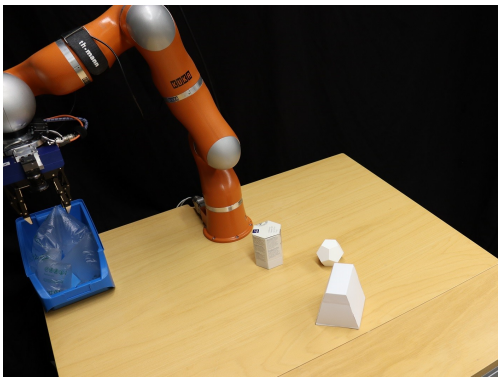
(b) Initiale Rekonstruktion



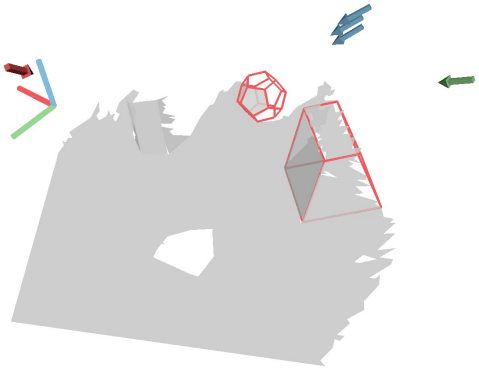
(c) Aufgreifen



(d) Transferbewegung

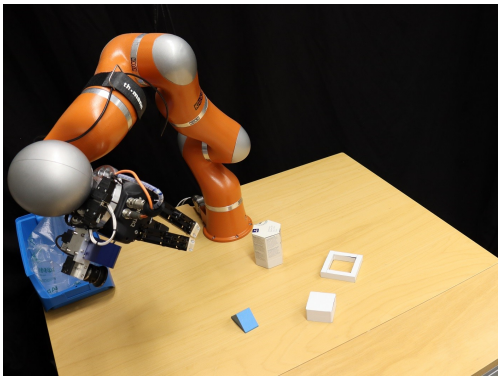


(e) Ablage

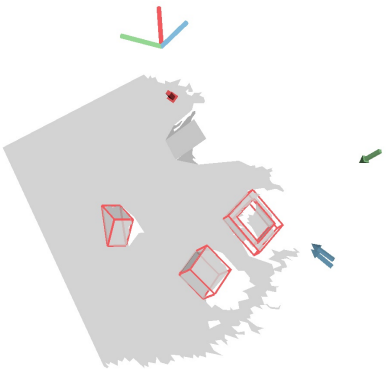


(f) Angepasste Rekonstruktion

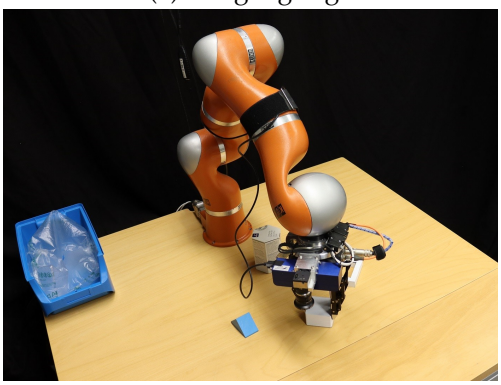
Abbildung 7.1: Wiedererkennen eines bestimmten Objekts mit anschließendem Versetzen und Aktualisierung der Umweltrepräsentation. Zu jeder Rekonstruktion ist zusätzlich das Weltkoordinatensystem in der Roboterbasis dargestellt.



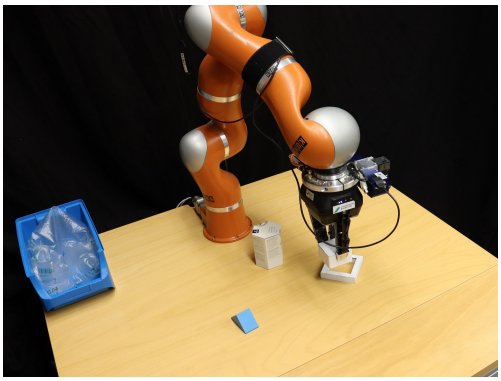
(a) Ausgangslage



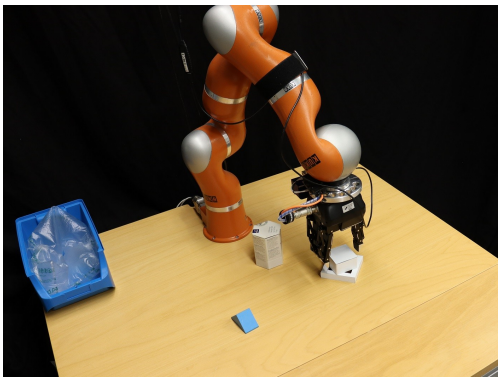
(b) Initiale Rekonstruktion



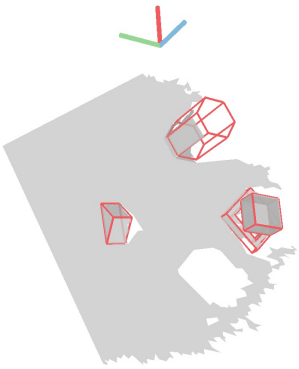
(c) Aufgreifen



(d) Transferbewegung



(e) Ablage



(f) Angepasste Rekonstruktion

Abbildung 7.2: Ausführung einer Pick-and-Place-Aufgabe: Nach erfolgreicher Wiedererkennung kann das Objekt Peg auf das Objekt Hole gestellt werden. Aus der Umweltrepräsentation wird das versetzte Objekt entfernt und an der neuen Position eingefügt.

7.2.2 Präzedenzgraphenerzeugung

Eine weitere Aufgabe, die auf dem erstellten Weltmodell basiert, ist die Erzeugung von Vorrangsbeziehungen (beispielsweise Präzedenzgraphen) für die Mensch-Roboter-Kollaboration [Rohner19a]. Der Ansatz ist es, aus dem bestehenden Weltmodell geometrische Abhängigkeiten zu extrahieren und dadurch eine Ordnung in den platzierten Objekten zu bestimmen.

Stand der Forschung

Für eine effiziente Zusammenarbeit zwischen Mensch und Roboter ist es hilfreich bis notwendig, dass ein gemeinsamer Plan existiert, an dem zusammen gearbeitet wird [Riedelbauch17, Riedelbauch20]. Das manuelle Erzeugen dieser Pläne ist zum einen aufwendig und erfordert zum anderen spezifisches Wissen sowohl über die Aufgabe als auch über die Aufgabenrepräsentation. Das Ziel ist, auf Basis einer rekonstruierten Szene den Arbeitsplan automatisch zu bestimmen. Die notwendige Rekonstruktion der Szene kann dabei von Hand erfolgen (beispielsweise durch einen von Hand geführten Roboter in Gravitationskompensation, wie es mit den hier verwendeten Robotern möglich ist) oder durch eine automatische Rekonstruktion. Anschließend können aus dieser Umweltrepräsentation *AND/OR*-Graphen als Zwischenstufe erzeugt werden, aus denen schließlich Präzedenzgraphen abgeleitet werden können.
Die grundlegende Problemstellung ist somit, eine Reihenfolge zu bestimmen, in der die Bauteile gehandhabt werden müssen (auch bekannt unter *Assembly Sequencing*). Im Rahmen der Robotik kann dies als eine Reihe von kollisionsfreien Operationen betrachtet werden, durch die sich die ursprüngliche Szene wiederherstellen lässt [Jiménez13]. Ebendort werden die Ansätze zur Lösung des Problems in *kombinatorische* und *geometrische* Ansätze differenziert. Kombinatorische Methoden lassen sich weiterhin gemäß der Repräsentation der Modellierung in *explizite* und *implizite* unterteilen [Homem de Mello91]. Bei expliziten Repräsentationen werden Bauteile und Operationen durch Elemente in einem Graphen dargestellt. Dazu können sowohl gerichtete Graphen [De Fazio87, Zhang02] als auch AND/OR-Graphen [Romney95, Thomas03] verwendet werden. Von impliziten Repräsentationen wird dagegen gesprochen, wenn Bedingungen an die valide Montagefolge gestellt [Wolter92] oder Vorrangsbeziehungen (Präzedenzen) und daraus abgeleitete Präzedenzgraphen [De Fazio87, Naphade00] aufgebaut werden. Um aus den gegebenen Präzedenzen und Bedingungen eine Montagefolge zu erzeugen, können übliche Such- und Optimierungsalgorithmen verwendet werden [Martelli73, Bagchi83, Cao98] mit Erweiterungen für Baugruppen aus vielen Objekten [Hong99, Chen08, Guan02]. Neben den kombinatorischen Methoden sind auch geometrische Ansätze möglich. Eine Gruppe sind dabei sogenannte *Assembly-by-Disassembly* Methoden, bei denen die vollständige Baugruppe dekonstruiert werden soll, um daraus mögliche Montagefolgen zu bestimmen [Wilson94, Romney95, Guibas95, Kaufman96, Niu03]. Alternativ zu den *Assembly-by-Disassembly* Ansätzen können potentielle geometrische Kontakte zwischen den einzelnen Bauteilen analysiert werden, um Rückschlüsse auf die verschiedenen Montageschritte zu erlangen [Ji99, Xiao00, Bruyninckx01]. Als dritte geometrische Methode kann für die einzelnen Objekte eine kollisionsfreie Bahn bestimmt werden, indem der Konfigurationsraum randomisiert abgetastet wird [Sundaram01].

Ansatz

Die Eingabe für die Erzeugung von Präzedenzgraphen ist eine vollständige Wiedererkennung aller Objekte in der Welt. Im vorgestellten Verfahren wird der Roboter von Hand durch die Szenen geführt. Da bei der Rekonstruktion einer komplexen Szene Verdeckungen möglich sind, ist es erforderlich, bei Bedarf eine Zwischenrekonstruktion durchzuführen. Sobald alle Objekte platziert sind und die Rekonstruktion vervollständigt ist, wird das in Kapitel 3 vorgestellte Verfahren zur Objektwiedererkennung durchgeführt.
Zur Berechnung der AND/OR-Graphen ist es zunächst erforderlich, zu untersuchen, welche Objekte kollisionsfrei aus der Szene entfernt werden können. Dafür wird für alle Objekt-Paare unter Beachtung der Arbeitsoberfläche untersucht, welche Bewegungsrichtungen (ausgehend vom Zentrum des Objekts) frei oder blockiert sind. Dazu wird für jedes Objektpaar die Minkowski-Differenz bestimmt. Mit Hilfe dieser Differenz werden diskretisierte Richtungen dahingehend geprüft, ob diese frei oder blockiert sind. Mathematisch lässt sich das pro Richtung durch einen Schnitt-Test einer Geraden durch den Ursprung auf einer Möbius-transformierten Riemannschen Kugel darstellen [Thomas03]. Diese Richtungen werden in sogenannten *Disassembly*-Karten verwaltet.
Um die hohe Komplexität der AND/OR-Graph Erzeugung [Kavraki93] zu reduzieren, wird ein Kontaktgraph verwendet, in dem jedes Objekt durch einen Knoten repräsentiert wird. Zwei Knoten sind durch eine Kante verbunden, wenn sich die beiden Objekte in der Szene berühren. Dieser Graph kann ebenfalls mit Hilfe der Minkowski-Differenz berechnet werden. Durch Nutzung einer Toleranz erhöht sich die Robustheit gegenüber Fehlern durch eine unpräzise Rekonstruktion (siehe dazu auch Abschnitt 3.2.3 bezüglich Hypothesenverifikation). Durch die Einschränkung, dass pro Aufbauschritt maximal zwei bereits zusammengehörige Baugruppen beteiligt sein können, reduziert sich, aufgrund der geringeren Anzahl an möglichen Kombinationen, die Komplexität weiter. Montageanweisungen, die drei oder mehrere Objekte beziehungsweise Gruppen verwenden, werden dadurch jedoch ausgeschlossen.
Aus diesem Kontaktgraph und den freien Bewegungsrichtungen kann schließlich ein AND/-OR-Graph erzeugt werden. Gemäß dem *Assembly-by-Disassembly* Ansatz wird von der vollständig aufgebauten Szene ausgegangen. Für jedes Objekt wird zunächst geprüft, ob es aus der Szene entfernt werden kann. Dazu werden alle diskretisierten Richtungen gemäß der eingangs berechneten Disassembly-Karten dahingehend geprüft, ob diese im Vergleich zu allen Objekten in der Szene frei sind. Für jedes Objekt das direkt entfernt werden kann, wird ein Knoten im Graphen eingefügt, der diesen Montageschritt repräsentiert. Dieser Prozess wird rekursiv weitergeführt, indem für die verbleibende Baugruppe geprüft wird, welche Objekte aus dieser direkt entfernt werden können. Falls ein Objekt in der aktuell betrachteten Baugruppe nicht vertreten ist, wird beim Test der freien Richtungen die entsprechende Disassembly-Karte nicht verwendet. Somit lässt sich diese Berechnung fortführen bis in den Knoten nur noch ein einzelnes Objekt vorhanden ist, welches als Blatt im Baum verbleibt. Analog ist die komplett aufgebaute Szene die Wurzel.
Im Rahmen dieser Arbeit war das Vorhandensein einer bekannten Arbeitsfläche gegeben. Diese kann ebenfalls als Blatt und als Komponente in der Baugruppe aufgefasst werden. Gemäß der Disassembly-Karten würde für dieses Objekt keine Richtung frei sein, solange bereits

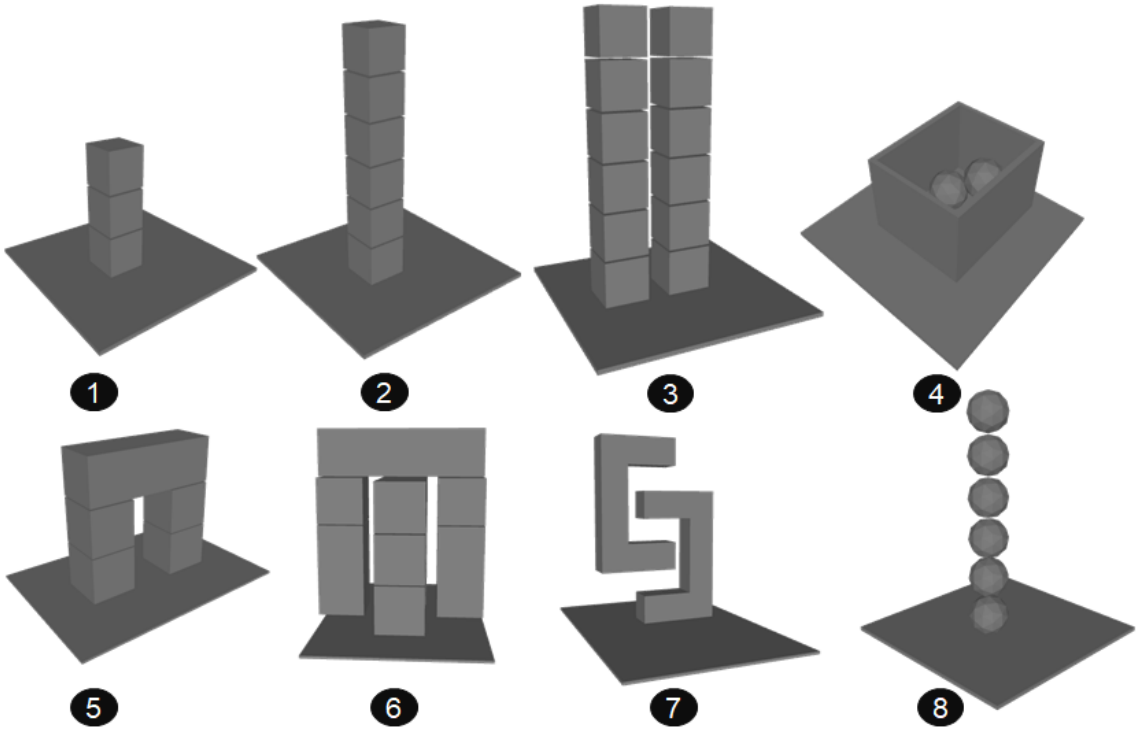
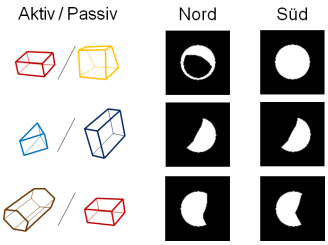


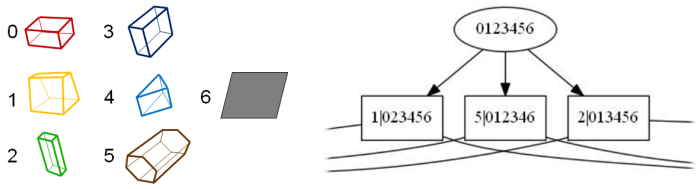
Abbildung 7.3: Evaluation der Präzedenzgraphenerzeugung anhand synthetischer Szenen unterschiedlicher Komplexität, aus [Rohner19a]

Abbildung 7.4: Beipspielhafte Szene zur Präzedenzgraphenerzeugung (links) mit dazugehöriger Rekonstruktion (mitte) und erkannten Objekten (rechts), aus [Rohner19a]

ein Objekt in der Szene steht. Damit ist sichergestellt, dass die Arbeitsfläche vor allen anderen Objekten vorhanden sein muss. An den Aufbau des AND/OR-Graphen können weitere Einschränkungen gestellt werden, beispielsweise spezifische Richtungen, in denen sich der Roboter bewegt. Diese lassen sich durch Einschränkungen in den Disassembly-Karten direkt umsetzen, indem einzelne Bereiche auf blockiert gesetzt werden.
Abschließend kann ein möglicher Präzedenzgraph erzeugt werden. Aus dem AND/OR-Graphen lässt sich ableiten, welche Objekte vor einem anderen platziert werden müssen und welche unabhängig voneinander sind. Diese Vorrangsbeziehungen können in der sogenannten *Präzedenz*-Matrix repräsentiert werden, woraus der eigentliche Präzedenzgraph bestimmt wird.



(a) Disassembly Karten für ausgewählte Objektpaare: Links die untersuchten Objekte, rechts in weiß freie und in schwarz blockierte Richtungen für beide Kugelhälften



(b) Wurzelknoten des resultierenden AND/OR-Graphen: Links die verwendeten Objekte und die Arbeitsoberfläche nummeriert für die im Graphen rechts. Die erste Zahl in den Folgeknoten beschreibt jeweils das Objekt, welches aus der aufgebauten Szene direkt nach oben kollisionsfrei entfernt werden kann.

Abbildung 7.5: Teilschritte der Präzedenzgraphenerzeugung mit den Objekten aus Abbildung 7.4, aus [Rohner19a]

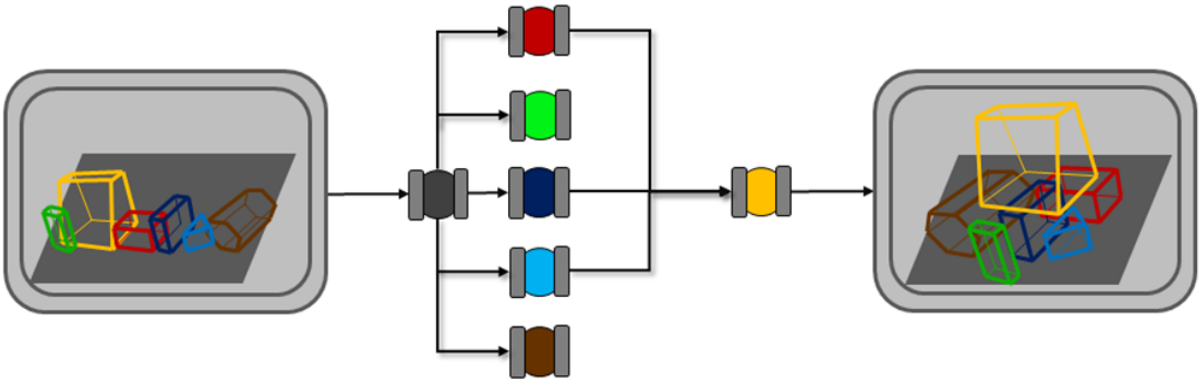


Abbildung 7.6: Resultierender Präzedenzgraph mit eingeschränkten Bewegungsrichtungen nach oben, aus [Rohner19a]

Ergebnisse

Der entwickelte Ansatz wird zunächst auf synthetischen Daten evaluiert, um die grundlegende Korrektheit zu validieren. Die betrachteten synthetischen Szenen sind in Abbildung 7.3 zu sehen. Die modellierten Objekte umfassen dabei sowohl konvexe als auch nicht konvexe Körper mit einer variablen Anzahl an Flächen. Die Szenen unterscheiden sich in der Komplexität gemäß der Anzahl an Objekten als auch in möglichen Bewegungsrichtungen. Als Ergebnis ist festzuhalten, dass die mit dem vorgestellten Verfahren berechneten AND/OR-Graphen sowie die daraus abgeleiteten Präzedenzgraphen korrekt sind und eine gültige Montagefolge abbilden. Bei zu starker Einschränkung der Bewegungsrichtungen kann das Verfahren allerdings scheitern, da diese Einschränkungen dazu führen, dass eigentlich gültige Bewegungsrichtungen verworfen werden. Wenn beispielsweise ausschließlich Bewegungen orthogonal zur Arbeitsfläche zugelassen sind, ist der Aufbau 7 in Abbildung 7.3 nicht lösbar, da die Objekte sich gegenseitig blockieren.
Neben der Validierung wurde der Ansatz an einer konkreten Szene evaluiert (zu sehen in Abbildung 7.4). Die dargestellte Szene (links) wurde in zwei Schritten aufgebaut und rekonstruiert (mitte). Darauf aufbauend konnten die korrekten Hypothesen bestimmt werden (rechts).

Die Rekonstruktion wurde genutzt um die Disassembly-Karten (Abbildung 7.5a) und den Kontaktgraphen zu berechnen. Davon ausgehend konnte der AND/OR-Graph erzeugt werden (Wurzel und erste Ebene in Abbildung 7.5b). Der abschließende Präzedenzgraph ist in Abbildung 7.6 zu sehen. Dabei ist zu erkennen, welche Objekte vor *Sweets3* (gelb dargestellt) platziert werden müssen. Analog gibt es Objekte, die unabhängig vom Rest des Aufbaus sind (*Flat cuboid* in Grün und *Sweets2* in Braun).

7.2.3 Einlernen von Objekten

Als dritte Aufgabe wird eine Möglichkeit vorgestellt, auf welche Art die Modelle in der Objektdatenbank zur Verfügung gestellt werden können. Im einfachsten Fall existiert bereits ein CAD-Modell vom relevanten Werkstück aus dem Fertigungsprozess. Alternativ ist es möglich, mit gängiger CAD-Software ein Oberflächenmodell vom Werkstück zu erstellen. Bei Objekten mit einfacher Geometrie sowie einer geringen Anzahl an Flächen ist diese Methode noch realistisch, bei komplexeren Objekten aber sowohl aufwendig als auch fehleranfällig. Für eine robuste und effiziente Lösung müssen die Modelle automatisiert erzeugt werden. Dazu wird im Weiteren ein Ansatz vorgestellt, der bereits in [Singer21] diskutiert wurde. Die grundlegende Idee ist dabei, Werkstücke automatisiert von allen Seiten zu betrachten und so ein B-Rep zu erzeugen.

Stand der Forschung

Ein Ansatz zum automatisierten Erstellen von Objektmodellen ist der Einsatz spezialisierter Hardware wie Drehteller. Das zu modellierende Objekt kann dann entweder von mehreren fixierten Kameras oder von einer einzelnen, beweglichen Kamera aufgenommen werden [Kasper12, Singh14, Banerjee18]. Der Nachteil dieser Ansätze ist der hohe Hardware-Aufwand, da nur zum Einlernen neue Hardware zur Verfügung stehen muss. Weiterhin kann das Objekt nicht vollständig erfasst werden, da die Unterseite fehlt. Eine Alternative ist das ausschließliche Verwenden von einem bereits vorhandenen Roboterarm und einem Sensor. Um das Objekt vollständig zu erfassen, kann der Manipulator das Objekt selbstständig aufgreifen, drehen und wieder ablegen oder direkt in der Szene bewegen, während der Sensor das Objekt erfasst. Dabei kann der Sensor sowohl am Roboter befestigt sein als auch stationär [Krainin11, Wang15, Bevec15, Fäulhammer17, Venkataraman19]. Falls mehrere Objekte eingelernt werden sollen, kann der Roboterarm diese selbstständig aufgreifen oder aus bestehenden Szenen entfernen und anschließend einlernen. Dabei ist allerdings wieder auf Verdeckungen zu achten: entweder durch die Positionierung des Objekts oder durch den Greifer.
Analog zum bisherigen Aufbau in dieser Arbeit wird ein Leichtbauroboter mit *Eye-in-Hand* Kamera verwendet. Konkret wird ein Franka-Emika-Panda sowie eine Intel ® Realsense™ D435 genutzt.

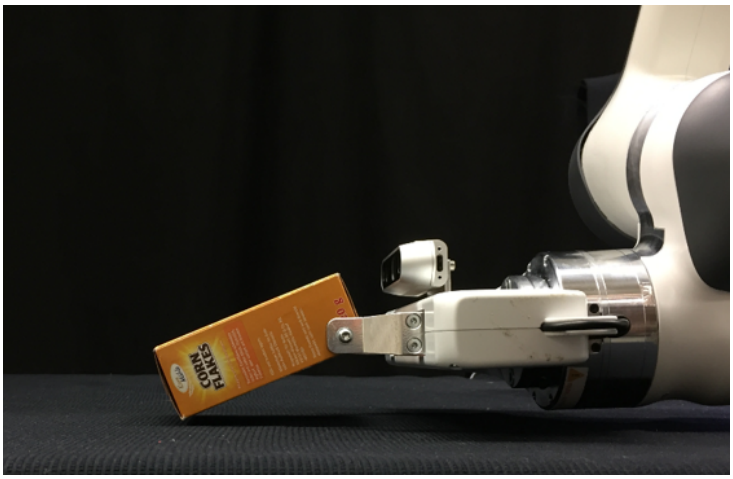


Abbildung 7.7: Greifen und Drehen des Objekts zum Einlernen, nachdem die Oberseite aufgezeichnet wurde. Ebenso ist die verwendete Hardware für das Einlernen zu sehen, aus [Singer21]

Ansatz

Der entwickelte Ansatz beginnt damit, ein Übersichtsbild von der Szene mit dem Tiefensensor aufzunehmen. Unter der Annahme, dass die Arbeitsoberfläche bekannt ist, wird diese zunächst entfernt und das verbleibende Tiefenbild mittels *Connected Components Labeling* [Wu09] segmentiert. Dabei ist das Ziel keine Segmentierung auf Objektebene, sondern ausschließlich die Information, welche Pixel zusammenhängen. Ausgehend von dieser Segmentierung einzelner Flächen wird versucht, ein Objekt zu greifen (in Anlehnung an [Katz13]). Wenn der Griff erfolgreich ist (gemäß Sensorik im Greifer), wird das Objekt zu einer bekannten, festen Vermessposition bewegt. Ist der Griff nicht erfolgreich, wird die Menge an Objekten manipuliert, indem mit dem Greifer ein Objekt in Richtung des Mittelpunktes der Szene bewegt wird, wodurch sich Objekte vereinzeln. Anschließend wird erneut ein Sensoreindruck aufgenommen, ausgewertet und ein Objekt gegriffen. Dieser Prozess wird wiederholt, bis ein Objekt erfolgreich an der Vermessposition angekommen ist. Dort wird das Objekt von allen Seiten aus festen, vordefinierten Posen vermessen. Die Ausgabe ist in diesem Fall eine texturierte Punktwolke pro Sicht. Die aufgenommenen Punktwolken werden mittels ICP-Algorithmus [Arun87] aufeinander registriert. Als Startschätzung wird die bekannte Pose der am Roboter montierten Kamera zum Moment der Aufnahme verwendet. Um den ICP zu unterstützen, besteht die Vermessposition aus einer bekannten, nicht symmetrischen Arbeitsoberfläche. Somit erhält man ein Teilmodell für die aktuelle Oberseite des Objektes.
Anschließend wird das Objekt gegriffen und um 180 Grad gedreht (siehe Abbildung 7.7). Bei flachen Objekten ist zu beachten, dass diese nicht vollständig horizontal gegriffen werden können aufgrund der räumlichen Ausdehnung des Roboterarms und Greifers. Diese Objekte können an den Rand der Arbeitsfläche befördert und dort gegriffen sowie gedreht werden. Danach kann die nun sichtbare Hälfte des Objekts aufgenommen und das zweite Teilmodell aufgezeichnet werden. Abschließend müssen die beiden Teilmodelle zueinander registriert werden. Da die Pose des Objekts nach dem Umdrehen aber nicht mehr gesichert ist (da es beispielsweise kippt), gibt es keine Startschätzung. Die Teilmodelle können somit (siehe dazu

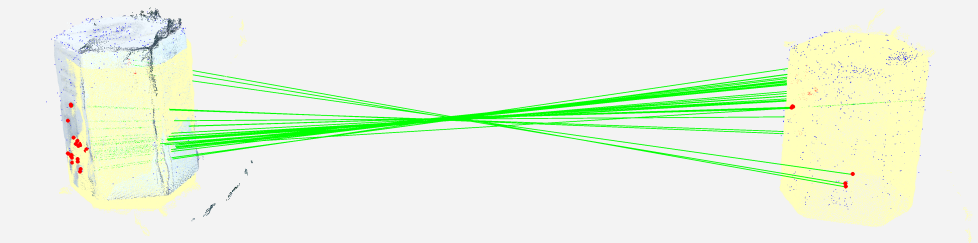
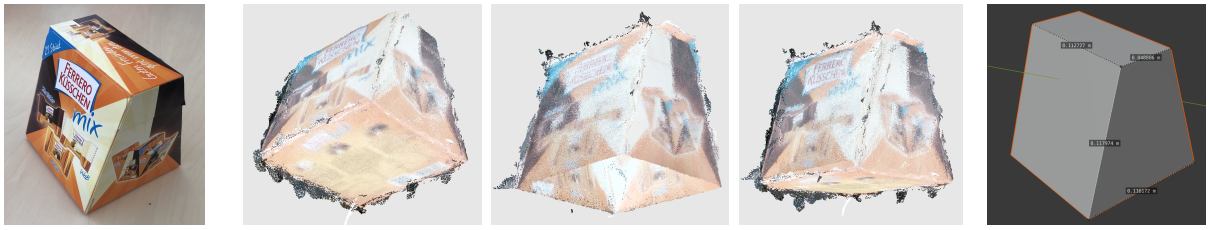


Abbildung 7.8: Zusammenfügen der Teilmodelle für Objekt *Pen*, aus [Singer21]



(a) Objekt Sweets2 (ohne Folie) (b) Unteres Teilmodell (c) Oberes Teilmodell (d) Kombinierte Punktwolke (e) Umgewandeltes B-Rep

Abbildung 7.9: Ergebnisse der Aufnahme des Objektes *Sweets2* (links) bestehend aus den zwei Teilmodellen und der zusammengefügten Punktwolke (mitte) sowie das resultierende B-Rep (rechts), aus [Singer21]

auch Kapitel 3) mittels eines geometrischen oder texturbasierten Ansatzes registriert werden. In diesem Fall werden mittels SIFT Keypoints bestimmt, aus denen durch *geometric consistency grouping* die Transformation zwischen beiden Teilmodellen bestimmt wird [Lowe04] (siehe Abbildung 7.8). Mit dieser Transformation kann abschließend das vollständige Modell bestimmt werden. Um das für die Objektdatenbank zur Verfügung zu stellen, wird abschließend die Umwandlung von Punktwolken in B-Reps darauf angewandt. Sollten Unvollständigkeiten vorliegen, können die bis zu einem gewissen Grad mit den Methoden aus [Sand19,Rohner20b] automatisch vervollständigt werden. Für die Rekonstruktion in ein B-Rep mit einer Intel ® Realsense™ D435 wird eine Strukturgröße von 1cm verwendet.

Ergebnisse

Der gesamte Ablauf wird in Abbildung 7.9 illustriert. Auf haushaltsüblichen Objekten mit Größen von einigen Zentimetern wurde eine Abweichung zur Ground Truth von circa 1 mm erreicht. Durch die Umwandlung in B-Reps wird zudem sensortypisches Rauschen entfernt [Sand19], wodurch die so erzeugten Modelle unabhängig vom verwendeten Sensor sind. Durch die Umwandlung in ein B-Rep geht die Farbinformation - und damit auch die Möglichkeit der Verwendung von farbbasierten Wiedererkennungsmethoden wie SIFT - jedoch verloren. Die erzeugten Modelle wurden anhand des Wiedererkennens von Objekten validiert: zum einen mit dem Ansatz aus Kapitel 3 (siehe Abbildung 7.10), zum anderen mit farbbasierten Merkmalen wie SIFT. Als Limitierung dieses Ansatzes ist die Notwendigkeit von Farbinformation festzuhalten, falls die Pose des Objektes nach dem Umdrehen nicht präzise bekannt ist. Da die beiden Teilmodelle zueinander registriert werden müssen und dies aktuell über SIFT Keypoints berechnet wird, ist eine Registrierung bei texturlosen Objekten

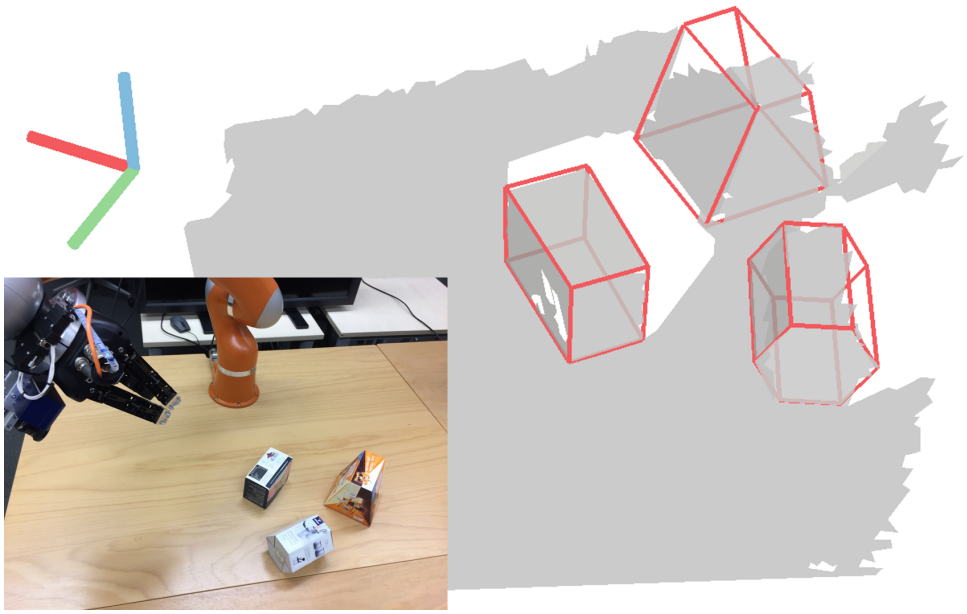


Abbildung 7.10: Wiedererkennung des eingelernten Objekts *Sweets2* in einer Szene (links im Bild) sowie der Rekonstruktion aus einer Perspektive mit erkannten Objekten (rechts), aus [Singer21]

nicht möglich. Für Objekte mit komplexer Geometrie sind oberflächenbasierte Ansätze zur Registrierung eine Alternative.
Neben der eigentlichen Methodik konnte die Invarianz der Ansätze gegenüber unterschiedlichen Tiefenkameras gezeigt werden. Der Großteil der bisherigen Evaluation wurde mit der ENSENSO N10 durchgeführt. Für das Einlernen der Objekte wurde mit der Intel ® Realsense™ eine Tiefenkamera mit geringer Auflösung und kleinerem Basisabstand verwendet.

7.3 Fazit

In diesem Kapitel wurden drei unterschiedliche Aufgaben beziehungsweise Anwendungen vorgestellt, welche mit den entwickelten Methoden dieser Arbeit gelöst werden können.
Als erste Aufgabe wurden klassische Pick-and-Place Aufgaben vorgestellt. Der Fokus liegt dabei nicht auf der Greif- oder Ablageplanung, sondern den Möglichkeiten, die sich aus der verwendeten Objektrepräsentation ergeben. Die Szene kann mit den in Kapitel 3 und 4 entwickelten Ansätzen automatisch rekonstruiert werden, und mit dem Verarbeiten von Änderungen und Störsignalen aus Kapitel 5 können auch dynamische Szenen korrekt in die Umweltrepräsentation eingearbeitet werden. Mit den dafür notwendigen Operationen kann schließlich das Weltmodell valide gehalten werden, wenn der Roboter Objekte in der Szene bewegt, da das Objekt an seinem ursprünglichen Ort aus der Rekonstruktion entfernt und am Ablageort neu eingefügt wird. Da nur Objekte gegriffen werden können, zu denen eine Hypothese vorliegt, ist ein minimaler Eingriff in die Rekonstruktion möglich. Als nächste Anwendung ist die Erzeugung von Präzedenzgraphen für die Mensch-Roboter-Kollaboration möglich. Als Eingabe hierfür ist wieder eine vollständig rekonstruierte Welt notwendig. Für die erkannten Objekte werden blockierende Richtungen berechnet und mittels eines *Assembly-by-Disassembly*

Ansatzes zunächst ein *AND/OR*-Graph und abschließend ein Präzedenzgraph erzeugt. Eine mögliche Anwendung ist die Koordination von mehreren Agenten in der Mensch-Roboter-Kollaboration. Abschließend wird eine Möglichkeit zum Erzeugen von Objektmodellen für eine Objektdatenbank beschrieben. Als Alternative zur händischen Erzeugung wurde ein voll automatisierter Ansatz dargestellt, der die identische Hardware nutzt, die bisher verwendet wurde. Dazu werden Objekte aus einer Menge heraus gegriffen und von mehreren Seiten betrachtet, sodass ein vollständiges Modell entsteht. Dabei werden sowohl Texturinformationen erzeugt, als auch geometrische Modelle, die als Eingabe für die Objektdatenbank in dieser Arbeit geeignet sind.

Kapitel 8

Fazit

Inhalt

Zum Abschluss dieser Arbeit werden die Methoden und Ergebnisse dieser Arbeit zusammengefasst. Im Hinblick auf die wissenschaftlichen Fragestellungen sowie die grundlegende Arbeitshypothese wird ein Fazit gezogen (Abschnitt 8.1). Darauf aufbauend werden abschließend mögliche nächste Schritte diskutiert (Abschnitt 8.2).

8.1 Zusammenfassung und Fazit

Aus der Popularität und dem Einsatz von Robotern im Haushalt sowie in kleinen und mittelständischen Unternehmen ergibt sich die Notwendigkeit, die Umwelt wahrnehmen zu können. Dafür stehen Robotern eine Vielzahl an Sensoren zur Verfügung, die unterschiedliche Zustände erfassen können: über den Zustand des Roboters und eines potentiellen Manipulators hin zum Erfassen der Szene und weiterer Agenten, die mit dieser interagieren. Aus der Notwendigkeit und der Möglichkeit der Wahrnehmung einer Szene lässt sich die Vision des vollständigen Verstehens ebendieser Umwelt ableiten. Ausgehend von unterschiedlichen Herausforderungen, die sowohl die Szene als auch die verwendeten Objekte betreffen, wurden wissenschaftliche Fragestellungen definiert.
Für den eigentlichen Grundansatz sowie die Arbeitshypothese wurde ein allgemeiner Stand der Forschung bezüglich Wahrnehmungsansätzen in der Robotik dargestellt. Durch die Festlegung auf eine einzelne, registrierte *Eye-in-Hand* Kamera sowie planar rekonstruierbare B-Reps als Repräsentationsform wurde die Arbeitshypothese formuliert: Inwieweit unterstützt eine abstrakte Repräsentationsform das Verständnis der Umwelt?
Für ein Verständnis der Umwelt ist zunächst das Wiedererkennen von bekannten Objekten notwendig, widergespiegelt in der ersten wissenschaftlichen Fragestellung.

F1 Inwieweit können geometrisch primitive, untexturierte Objekte wiedererkannt werden?

Dazu wurde in Kapitel 3 ein oberflächenbasierter Wiedererkennungsansatz entwickelt, der unter Verwendung einer Objektdatenbank alle Objekte in einer Szene wiedererkennt. Dazu bedarf es vorab keiner Segmentierung der Szene beziehungsweise der Umwelt. Anhand von linear unabhängigen Flächen kann ein Deskriptor berechnet und mit diesem mögliche Objekthypothesen bestimmt werden. Aus der Menge aller möglichen Hypothesen wird die Teilmenge bestimmt, die die Szene am besten erklärt, gemessen am Überschneidungsgrads aller Flächen. Die Verwendung von B-Reps ist dabei von Vorteil, da aufgrund der geringen Menge an geometrischen Objekten eine ausführliche Suche nach Hypothesen möglich ist. Zusätzlich ist die Berechnung von Qualitätsmaßen für die Hypothesen und den kompletten Erklärungsgrad der Welt durch die Objektinstanzen leicht möglich. Aufgrund des verwendeten Deskriptors reicht je nach Oberflächenkomplexität des Objekts bereits ein sehr geringer Rekonstruktionsgrad, um Objekte korrekt wiederzuerkennen. Limitiert ist der Ansatz vor allem durch die Notwendigkeit von drei sichtbaren, linear unabhängigen Flächen. Bei Objekten mit geringer Oberflächenkomplexität ist das nur aus mehreren Sichten möglich. Zudem steigt die Komplexität mit steigender Anzahl an Flächen sowohl in der Szene als auch der Datenbank quadratisch an, was einen Nutzen für große Szenen und viele Objekte einschränkt. Abschließend muss festgehalten werden, dass Gegenstände, die nicht planar rekonstruierbar sind, nicht wiedererkannt werden können.
Da beim entwickelten Wiedererkennungsansatz häufig mehrere Sichten auf die Szene notwendig sind, wurde in Kapitel 4 ein Verfahren vorgestellt, mit dem für eine gegebene Szene und die dazugehörigen Hypothesen neue Sichten für einen Wissenszugewinn bestimmt werden können. Diese Notwendigkeit wird in der zweiten Fragestellung formuliert:

F2 Inwieweit unterstützen zusätzliche lokale Sichten die Objektwiedererkennung? Inwieweit können neue Sichten szenenspezifisch bestimmt werden?

Die Sichten werden mit dem Ziel erzeugt, die Wiedererkennung aller Objekte sicherzustellen. Dazu werden zum einen Explorationssichten definiert, die gezielte Posen für den Roboter enthalten, um weitere Hypothesen zu erzeugen. Im Gegensatz dazu wurden Validierungssichten eingeführt, die Hypothesen von erkannten Objekten bestätigen. Dieser Ansatz wird durch die verwendeten B-Reps ermöglicht. Zum einen ist berechenbar, an welchen Kanten weitere Flächen existieren, die bisher nicht rekonstruiert wurden (für die Exploration). Zum anderen ist für alle Flächen von Objekthypothesen ableitbar, ob diese in der Szene existieren. Die Sichten werden dabei so bestimmt, dass die gewünschte Information gezielt erfasst werden kann. Für die Exploration wird dabei die Sicht von der ursprünglichen Kante weg rotiert, um angrenzende Flächen rekonstruieren zu können. Für die Validierung hingegen wird direkt auf die entsprechenden Flächen geblickt. Durch die Verwendung von einer Kamera als Sensor werden im Allgemeinen durch eine Sicht mehrere Flächen gleichzeitig wahrgenommen, wodurch sowohl bei der Exploration als auch der Validierung mehrere Flächen durch eine Sicht rekonstruiert werden. Die möglichen Sichten werden abschließend auf Plausibilität und mögliche Kollisionen geprüft. Eine Einschränkung dieses Ansatzes sind Szenen, in denen für eine vollständige Wiedererkennung Sichten notwendig wären, die nicht angefahren werden

können (beispielsweise aufgrund der Robotergeometrie). Weiterhin basiert die Filterung der möglichen Sichten auf Heuristiken, wodurch es auftreten kann, dass zu viele Sichten entfernt und zum anderen Posen berechnet werden, die nicht vom Roboter erreichbar sind oder keine zusätzliche Information tragen.
Im Fall, dass der Roboter mit einem Menschen (oder einem anderen autonom agierenden Roboter) zusammenarbeitet, ist es möglich, dass Szenen sich dynamisch ändern, während diese gerade vom Roboter wahrgenommen werden. Da dies beispielsweise auch während des *Active Vision* Prozesses auftreten kann, motiviert das die dritte Fragestellung, welche in Kapitel 6 beantwortet wird.

F3 Inwieweit können Änderungen an der Szene zwischen zwei Aufnahmen erkannt und verarbeitet werden?

Zunächst wurden alle möglichen Änderungen an Flächen (da diese sämtliche geometrische Information tragen) gemäß der Umweltrekonstruktion dargestellt. Anhand der aktuellen Pose der Tiefenkamera und des bisherigen Umweltmodells kann berechnet werden, was aus der aktuellen Perspektive sichtbar sein müsste. Da keine Punktmengen transformiert werden müssen, sondern lediglich die Eckpunkte aller Flächen, ist diese Operation schnell berechenbar, da sich an der Geometrie der Objekte durch die Transformation nichts ändert und alle Kanten und Flächen direkt abhängig davon sind. Anschließend kann die so berechnete erwartete Rekonstruktion mit der tatsächlichen Sicht verglichen werden. Bei Abweichungen (wenn beispielsweise ein Objekt entfernt oder hinzugefügt wurde) wird dies in die globale Umweltrepräsentation eingearbeitet. Falls die geänderten Flächen mit bestehenden Hypothesen korrespondieren, kann die Umweltrepräsentation durch Anpassung aller zur Hypothese gehörenden Flächen konsistent gehalten werden. Die Vorteile durch die Nutzung von B-Reps zeigen sich sowohl bei der Erkennung als auch beim Verarbeiten der Änderungen. Allerdings ist die Repräsentation durch B-Reps gleichzeitig eine Einschränkung, da beim Entfernen von Objekten mehr Flächen gelöscht werden müssen, um die Datenstruktur valide zu halten, als geometrisch notwendig wäre. Durch die rein geometrische Betrachtung ist es zudem möglich, dass kleine Flächen nach dem Entfernen in der Rekonstruktion verbleiben.
Um das Verständnis der Umwelt zu vervollständigen, ist es notwendig, weiterführende Informationen über die Objekte in einer konkreten Szene zu extrahieren. Dazu muss zunächst analysiert werden, welche Art von semantischer Information relevant ist, wie diese repräsentiert und schließlich aus einer Szene heraus bestimmt werden kann. Zusammengefasst ergibt das die letzte wissenschaftliche Fragestellung.

F4 Inwieweit kann semantische Information definiert und extrahiert werden?

Ein Hindernis bei der Beantwortung dieser Frage ist der bestehende Stand der Forschung. Zum einen ist Semantik ein sehr breit aufgestelltes Thema mit Nutzen in diversen Aufgabenstellungen, sowie einer gezielten Betrachtung für bestimmte Aufgaben. Zum anderen ist Semantik kein eindeutig definierter Begriff, der dadurch für unterschiedliche Informationen verwendet wird. Aus dieser Lage heraus wurde zunächst eine allgemeine Definition von Semantik erarbeitet, die unterschiedliche Grade an Abstraktionsniveau repräsentiert. Von den

bisherigen Ergebnissen dieser Arbeit ausgehend, wurde schließlich ein allgemeines Konzept zur Definition von Beziehungen von und zwischen Objekten vorgeschlagen. Für die Extraktion von Relationen wurde dieses Konzept beispielhaft durchgeführt. Um die Allgemeingültigkeit zu gewährleisten, ist dieses nicht direkt auf B-Reps ausgerichtet. Vorteile durch die Nutzung von B-Reps ergeben sich vor allem dadurch, dass notwendige Operationen direkter ausgeführt werden können als mit anderen Repräsentationsarten wie Punktwolken, beispielsweise bei Kontaktbeziehungen oder Größenbestimmung von Objekten.
Die in dieser Arbeit vorgestellten Methoden wurden abschließend anhand einiger Anwendungen evaluiert. Das sind zunächst grundlegende Pick-and-Place Aufgaben, für die die entsprechenden Objekte in der Szene wiedererkannt werden mussten. Dies erfolgt durch die Kombination des Wiedererkennungsansatzes und *Active Vision*. Mit Hilfe von gegebenen Greif- und Ablageposen in der Objektdatenbank können die Objekte gehandhabt werden. Die Szene wird dabei automatisch angepasst, wenn ein Objekt bewegt wird. Dazu wird dieses an der Greifposition aus der Repräsentation entfernt und am Ablageort neu eingefügt. Eine weitere Anwendung ist die Erzeugung von Präzedenzgraphen. Dazu ist es notwendig, dass in einer Umweltrekonstruktion alle vorhandenen Objekte erkannt werden. Diese Rekonstruktion kann durch manuelles Führen des Roboters im Gravitationskompensationsmodus, oder durch *Active Vison* erzeugt werden. Aus den erkannten Objekten wird über einen *AND/OR*-Graphen als Zwischenschritt schließlich ein Präzedenzgraph mit Anwendung in der Mensch-Roboter-Kollaboration erzeugt. Als letzte Anwendung wurde das Einlernen von unbekannten Objekten durchgeführt. Dazu wird aus einer Menge an Objekten automatisch eines extrahiert und analysiert. Zunächst wird aus mehreren Sichten eine Hälfte des Objektes aufgenommen, anschließend das Objekt gedreht und die zweite Hälfte rekonstruiert. Die beiden so erzeugten Teilmodelle werden zu einem Gesamtmodell zusammengefügt und in ein B-Rep transformiert. Diesem können weitere Informationen wie die Greifpose hinzugefügt werden.
Insgesamt lässt sich somit ein Mehrwert für die Wahl einer abstrakteren Umweltrepräsentation feststellen. Durch die gegebenen Informationen, welche Komponenten zu einer Fläche gehören, wird zum einen sensortypisches Rauschen reduziert und zum anderen Information nicht mehrfach verwaltet, wenn ein Teil der Szene mehrmals rekonstruiert wird. Bei Punktwolken beispielsweise würden bei jeder neuen Rekonstruktion auch immer mehr Punkte erzeugt werden, welche zum Beispiel mittels *Downsampling* zuerst wieder entfernt werden müssten. Darüber hinaus werden manche Operationen vereinfacht beziehungsweise erst ermöglicht, wie beispielsweise in der Hypothesenerzeugung oder beim Verarbeiten von Änderungen. Dabei ist zu beachten, dass das B-Rep während dieser Berechnungen nicht abgetastet werden sollte, um die Vorteile der abstrakteren Repräsentation nicht zu verlieren. Vor allem gegenüber lernbasierten Methoden mit Hilfe von neuronalen Netzen ergeben sich unterschiedliche Vorteile. Allem voran ist das Hinzufügen neuer Objekte leicht möglich: entweder durch ein existierendes CAD-Modell oder die automatische Generierung dieser Daten. Für spezifische, individuell trainierte neuronale Netze würde das Hinzufügen neuer Objekte mit großem Aufwand für das Sammeln von Trainingsdaten sowie dem eigentlichen Training einhergehen. Das ist vor allem im Haushalt oder in kleinen und mittelständischen Unternehmen eine Herausforderung. Große neuronale Netze wie *Foundation Models* können mit einer Vielzahl an

alltäglichen Objekten direkt umgehen, spezifische und einzigartige Objekte sind diesen aber unbekannt. Darüber hinaus haben modellbasierte Ansätze den Vorteil von nachvollziehbaren Erkennungen, während lernbasierte Verfahren Schwierigkeiten mit der Erklärbarkeit von Ergebnissen haben.

8.2 Ausblick

Neben den individuellen Ausblicken jedes Kapitels werden hier mögliche Erweiterungen für die komplette Vision beziehungsweise Fragestellungen gegeben.
Eine erste Möglichkeit ist die Erweiterung der Repräsentation auf allgemeine Formen und nicht nur planare Flächen (beispielsweise [Bloeß22]). Durch diese Erweiterung können weitere Objektgeometrien eindeutig rekonstruiert werden, was eine Anpassung der hier entwickelten Ansätze erfordert. So müsste der Deskriptor des Wiedererkennungsansatzes modifiziert werden, wodurch gegebenenfalls auch die Einschränkung auf drei linear unabhängige Flächen gelöst würde. Durch Änderungen an der Hypothesenerzeugung muss schließlich geprüft werden, wie zusätzliche Sichten gezielt auf die Unterstürzung der Wiedererkennung erzeugt werden können. Im Zusammenhang der dynamischen Szenen muss die Anforderung bezüglich Entfernen von Flächen neu evaluiert werden.
Eine Alternative um die Repräsentation mit weiteren Informationen anzureichern, ist die Verwendung von Textur. Während in dieser Arbeit bewusst möglichst wenig Information verwendet wurde, sind die entwickelten Ansätze ausschließlich auf geometrischer Information an Grenzen gestoßen, beispielsweise bei Verdeckungen. Beim Ergänzen um Farbinformation ist zu beachten, dass diese ähnlich kompakt dargestellt werden sollte wie die Repräsentation mittels B-Reps. Sollte die Textur nur diskretisiert mit hoher Auflösung verfügbar sein, würden die Vorteile der abstrakteren Repräsentation verloren gehen. Analog zur Erweiterung auf nicht planare Flächen sind einige Anpassungen möglich: Für die Wiedererkennung muss die Farb- und Oberflächeninformation kombiniert werden. Dadurch ist es möglich, dass die Anzahl an notwendigen Sichten auf eine Szene reduziert werden kann. Im Rahmen der dynamischen Szenen kann es helfen, geometrisch nicht zu unterscheidenden Flächen sicher zu differenzieren um so die Umweltrepräsentation korrekt zu halten.
Neben der Ergänzung um Textur kann ein *multimodaler* Ansatz entwickelt werden, unter Verwendung von mehr und weiterer Sensorik. Für einen Überblick bezüglich möglicher Sensorik siehe Kapitel 1. In der Robotik sind vor allem LIDAR-Sensoren verbreitet: sowohl in der mobilen Robotik als auch als unterstützender Sensor bei Leichtbaurobotern. Alternativ sind Methoden der *Sichtprüfung* möglich, die sowohl spezialisierte Kamerasysteme als auch akustische Sensoren verwenden. Eine letzte Möglichkeit sind weitere Tiefenkameras, sowohl als *Overhead* als auch *Eye-in-Hand*, um mehr Information in gleicher Zeit aus weiteren Sichten aufzuzeichnen. Bei zeitlicher Synchronisierung können die aufgenommenen Punktwolken gleichzeitig in ein B-Rep umgerechnet und als Eingabe für die Methoden dieser Arbeit verwendet werden. Durch diese zusätzlichen Sensoren kann beispielsweise das Problem der Objektwiedererkennung vereinfacht werden, da pro Objekt mehr Information vorhanden ist. Analog sind für eine Wiedererkennung aller Objekte gegebenenfalls weniger Sichten notwendig.

Schließlich ist eine mögliche Erweiterung, die in dieser Arbeit entwickelten Methoden weiter miteinander zu verknüpfen. Bei der Berechnung von weiteren Sichten kann die Information aus der Handhabung dynamischer Szenen einfließen, um so gezielter nächste Sichten auszuwählen. Mit Hilfe von semantischer Information kann zusätzliches Wissen die Handhabung von Änderungen in der Szene vereinfachen, beispielsweise welche Flächen gelöscht werden müssen. Analog kann die Semantik beim Erzeugen und Auswählen weiterer Sichten unterstützen, indem Zusammenhänge der Szene genutzt werden. Anhand einer konkreten Aufgabe kann gezielt Semantik beschrieben und im Gesamtprozess verwendet werden.

Abbildungsverzeichnis

Tabellenverzeichnis

Quellenverzeichnis

[Abdelaziz24] O. Abdelaziz, M. Shehata & M. Mohamed. *Beyond Traditional Single Object Tracking: A Survey*. arxiv:2405.10439, 2024. Zitiert auf Seite 79.

[Achlioptas20] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny & L. Guibas. *ReferIt3D: Neural Listeners for Fine-Grained 3D Object Identification in Real-World Scenes*. In Computer Vision – ECCV 2020. Springer International Publishing, 2020. Zitiert auf Seite 95.

[Aghajan09] H. Aghajan & A. Cavallaro. Multi-camera networks: Principles and applications. Academic Press, Inc., 2009. Zitiert auf Seite 57.

[Ahn22] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan & A. Zeng. *Do As I Can, Not As I Say: Grounding Language in Robotic Affordances*. arxiv:2204.01691, 2022. Zitiert auf Seite 24.

[Akkaladevi21] S. Akkaladevi, M. Plasch, M. Hofmann & A. Pichler. *Semantic knowledge based reasoning framework for human robot collaboration*. Procedia CIRP, Band 97, 2021. Zitiert auf Seite 98.

[Aksoy15] E. E. Aksoy, M. Tamosiunaite & F. Wörgötter. *Model-free incremental learning of the semantics of manipulation actions*. Robotics and Autonomous Systems, Band 71, 2015. Zitiert auf Seite 96.

[Al Haj11] M. Al Haj, C. Fernández, Z. Xiong, I. Huerta, J. Gonzàlez & X. Roca. Beyond the static camera: Issues and trends in active vision. Springer London, 2011. Zitiert auf Seite 57.

[Aldoma11] A. Aldoma, M. Vincze, N. Blodow, D. Gossow, S. Gedikli, R. B. Rusu & G. Bradski. *CAD-model recognition and 6DOF pose estimation using 3D cues*. In 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 2011. Zitiert auf Seite 38.

[Aldoma12a] A. Aldoma, Z.-C. Marton, F. Tombari, W. Wohlkinger, C. Potthast, B. Zeisl, R. B. Rusu, S. Gedikli & M. Vincze. *Tutorial: Point Cloud Library: Three-Dimensional Object Recognition and 6 DOF Pose Estimation*. IEEE Robotics Automation Magazine, Band 19, 2012. Zitiert auf Seite 38.

[Aldoma12b] A. Aldoma, F. Tombari, R. B. Rusu & M. Vincze. *OUR-CVFH – Oriented, Unique and Repeatable Clustered Viewpoint Feature Histogram for Object Recognition and 6DOF Pose Estimation*. In Pattern Recognition, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. Zitiert auf Seite 38.

[Aldoma16] A. Aldoma, F. Tombari, L. D. Stefano & M. Vincze. *A Global Hypothesis Verification Framework for 3D Object Recognition in Clutter*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Band 38, 2016. Zitiert auf Seite 40.

[Aloimonos88] Y. Aloimonos, I. Weiss & A. Bandyopadhyay. *Active vision*. International Journal of Computer Vision, Band 1, 1988. Zitiert auf Seite 57.

[Arruda16] E. Arruda, J. Wyatt & M. Kopicki. *Active vision for dexterous grasping of novel objects*. In 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2016. Zitiert auf Seite 58.

[Arun87] K. S. Arun, T. S. Huang & S. D. Blostein. *Least-Squares Fitting of Two 3-D Point Sets*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1987. Zitiert auf Seite 117.

[Ataer-Cansizoglu13] E. Ataer-Cansizoglu, Y. Taguchi, S. Ramalingam & T. Garaas. *Tracking an RGB-D Camera Using Points and Planes*. In 2013 IEEE International Conference on Computer Vision Workshops, 2013. Zitiert auf Seite 27.

[Badrinarayanan17] V. Badrinarayanan, A. Kendall & R. Cipolla. *SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Band 39, 2017. Zitiert auf Seite 38.

[Bagchi83] A. Bagchi & A. Mahanti. *Admissible heuristic search in and/or graphs*. Theoretical Computer Science, Band 24, 1983. Zitiert auf Seite 112.

[Bahmanyar15] R. Bahmanyar, A. Murillo Montes de Oca & M. Datcu. *The Semantic Gap: An Exploration of User and Computer Perspectives in Earth Observation Images*. IEEE Geoscience and Remote Sensing Letters, Band 12, 2015. Zitiert auf Seite 95.

[Banerjee18] D. Banerjee, K. Yu & G. Aggarwal. *Robotic Arm Based 3D Reconstruction Test Automation*. IEEE Access, Band 6, 2018. Zitiert auf Seite 116.

[Baral00] C. Baral. *Reasoning about actions: Non-deterministic effects, Constraints, and Qualification*. International Joint Conference on Artificial Intelligence, 2000. Zitiert auf Seite 97.

[Bauer20] D. Bauer, T. Patten & M. Vincze. *VeREFINE: Integrating Object Pose Verification With Physics-Guided Iterative Refinement*. IEEE Robotics and Automation Letters, Band 5, 2020. Zitiert auf Seite 40.

[Bay06] H. Bay, T. Tuytelaars & L. Van Gool. *SURF: Speeded Up Robust Features*. In Computer Vision – ECCV 2006. Springer Berlin Heidelberg, 2006. Zitiert auf Seite 38.

[Benjdira19] B. Benjdira, Y. Bazi, A. Koubaa & K. Ouni. *Unsupervised Domain Adaptation Using Generative Adversarial Networks for Semantic Segmentation of Aerial Images*. Remote Sensing, Band 11, 2019. Zitiert auf Seite 98.

[Benkő01] P. Benkő, R. Martin & T. Varady. *Algorithms for reverse engineering boundary representation models*. Computer-Aided Design, Band 33, 2001. Zitiert auf Seite 27.

[Bennamoun02] M. Bennamoun & G. J. Mamic. Object recognition - fundamentals and case studies. Springer-Verlag, 2002. Zitiert auf Seite 16, 17 und 28.

[Besl88] P. Besl & R. Jain. *Segmentation through variable-order surface fitting*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Band 10, 1988. Zitiert auf Seite 27.

[Bevec15] R. Bevec & A. Ude. *Pushing and grasping for autonomous learning of object models with foveated vision*. In 2015 International Conference on Advanced Robotics (ICAR), 2015. Zitiert auf Seite 116.

[Blodow11] N. Blodow, L. C. Goron, Z.-C. Marton, D. Pangercic, T. Rühr, M. Tenorth & M. Beetz. *Autonomous semantic mapping for robots performing everyday manipulation tasks in kitchen environments*. In 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2011. Zitiert auf Seite 96.

[Bloeß22] J. Bloeß & D. Henrich. *Incremental Online Reconstruction of Locally Quadric Surfaces*. In Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. SciTePress, 2022. Zitiert auf Seite 126.

[Bénière13] R. Bénière, G. Subsol, G. Gesquière, F. Le Breton & W. Puech. *A comprehensive process of reverse engineering from 3D meshes to CAD models*. Computer-Aided Design, Band 45, 2013. Zitiert auf Seite 27.

[Bohg12] J. Bohg, K. Welke, B. León, M. Do, D. Song, W. Wohlkinger, M. Madry, A. Aldóma, M. Przybylski, T. Asfour, H. Martí, D. Kragic, A. Morales & M. Vincze. *Task-based Grasp Adaptation on a Humanoid Robot*. IFAC Proceedings Volumes, Band 45, 2012. Zitiert auf Seite 58.

[Bohg17] J. Bohg, K. Hausman, B. Sankaran, O. Brock, D. Kragic, S. Schaal & G. S. Sukhatme. *Interactive Perception: Leveraging Action in Perception and Perception in Action*. IEEE Transactions on Robotics, Band 33, 2017. Zitiert auf Seite 58.

[Bore18] N. Bore, P. Jensfelt & J. Folkesson. *Multiple Object Detection, Tracking and Long-Term Dynamics Learning in Large 3D Maps*. arxiv:1801.09292, 2018. Zitiert auf Seite 79.

[Bousmalis23] K. Bousmalis, G. Vezzani, D. Rao, C. Devin, A. X. Lee, M. Bauza, T. Davchev, Y. Zhou, A. Gupta, A. Raju, A. Laurens, C. Fantacci, V. Dalibard, M. Zambelli, M. Martins, R. Pevceviciute, M. Blokzijl, M. Denil, N. Batchelor, T. Lampe, E. Parisotto, K. Żołna, S. Reed, S. G. Colmenarejo, J. Scholz, A. Abdolmaleki, O. Groth, J.-B. Regli, O. Sushkov, T. Rothörl, J. E. Chen, Y. Aytar, D. Barker, J. Ortiz, M. Riedmiller, J. T. Springenberg, R. Hadsell, F. Nori & N. Heess. *RoboCat: A Self-Improving Generalist Agent for Robotic Manipulation*. arxiv:2306.11706, 2023. Zitiert auf Seite 25.

[Breazeal09] C. Breazeal. *Role of expressive behaviour for robots that learn from people*. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, Band 364, 2009. Zitiert auf Seite 98.

[Brohan23a] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu & B. Zitkovich. *RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control*. arxiv:2307.15818, 2023. Zitiert auf Seite 24 und 25.

[Brohan23b] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. Ryoo, G. Salazar, P. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu & B. Zitkovich. *RT-1: Robotics Transformer for Real-World Control at Scale*. arxiv:2212.06817, 2023. Zitiert auf Seite 24.

[Bruyninckx01] H. Bruyninckx, T. Lefebvre, L. Mihaylova, E. Staffetti, J. De Schutter & J. Xiao. *A roadmap for autonomous robotic assembly*. In Proceedings of the 2001 IEEE International Symposium on Assembly and Task Planning (ISATP2001), 2001. Zitiert auf Seite 112.

[Buchholz13] D. Buchholz, M. Futterlieb, S. Winkelbach & F. M. Wahl. *Efficient bin-picking and grasp planning based on depth data*. 2013 IEEE International Conference on Robotics and Automation, 2013. Zitiert auf Seite 38.

[Buchholz15] D. Buchholz. *Bin-Picking - New Approaches for a Classical Problem (Griff-in-die-Kiste - Neue Ansätze für ein klassisches Problem)*. Dissertation, Braunschweig, 2015. Zitiert auf Seite 23.

[Bundesen90] C. Bundesen. *A theory of visual attention*. Psychological Review, 1990. Zitiert auf Seite 57.

[Cao98] T. Cao & A. Sanderson. *AND/OR net representation for robotic task sequence planning*. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Band 28, 1998. Zitiert auf Seite 112.

[Carreira17] J. Carreira & A. Zisserman. *Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset*. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. Zitiert auf Seite 96.

[Cencen18] A. Cencen, J. C. Verlinden & J. M. P. Geraedts. *Design Methodology to Improve Human-Robot Coproduction in Small- and Medium-Sized Enterprises*. IEEE/ASME Transactions on Mechatronics, Band 23, 2018. Zitiert auf Seite 26.

[Chen08] W.-C. Chen, P.-H. Tai, W.-J. Deng & L.-F. Hsieh. *A three-stage integrated approach for assembly sequence planning using neural networks*. Expert Systems with Applications, Band 34, 2008. Zitiert auf Seite 112.

[Chen10] T. Chen & M. Schneider. *Modeling Cardinal Directions in the 3D Space with the Objects Interaction Cube Matrix*. In Proceedings of the 2010 ACM Symposium on Applied Computing, 2010. Zitiert auf Seite 96.

[Chen11] S. Chen, Y. Li & N. M. Kwok. *Active vision in robotic systems: A survey of recent developments*. The International Journal of Robotics Research, Band 30, 2011. Zitiert auf Seite 57.

[Chen23a] W. Chen, Y. Li, Z. Tian & F. Zhang. *2D and 3D object detection algorithms from images: A Survey*. Array, Band 19, 2023. Zitiert auf Seite 38 und 39.

[Chen23b] W. Chen, S. Hu, R. Talak & L. Carlone. *Leveraging Large (Visual) Language Models for Robot 3D Scene Understanding*. arxiv:2209.05629, 2023. Zitiert auf Seite 96.

[Cheng18] J. Cheng, H. Cheng, M. Q.-H. Meng & H. Zhang. *Autonomous Navigation by Mobile Robots in Human Environments: A Survey*. In 2018 IEEE International Conference on Robotics and Biomimetics (ROBIO), 2018. Zitiert auf Seite 98.

[Cheng24] G. Cheng, Y. Huang, X. Li, S. Lyu, Z. Xu, H. Zhao, Q. Zhao & S. Xiang. *Change Detection Methods for Remote Sensing in the Last Decade: A Comprehensive Review*. Remote Sensing, Band 16, 2024. Zitiert auf Seite 79.

[Chowdhery23] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov & N. Fiedel. *PaLM: scaling language modeling with pathways*. Journal of Machine Learning Research, Band 24, 2023. Zitiert auf Seite 24.

[Clementini93] E. Clementini, P. D. Felice & P. v. Oosterom. *A Small Set of Formal Topological Relationships Suitable for End-User Interaction*. In Proceedings of the Third International Symposium on Advances in Spatial Databases. Springer-Verlag, 1993. Zitiert auf Seite 96.

[Cohen24] V. Cohen, J. X. Liu, R. Mooney, S. Tellex & D. Watkins. *A Survey of Robotic Language Grounding: Tradeoffs between Symbols and Embeddings*. arxiv:2405.13245, 2024. Zitiert auf Seite 97.

[Connolly85] C. Connolly. *The determination of next best views*. In Proceedings. 1985 IEEE International Conference on Robotics and Automation, Band 2, 1985. Zitiert auf Seite 57.

[Coradeschi03] S. Coradeschi & A. Saffiotti. *An introduction to the anchoring problem*. Robotics and Autonomous Systems, Band 43, 2003. Zitiert auf Seite 97.

[Coradeschi13] S. Coradeschi, A. Loutfi & B. Wrede. *A Short Review of Symbol Grounding in Robotic and Intelligent Systems*. KI, Band 27, 2013. Zitiert auf Seite 97.

[Curless96] B. Curless & M. Levoy. *A volumetric method for building complex models from range images*. In Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '96. Association for Computing Machinery, 1996. Zitiert auf Seite 27.

[Cusumano-Towner11] M. Cusumano-Towner, A. Singh, S. Miller, J. F. O'Brien & P. Abbeel. *Bringing clothing into desired configurations with limited perception*. In 2011 IEEE International Conference on Robotics and Automation, 2011. Zitiert auf Seite 58.

[de Berg08] M. de Berg, M. van Kreveld, M. Overmars & O. Cheong. Computational geometry: Algorithms and applications. Springer-Verlag, third edition, 2008. Zitiert auf Seite 32.

[De Fazio87] T. De Fazio & D. Whitney. *Simplified generation of all mechanical assembly sequences*. IEEE Journal on Robotics and Automation, Band 3, 1987. Zitiert auf Seite 112.

[Dehghani23] M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. P. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin, R. Jenatton, L. Beyer, M. Tschannen, A. Arnab, X. Wang, C. Riquelme Ruiz, M. Minderer, J. Puigcerver, U. Evci, M. Kumar, S. V. Steenkiste, G. F. Elsayed, A. Mahendran, F. Yu, A. Oliver, F. Huot, J. Bastings, M. Collier, A. A. Gritsenko, V. Birodkar, C. N. Vasconcelos, Y. Tay, T. Mensink, A. Kolesnikov, F. Pavetic, D. Tran, T. Kipf, M. Lucic, X. Zhai, D. Keysers, J. J. Harmsen & N. Houlsby. *Scaling Vision Transformers to 22 Billion Parameters*. In Proceedings of the 40th International Conference on Machine Learning, Band 202 of *Proceedings of Machine Learning Research*. PMLR, 2023. Zitiert auf Seite 24.

[Denker13] K. Denker, D. Hagel, J. Raible, G. Umlauf & B. Hamann. *On-Line Reconstruction of CAD Geometry*. In 2013 International Conference on 3D Vision - 3DV 2013, 2013. Zitiert auf Seite 27.

[Derner21] E. Derner, C. Gomez, A. C. Hernandez, R. Barber & R. Babuška. *Change detection using weighted features for image-based localization*. Robotics and Autonomous Systems, Band 135, 2021. Zitiert auf Seite 79.

[Ding22] Y. Ding. *Fast Perception-Action Loops with Proximity Sensors for Robotic Manipulators*. Dissertation, 2022. Zitiert auf Seite 24.

[Drews13a] P. Drews, S. C. da Silva Filho, L. F. Marcolino & P. Núñez. *Fast and adaptive 3D change detection algorithm for autonomous robots based on Gaussian Mixture Models*. In 2013 IEEE International Conference on Robotics and Automation, 2013. Zitiert auf Seite 79.

[Drews13b] P. Drews, L. Manso, S. da Silva Filho & P. Núñez. *Improving change detection using Vertical Surface Normal Histograms and Gaussian Mixture Models in structured environments*. In 2013 16th International Conference on Advanced Robotics (ICAR), 2013. Zitiert auf Seite 79.

[Driess23] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch & P. Florence.

PaLM-E: an embodied multimodal language model. In Proceedings of the 40th International Conference on Machine Learning, 2023. Zitiert auf Seite 24.

[Duda01] R. O. Duda, P. E. Hart, D. G. Stork, C. R. O. Duda, P. E. Hart & D. G. Stork. *Pattern Classification, 2nd Ed*, 2001. Zitiert auf Seite 43 und 105.

[Dunn09] E. Dunn & J.-M. Frahm. *Next Best View Planning for Active Model Improvement*. British Machine Vision Conference, BMVC 2009 - Proceedings, 2009. Zitiert auf Seite 57.

[Elzaiady17] M. E. Elzaiady & A. Elnagar. *Next-best-view planning for environment exploration and 3D model construction*. In 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), 2017. Zitiert auf Seite 57.

[Enser03] P. Enser & C. Sandom. *Towards a Comprehensive Survey of the Semantic Gap in Visual Image Retrieval*. In Image and Video Retrieval. Springer Berlin Heidelberg, 2003. Zitiert auf Seite 95.

[Fang19] C. Fang, X. Ding, C. Zhou & N. Tsagarakis. *A2ML: A general human-inspired motion language for anthropomorphic arms based on movement primitives*. Robotics and Autonomous Systems, Band 111, 2019. Zitiert auf Seite 79 und 98.

[Feng19] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, F. Duffhauss, C. Gläser, W. Wiesbeck & K. Dietmayer. *Deep Multi-modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges*. arxiv:1902.07830, 2019. Zitiert auf Seite 98.

[Fischer15] J. Fischer. *A user-oriented, comprehensive system for the 6 DoF recognition of arbitrary rigid household objects*. Dissertation, Fraunhofer-Verlag, 2015. Zitiert auf Seite 23.

[Fischler81] M. A. Fischler & R. C. Bolles. *Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography*. Commun. ACM, Band 24, 1981. Zitiert auf Seite 27.

[Freeman75] J. Freeman. *The modelling of spatial relations*. Computer Graphics and Image Processing, Band 4, 1975. Zitiert auf Seite 96.

[Fäulhammer17] T. Fäulhammer, R. Ambruş, C. Burbridge, M. Zillich, J. Folkesson, N. Hawes, P. Jensfelt & M. Vincze. *Autonomous Learning of Object Models on a Mobile Robot*. IEEE Robotics and Automation Letters, Band 2, 2017. Zitiert auf Seite 116.

[Gandhi17] D. Gandhi, L. Pinto & A. Gupta. *Learning to fly by crashing*. In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017. Zitiert auf Seite 98.

[Gao21] R. Gao, Y.-Y. Chang, S. Mall, L. Fei-Fei & J. Wu. *Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations*. arxiv:2109.07991, 2021. Zitiert auf Seite 95.

[Garg20] S. Garg, N. Sünderhauf, F. Dayoub, D. Morrison, A. Cosgun, G. Carneiro, Q. Wu, T.-J. Chin, I. D. Reid, S. Gould, P. Corke & M. Milford. *Semantics for Robotic Mapping, Perception and Interaction: A Survey*. Foundations and Trends in Robotics, Band 8, 2020. Zitiert auf Seite 95, 96, 97, 98 und 100.

[Garland01] M. Garland, A. Willmott & P. Heckbert. *Hierarchical Face Clustering on Polygonal Meshes*. Proceedings of the Symposium on Interactive 3D Graphics, 2001. Zitiert auf Seite 27.

[Garrett21] C. R. Garrett, R. Chitnis, R. Holladay, B. Kim, T. Silver, L. P. Kaelbling & T. Lozano-Pérez. *Integrated task and motion planning*. Annual review of control, robotics, and autonomous systems, Band 4, 2021. Zitiert auf Seite 96.

[Giambattista16] A. Giambattista, L. Teixeira, H. Ayanoğlu, M. Saraiva & E. Duarte. *Expression of Emotions by a Service Robot: A Pilot Study*. In Design, User Experience, and Usability: Technological Contexts. Springer International Publishing, 2016. Zitiert auf Seite 98.

[Gibson79] J. J. Gibson. The ecological approach to visual perception. Houghton Mifflin, 1979. Zitiert auf Seite 96.

[Ginsberg88a] M. L. Ginsberg & D. E. Smith. *Reasoning about action I: A possible worlds approach*. Artificial Intelligence, Band 35, 1988. Zitiert auf Seite 97.

[Ginsberg88b] M. L. Ginsberg & D. E. Smith. *Reasoning about action II: The qualification problem*. Artificial Intelligence, Band 35, 1988. Zitiert auf Seite 97.

[Ginting24] M. Ginting, S.-K. Kim, D. Fan, M. Palieri, M. Kochenderfer & A.-a. Agha-Mohammadi. *SEEK: Semantic Reasoning for Object Goal Navigation in Real World Inspection Tasks*. arXiv:2405.09822, 2024. Zitiert auf Seite 98.

[Gratal10] X. Gratal, J. Bohg, M. Björkman & D. Kragic. *Scene Representation and Object Grasping Using Active Vision*. In IROS'10 Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics, 2010. Zitiert auf Seite 58.

[Grotz21] M. Grotz. *Active Vision for Scene Understanding*. Dissertation, KIT Scientific Publishing, 2021. Zitiert auf Seite 23 und 57.

[Guan02] Q. Guan, J. H. Liu & Y. F. Z. and. *A concurrent hierarchical evolution approach to assembly process planning*. International Journal of Production Research, Band 40, 2002. Zitiert auf Seite 112.

[Guibas95] L. Guibas, D. Halperin, H. Hirukawa, J.-C. Latombe & R. Wilson. *A simple and efficient procedure for polyhedral assembly partitioning under infinitesimal motions*. In Proceedings of 1995 IEEE International Conference on Robotics and Automation, Band 3, 1995. Zitiert auf Seite 112.

[Harnad01] S. Harnad. *What's Wrong and Right About Searle's Chinese Room Argument?* Essays on Searle's Chinese Room Argument, 2001. Zitiert auf Seite 97.

[Harnad90] S. Harnad. *The symbol grounding problem*. Physica D: Nonlinear Phenomena, Band 42, 1990. Zitiert auf Seite 97.

[Hayes81] P. Hayes. *The Frame Problem and Related Problems in Artificial Intelligence*. In Readings in Artificial Intelligence. Morgan Kaufmann, 1981. Zitiert auf Seite 97.

[He17] W. He, Z. Li & C. L. P. Chen. *A survey of human-centered intelligent robots: issues and challenges*. IEEE/CAA Journal of Automatica Sinica, Band 4, 2017. Zitiert auf Seite 98.

[Hebel13] M. Hebel, M. Arens & U. Stilla. *Change detection in urban areas by object-based analysis and on-the-fly comparison of multi-view ALS data*. ISPRS Journal of Photogrammetry and Remote Sensing, Band 86, 2013. Zitiert auf Seite 79.

[Hernandez94] D. Hernandez. Qualitative representation of spatial knowledge. Springer Science & Business Media, 1994. Zitiert auf Seite 96.

[Holz13] D. Holz, M. Nieuwenhuisen, D. Droeschel, J. Stückler, A. Berner, J. Li, R. Klein & S. Behnke. *Active Recognition and Manipulation for Mobile Robot Bin Picking*. Springer Tracts in Advanced Robotics, Band 94, 2013. Zitiert auf Seite 58.

[Homem de Mello91] L. Homem de Mello & A. Sanderson. *Representations of mechanical assembly sequences*. IEEE Transactions on Robotics and Automation, Band 7, 1991. Zitiert auf Seite 112.

[Hong99] D. Hong & H. Cho. *Generation of robotic assembly sequences using a simulated annealing*. In Proceedings 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems, Band 2, 1999. Zitiert auf Seite 112.

[Hoseini22] P. Hoseini, S. K. Paul, M. Nicolescu & M. Nicolescu. *A one-shot next best view system for active object recognition*. Applied Intelligence, Band 52, 2022. Zitiert auf Seite 58.

[Huang03] J. Huang & C.-H. Menq. *Automatic CAD Model Reconstruction from Multiple Point Clouds for Reverse Engineering*. Journal of Computing and Information Science in Engineering, Band 2, 2003. Zitiert auf Seite 27.

[Huang22] R. Huang, Y. Xu, L. Hoegner & U. Stilla. *Semantics-aided 3D change detection on construction sites using UAV-based photogrammetric point clouds*. Automation in Construction, Band 134, 2022. Zitiert auf Seite 79.

[Ilie14] A. Ilie & G. Welch. *Online control of active camera networks for computer vision tasks*. ACM Transactions on Sensor Networks, Band 10, 2014. Zitiert auf Seite 57.

[Isler16] S. Isler, R. Sabzevari, J. Delmerico & D. Scaramuzza. *An information gain formulation for active volumetric 3D reconstruction*. In 2016 IEEE International Conference on Robotics and Automation (ICRA), 2016. Zitiert auf Seite 57.

[Izadi11] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison & A. Fitzgibbon. *KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera*. UIST'11 - Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, 2011. Zitiert auf Seite 79.

[Ji99] X. Ji & J. Xiao. *Automatic generation of high-level contact state space*. In Proceedings 1999 IEEE International Conference on Robotics and Automation, Band 1, 1999. Zitiert auf Seite 112.

[Jiménez13] P. Jiménez. *Survey on assembly sequencing: a combinatorial and geometrical perspective*. Journal of Intelligent Manufacturing, Band 24, 2013. Zitiert auf Seite 112.

[Kaelbling11] L. P. Kaelbling & T. Lozano-Pérez. *Hierarchical task and motion planning in the now*. In 2011 IEEE International Conference on Robotics and Automation, 2011. Zitiert auf Seite 95.

[Kartmann21] R. Kartmann, D. Liu & T. Asfour. *Semantic Scene Manipulation Based on 3D Spatial Object Relations and Language Instructions*. In 2020 IEEE-RAS 20th international conference on humanoid robots, 2021. Zitiert auf Seite 98.

[Kasper12] A. Kasper, Z. Xue & R. Dillmann. *The KIT object models database: An object model database for object recognition, localization and manipulation in service robotics*. The International Journal of Robotics Research, Band 31, 2012. Zitiert auf Seite 116.

[Katz03] S. Katz & A. Tal. *Hierarchical mesh decomposition using fuzzy clustering and cuts*. ACM Transaction on Graphics, Band 22, 2003. Zitiert auf Seite 27.

[Katz13] D. Katz, M. Kazemi, J. A. Bagnell & A. Stentz. *Clearing a pile of unknown objects using interactive perception*. In 2013 IEEE International Conference on Robotics and Automation, 2013. Zitiert auf Seite 117.

[Kaufman96] S. Kaufman, R. Wilson, R. Jones, T. Calton & A. Ames. *The Archimedes 2 mechanical assembly planning system*. In Proceedings of IEEE International Conference on Robotics and Automation, Band 4, 1996. Zitiert auf Seite 112.

[Kavraki93] L. Kavraki, J.-C. Latombe & R. H. Wilson. *On the complexity of assembly partitioning*. Information Processing Letters, Band 48, 1993. Zitiert auf Seite 113.

[Ke04] Y. Ke & R. Sukthankar. *PCA-SIFT: a more distinctive representation for local image descriptors*. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Band 2, 2004. Zitiert auf Seite 38.

[Keller13] M. Keller, D. Lefloch, M. Lambers, S. Izadi, T. Weyrich & A. Kolb. *Real-Time 3D Reconstruction in Dynamic Scenes Using Point-Based Fusion*. 2013 International Conference on 3D Vision, 2013. Zitiert auf Seite 27.

[Kim02] Y. J. Kim, M. A. Otaduy, M. C. Lin & D. Manocha. *Fast penetration depth computation for physically-based animation*. In Proceedings of the 2002 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, 2002. Zitiert auf Seite 43.

[Kim07] H. Y. Kim & S. A. de Araújo. *Grayscale Template-Matching Invariant to Rotation, Scale, Translation, Brightness and Contrast*. In Advances in Image and Video Technology. Springer Berlin Heidelberg, 2007. Zitiert auf Seite 38.

[Klasing10] K. Klasing. *Aspects of 3D Perception, Abstraction, and Interpretation in Autonomous Mobile Robotics*. Dissertation, München, 2010. Zitiert auf Seite 23.

[Kollmitz21] M. Kollmitz. *Perception and learning for mobile robots in populated environments*. Dissertation, Freiburg, 2021. Zitiert auf Seite 24.

[Koppula11] H. S. Koppula, A. Anand, T. Joachims & A. Saxena. *Labeling 3D scenes for Personal Assistant Robots*. arxiv:1106.5551, 2011. Zitiert auf Seite 38.

[Korman13] S. Korman, D. Reichman, G. Tsur & S. Avidan. *FasT-Match: Fast Affine Template Matching*. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2013. Zitiert auf Seite 38.

[Kostavelis17] I. Kostavelis & A. Gasteratos. *Robots in Crisis Management: A Survey*. In Information Systems for Crisis Response and Management in Mediterranean Countries. Springer International Publishing, 2017. Zitiert auf Seite 98.

[Kragic02] D. Kragic, H. Christensen & F. A. *Survey on Visual Servoing for Manipulation*. Computational Vision and Perception Laboratory, Band 15, 2002. Zitiert auf Seite 79.

[Krainin11] M. Krainin, P. Henry, X. Ren & D. Fox. *Manipulator and object tracking for in-hand 3D object modeling*. The International Journal of Robotics Research, Band 30, 2011. Zitiert auf Seite 116.

[Krizhevsky12] A. Krizhevsky, I. Sutskever & G. E. Hinton. *ImageNet Classification with Deep Convolutional Neural Networks*. Advances in Neural Information Processing Systems. Curran Associates Inc., 2012. Zitiert auf Seite 38.

[Krüger11] N. Krüger, C. Geib, J. Piater, R. Petrick, M. Steedman, F. Wörgötter, A. Ude, T. Asfour, D. Kraft, D. Omrčen, A. Agostini & R. Dillmann. *Object–action complexes: Grounded abstractions of sensory–motor processes*. Robotics and Autonomous Systems, Band 59, 2011. Zitiert auf Seite 96.

[Kunze11] L. Kunze, T. Roehm & M. Beetz. *Towards semantic robot description languages*. In 2011 IEEE International Conference on Robotics and Automation, 2011. Zitiert auf Seite 95 und 96.

[Kunze14] L. Kunze, C. Burbridge, M. Alberti, A. Thippur, J. Folkesson, P. Jensfelt & N. Hawes. *Combining top-down spatial reasoning and bottom-up object class recognition for scene understanding*. In 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2014. Zitiert auf Seite 96.

[Kunze18] L. Kunze, H. Karaoguz, J. Young, F. Jovan, J. Folkesson, P. Jensfelt & N. Hawes. *SOMA: A framework for understanding change in everyday environments using semantic object maps*. Reasoning and Learning in Real-World Systems for Long-Term Autonomy - AAAI 2018 Fall Symposium, 2018. Zitiert auf Seite 79.

[Kyrkou20] C. Kyrkou. *Imitation-Based Active Camera Control with Deep Convolutional Neural Network*. In 2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS), 2020. Zitiert auf Seite 57.

[Lai18] Y.-H. Lai & S.-H. Lai. *Emotion-Preserving Representation Learning via Generative Adversarial Network for Multi-View Facial Expression Recognition*. In 2018 13th IEEE International Conference on Automatic Face and Gesture Recognition, 2018. Zitiert auf Seite 98.

[Langer17] E. Langer, B. Ridder, M. Cashmore, D. Magazzeni, M. Zillich & M. Vincze. *On-the-fly detection of novel objects in indoor environments*. IEEE International Conference on Robotics and Biomimetics (ROBIO), 2017. Zitiert auf Seite 58.

[Langer20] E. Langer, T. Patten & M. Vincze. *Robust and Efficient Object Change Detection by Combining Global Semantic Information and Local Geometric Verification*. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020. Zitiert auf Seite 79.

[Le17] T. Le & Y. Duan. *A primitive-based 3D segmentation algorithm for mechanical CAD models*. Computer Aided Geometric Design, Band 52-53, 2017. Zitiert auf Seite 27.

[Lemaignan17] S. Lemaignan, M. Warnier, E. A. Sisbot, A. Clodic & R. Alami. *Artificial cognition for social human–robot interaction: An implementation*. Artificial Intelligence, Band 247, 2017. Zitiert auf Seite 95 und 96.

[Li09] C. Li, J. Lu, C. Yin & L. Ma. *Qualitative Spatial Representation and Reasoning in 3D Space*. In International Conference on Intelligent Computation Technology and Automation, Band 1, 2009. Zitiert auf Seite 96.

[Li11] Y. Li, X. Wu, Y. Chrysathou, A. Sharf, D. Cohen-Or & N. J. Mitra. *GlobFit: consistently fitting primitives by discovering global relations*. ACM Transactions on Graphics, Band 30, 2011. Zitiert auf Seite 27.

[Li16] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. M. Snoek & A. D. Bimbo. *Socializing the Semantic Gap: A Comparative Survey on Image Tag Assignment, Refinement, and Retrieval*. ACM Computing Surveys, Band 49, 2016. Zitiert auf Seite 95.

[Li21] C. Li, B. Li, R. Wang & X. Zhang. *A survey on visual servoing for wheeled mobile robots*. International Journal of Intelligent Robotics and Applications, Band 5, 2021. Zitiert auf Seite 79.

[Li22] J. Li & H. Mao. *The Difficulties in Symbol Grounding Problem and the Direction for Solving It*. Philosophies, Band 7, 2022. Zitiert auf Seite 97.

[Lin94] F. Lin & R. Reiter. *State Constraints Revisited 1*. Band 4, 1994. Zitiert auf Seite 97.

[Litomisky13] K. Litomisky & B. Bhanu. *Removing Moving Objects from Point Cloud Scenes*. In Advances in Depth Image Analysis and Applications. Springer Berlin Heidelberg, 2013. Zitiert auf Seite 79.

[Liu07] Y. Liu, D. Zhang, G. Lu & W.-Y. Ma. *A survey of content-based image retrieval with high-level semantics*. Pattern Recognition, Band 40, 2007. Zitiert auf Seite 95.

[Liu22] W. Liu, C. Paxton, T. Hermans & D. Fox. *StructFormer: Learning Spatial Structure for Language-Guided Semantic Rearrangement of Novel Objects*. In 2022 International Conference on Robotics and Automation (ICRA). IEEE Press, 2022. Zitiert auf Seite 98.

[Liu23] W. Liu, A. Daruna, M. Patel, K. Ramachandruni & S. Chernova. *A survey of Semantic Reasoning frameworks for robotic systems*. Robotics and Autonomous Systems, Band 159, 2023. Zitiert auf Seite 95 und 98.

[Loquercio18] A. Loquercio, A. I. Maqueda, C. R. del Blanco & D. Scaramuzza. *DroNet: Learning to Fly by Driving*. IEEE Robotics and Automation Letters, Band 3, 2018. Zitiert auf Seite 98.

[Lowe04] D. G. Lowe. *Distinctive Image Features from Scale-Invariant Keypoints*. International Journal of Computer Vision, Band 60, 2004. Zitiert auf Seite 38 und 118.

[Lowe99] D. Lowe. *Object recognition from local scale-invariant features*. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Band 2, 1999. Zitiert auf Seite 38.

[Luperto16] M. Luperto & F. Amigoni. *Exploiting Structural Properties of Buildings Towards General Semantic Mapping Systems*. In Intelligent Autonomous Systems 13. Springer International Publishing, 2016. Zitiert auf Seite 96.

[Mäntylä87] M. Mäntylä. An introduction to solid modeling. Computer Science Press, Inc., 1987. Zitiert auf Seite 28.

[Marques25] J. M. C. Marques, N. Dengler, T. Zaenker, J. Mucke, S. Wang, M. Bennewitz & K. Hauser. *Map Space Belief Prediction for Manipulation-Enhanced Mapping*. arxiv:2502.20606, 2025. Zitiert auf Seite 58.

[Martelli73] A. Martelli & U. Montanari. *Additive AND/OR graphs*. In Proceedings of the 3rd International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc., 1973. Zitiert auf Seite 112.

[Maturana15] D. Maturana & S. Scherer. *VoxNet: A 3D Convolutional Neural Network for real-time object recognition*. In 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015. Zitiert auf Seite 39.

[McCarthy63] J. McCarthy. *Situations, actions, and causal laws*. Technical report, Stanford Univ CA Department of Computer Science, 1963. Zitiert auf Seite 97.

[McCarthy80] J. McCarthy. *Circumscription—A form of non-monotonic reasoning*. Artificial Intelligence, Band 13, 1980. Zitiert auf Seite 97.

[McCarthy81] J. McCarthy & P. Hayes. *Some Philosophical Problems from the Standpoint of Artificial Intelligence*. In Readings in Artificial Intelligence. Morgan Kaufmann, 1981. Zitiert auf Seite 97.

[McCormac16] J. McCormac, A. Handa, A. Davison & S. Leutenegger. *SemanticFusion: Dense 3D Semantic Mapping with Convolutional Neural Networks*. arxiv:1609.05130, 2016. Zitiert auf Seite 38.

[McGreavy16] C. McGreavy, L. Kunze & N. Hawes. *Next Best View Planning for Object Recognition in Mobile Robotics*. In Proceedings of the 34th Workshop of the UK Planning and Scheduling Special Interest Group, 2016. Zitiert auf Seite 58.

[Mei20] J. Mei, B. Gao, D. Xu, W. Yao, X. Zhao & H. Zhao. *Semantic Segmentation of 3D LiDAR Data in Dynamic Scene Using Semi-Supervised Learning*. IEEE Transactions on Intelligent Transportation Systems, Band 21, 2020. Zitiert auf Seite 79.

[Menon23] R. Menon, T. Zaenker, N. Dengler & M. Bennewitz. *NBV-SC: Next Best View Planning Based on Shape Completion for Fruit Mapping and Reconstruction*. In 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2023. Zitiert auf Seite 57.

[Miandarhoie17] A. R. Miandarhoie, K. Khalili & H. Mohammadinejad. *CAD mesh models segmentation into swept surfaces*. The International Journal of Advanced Manufacturing Technology, Band 92, 2017. Zitiert auf Seite 27.

[Miyajima94] K. Miyajima & A. Ralescu. *Spatial Organization in 2D Segmented Images: Representation and Recognition of Primitive Spatial Relations*. Fuzzy Sets Syst., Band 65, 1994. Zitiert auf Seite 96.

[Mojtahedzadeh13] R. Mojtahedzadeh, A. Bouguerra & A. Lilienthal. *Automatic relational scene representation for safe robotic manipulation tasks*. In 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2013. Zitiert auf Seite 105.

[Mojtahedzadeh14] R. Mojtahedzadeh, A. Bouguerra, E. Schaffernicht & A. Lilienthal. *Probabilistic relational scene representation and decision making under incomplete information for robotic manipulation tasks*. In 2014 IEEE International Conference on Robotics and Automation (ICRA), 2014. Zitiert auf Seite 105.

[Mojtahedzadeh16] R. Mojtahedzadeh. *Safe Robotic Manipulation to Extract Objects from Piles: From 3D Perception to Object Selection*. Dissertation, Örebro, 2016. Zitiert auf Seite 105.

[Monica16] R. Monica, J. Aleotti & S. Caselli. *A KinFu based approach for robot spatial attention and view planning*. Robotics and Autonomous Systems, Band 75, 2016. Zitiert auf Seite 57 und 79.

[Moratz02] R. Moratz, B. Nebel & C. Freksa. *Qualitative spatial reasoning about relative position*. In Spatial cognition III. Springer, 2002. Zitiert auf Seite 96.

[Muzahid24] A. Muzahid, H. Han, Y. Zhang, D. Li, Y. Zhang, J. Jamshid & F. Sohel. *Deep learning for 3D object recognition: A survey*. Neurocomputing, Band 608, 2024. Zitiert auf Seite 39.

[Naphade00] K. Naphade, R. Storer & S. Wu. *Graph-Theoretic Generation of Assembly Plans - Part I: Correct Generation of Precedence Graphs*. Technical Report IMSE, Lehigh University, 2000. Zitiert auf Seite 112.

[Nüchter08] A. Nüchter & J. Hertzberg. *Towards semantic maps for mobile robots*. Robotics and Autonomous Systems, Band 56, 2008. Zitiert auf Seite 96.

[Newcombe11] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges & A. Fitzgibbon. *KinectFusion: Real-time dense surface mapping and tracking*. In 2011 10th IEEE International Symposium on Mixed and Augmented Reality, 2011. Zitiert auf Seite 27.

[Newcombe15] R. A. Newcombe, D. Fox & S. M. Seitz. *DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time*. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. Zitiert auf Seite 79.

[Niu03] X. Niu, H. Ding & Y. Xiong. *A hierarchical approach to generating precedence graphs for assembly planning*. International Journal of Machine Tools and Manufacture, Band 43, 2003. Zitiert auf Seite 112.

[Park19] K. Park, T. Patten, J. Prankl & M. Vincze. *Multi-Task Template Matching for Object Detection, Segmentation and Pose Estimation Using Depth Images*. 2019 International Conference on Robotics and Automation (ICRA), 2019. Zitiert auf Seite 38.

[Postica16] G. Postica, A. Romanoni & M. Matteucci. *Robust moving objects detection in lidar data exploiting visual cues*. In 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2016. Zitiert auf Seite 79.

[Potapova17] E. Potapova, M. Zillich & M. Vincze. *Survey of recent advances in 3D visual attention for robotics*. The International Journal of Robotics Research, Band 36, 2017. Zitiert auf Seite 57.

[Radke05] R. Radke, S. Andra, O. Al-Kofahi & B. Roysam. *Image change detection algorithms: a systematic survey*. IEEE Transactions on Image Processing, Band 14, 2005. Zitiert auf Seite 79.

[Ramirez-Amaro14] K. Ramirez-Amaro, M. Beetz & G. Cheng. *Automatic segmentation and recognition of human activities from observation based on semantic reasoning*. In 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2014. Zitiert auf Seite 96.

[Ramirez-Amaro19] K. Ramirez-Amaro, Y. Yang & G. Cheng. *A survey on semantic-based methods for the understanding of human movements*. Robotics and Autonomous Systems, Band 119, 2019. Zitiert auf Seite 96.

[Redmon17] J. Redmon & A. Farhadi. *YOLO9000: Better, Faster, Stronger*. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. Zitiert auf Seite 38.

[Reiter80] R. Reiter. *A logic for default reasoning*. Artificial Intelligence, Band 13, 1980. Zitiert auf Seite 97.

[Riedelbauch17] D. Riedelbauch & D. Henrich. *Coordinating flexible human-robot teams by local world state observation*. In 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), 2017. Zitiert auf Seite 79 und 112.

[Riedelbauch20] D. Riedelbauch. *Dynamic Task Sharing for Flexible Human-Robot Teaming under Partial Workspace Observability*. Dissertation, Bayreuth, 2020. Zitiert auf Seite 112.

[Roa15] M. A. Roa & R. Suárez. *Grasp quality measures: review and performance*. Autonomous Robots, Band 38, 2015. Zitiert auf Seite 96.

[Rohner19a] D. Rohner, M. Fichtner & D. Henrich. *Vision-based Generation of Precedence Graphs*. Annals of Scientific Society for Assembly, Handling and Industrial Robotics, 2019. Zitiert auf Seite 112, 114 und 115.

[Rohner19b] D. Rohner & D. Henrich. *Object Recognition for Robotics based on Planar Reconstructed B-Rep Models*. The Third IEEE International Conference on Robotic Computing, 2019. Zitiert auf Seite 36.

[Rohner20a] D. Rohner & D. Henrich. *Using Active Vision for Enhancing an Surface-based Object Recognition Approach*. The Fourth IEEE International Conference on Robotic Computing, 2020. Zitiert auf Seite 56, 60, 62, 64, 65, 71 und 72.

[Rohner20b] D. Rohner, D. Henrich & M. Sand. *User Guidance and Automatic Completion for Generating Planar B-Rep Models*. Annals of Scientific Society for Assembly, Handling and Industrial Robotics, 2020. Zitiert auf Seite 30, 109 und 118.

[Rohner22] D. Rohner, J. Hartwig & D. Henrich. *Detection and Handling of Dynamic Scenes during an Active Vision Process for Object Recognition using a Boundary Representation*. Annals of Scientific Society for Assembly, Handling and Industrial Robotics, 2022. Zitiert auf Seite 78, 85, 89 und 90.

[Romney95] B. Romney, C. Godard, M. Goldwasser & G. Ramkumar. *An Efficient System for Geometric Assembly Sequence Generation and Evaluation*. Band ASME 1995 15th International Computers in Engineering Conference and the ASME 1995 9th Annual Engineering Database Symposium, 1995. Zitiert auf Seite 112.

[Rottmann05] A. Rottmann, O. M. Mozos, C. Stachniss & W. Burgard. *Semantic place classification of indoor environments with mobile robots using boosting*. In Proceedings of the 20th National Conference on Artificial Intelligence - Volume 3. AAAI Press, 2005. Zitiert auf Seite 96.

[Rusu08] R. B. Rusu, N. Blodow, Z. C. Marton & M. Beetz. *Aligning point cloud views using persistent feature histograms*. In 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008. Zitiert auf Seite 38.

[Rusu09a] R. B. Rusu. *Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments*. Dissertation, München, 2009. Zitiert auf Seite 23.

[Rusu09b] R. B. Rusu, N. Blodow & M. Beetz. *Fast Point Feature Histograms (FPFH) for 3D registration*. In 2009 IEEE International Conference on Robotics and Automation, 2009. Zitiert auf Seite 38.

[Rusu10] R. B. Rusu, G. Bradski, R. Thibaux & J. Hsu. *Fast 3D recognition and pose using the Viewpoint Feature Histogram*. In 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2010. Zitiert auf Seite 38.

[Sahin07] E. Sahin, M. Cakmak, M. Dogar, E. Ugur & G. Üçoluk. *To Afford or Not to Afford: A New Formalization of Affordances Toward Affordance-Based Robot Control*. Adaptive Behavior, Band 15, 2007. Zitiert auf Seite 96.

[Sand16] M. Sand & D. Henrich. *Incremental reconstruction of planar B-Rep models from multiple point clouds*. The Visual Computer, Springer, 2016. Zitiert auf Seite 30.

[Sand17] M. Sand & D. Henrich. *Matching and Pose Estimation of Noisy, Partial and Planar B-Rep Models*. Computer Graphics International 2017, 2017. Zitiert auf Seite 30 und 41.

[Sand19] M. Sand. *Inkrementelle Rekonstruktion von planaren Volumenmodellen mit handgehaltenen Tiefenkameras*. Dissertation, Bayreuth, 2019. Zitiert auf Seite 27, 28, 30, 32, 33, 37, 40, 41, 42, 43, 44, 45, 48, 61, 80, 81, 84, 85, 99 und 118.

[Savage19] J. Savage, D. A. Rosenblueth, M. Matamoros, M. Negrete, L. Contreras, J. Cruz, R. Martell, H. Estrada & H. Okada. *Semantic reasoning in service robots using expert systems*. Robotics and Autonomous Systems, Band 114, 2019. Zitiert auf Seite 98.

[Savela18] N. Savela, T. Turja & A. Oksanen. *Social Acceptance of Robots in Different Occupational Fields: A Systematic Literature Review*. International Journal of Social Robotics, Band 10, 2018. Zitiert auf Seite 98.

[Searle80] J. R. Searle. *Minds, brains, and programs*. Behavioral and Brain Sciences, Band 3, 1980. Zitiert auf Seite 97.

[Sengar19] S. S. Sengar & S. Mukhopadhyay. *Moving object detection using statistical background subtraction in wavelet compressed domain*. Multimedia Tools and Applications, Band 79, 2019. Zitiert auf Seite 79.

[Shanahan97] M. Shanahan. Solving the frame problem: a mathematical investigation of the common sense law of inertia. MIT Press, Cambridge, 1997. Zitiert auf Seite 97.

[Silberman12] N. Silberman, D. Hoiem, P. Kohli & R. Fergus. *Indoor Segmentation and Support Inference from RGBD Images*. In Computer Vision – ECCV 2012. Springer Berlin Heidelberg, 2012. Zitiert auf Seite 38.

[Sinapov13] J. Sinapov & A. Stoytchev. *Grounded object individuation by a humanoid robot*. In 2013 IEEE International Conference on Robotics and Automation, 2013. Zitiert auf Seite 58.

[Singer21] D. Singer, D. Henrich & D. Rohner. *Robot-Based Creation of Complete 3D Workpiece Models*. Annals of Scientific Society for Assembly, Handling and Industrial Robotics, 2021. Zitiert auf Seite 116, 117, 118 und 119.

[Singh14] A. Singh, J. Sha, K. S. Narayan, T. Achim & P. Abbeel. *BigBIRD: A large-scale 3D database of object instances*. In 2014 IEEE International Conference on Robotics and Automation (ICRA), 2014. Zitiert auf Seite 116.

[Somani13] N. Somani, E. C. Dean-León, C. Cai & A. Knoll. *Scene Perception and Recognition in Industrial Environments for Human-Robot Interaction*. In Advances in Visual Computing. Springer Berlin Heidelberg, 2013. Zitiert auf Seite 38.

[Somani15] N. Somani, A. Perzylo, C. Cai, M. Rickert & A. Knoll. *Object detection using boundary representations of primitive shapes*. In 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO), 2015. Zitiert auf Seite 38.

[Spangenberg17] M. Spangenberg. *Intuitive Roboterkommandierung basierend auf Verbalisierten Physikalischen Effekten*. Dissertation, Bayreuth, 2017. Zitiert auf Seite 97.

[Stachniss06] C. Stachniss, O. Martinez Mozos & W. Burgard. *Speeding-up multi-robot exploration by considering semantic place information*. In Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. Zitiert auf Seite 96.

[Stückler14] J. Stückler & S. Behnke. *Multi-resolution surfel maps for efficient dense 3D modeling and tracking*. Journal of Visual Communication and Image Representation, Band 25, 2014. Zitiert auf Seite 27.

[Steinbrucker13] F. Steinbrucker, C. Kerl, D. Cremers & J. Sturm. *Large-Scale Multi-resolution Surface Reconstruction from RGB-D Sequences*. In 2013 IEEE International Conference on Computer Vision, 2013. Zitiert auf Seite 27.

[Sudha15] D. D. Sudha & P. Jayaraju. *Reducing Semantic Gap in Video Retrieval with Fusion: A Survey*. Procedia Computer Science, Band 50, 2015. Zitiert auf Seite 95.

[Sundaram01] S. Sundaram, I. Remmler & N. Amato. *Disassembly sequencing using a motion planning approach*. In IEEE International Conference on Robotics and Automation, Band 2, 2001. Zitiert auf Seite 112.

[Suppa04] M. Suppa, P. Wang, K. Gupta & G. Hirzinger. *C-space exploration using noisy sensor models*. In IEEE International Conference on Robotics and Automation, Band 5, 2004. Zitiert auf Seite 57.

[Szegedy15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke & A. Rabinovich. *Going deeper with convolutions*. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 2015. Zitiert auf Seite 38.

[Taddeo05] M. Taddeo & L. F. and. *Solving the symbol grounding problem: a critical review of fifteen years of research*. Journal of Experimental & Theoretical Artificial Intelligence, Band 17, 2005. Zitiert auf Seite 97.

[Taguchi13] Y. Taguchi, Y.-D. Jian, S. Ramalingam & C. Feng. *Point-plane SLAM for hand-held 3D sensors*. In 2013 IEEE International Conference on Robotics and Automation, 2013. Zitiert auf Seite 27.

[Taketomi17] T. Taketomi, H. Uchiyama & S. Ikeda. *Visual SLAM algorithms: A survey from 2010 to 2016*. IPSJ Transactions on Computer Vision and Applications, Band 9, 2017. Zitiert auf Seite 27.

[Takikawa19] T. Takikawa, D. Acuna, V. Jampani & S. Fidler. *Gated-SCNN: Gated Shape CNNs for Semantic Segmentation*. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019. Zitiert auf Seite 38.

[Talbot21] B. Talbot, F. Dayoub, P. Corke & G. Wyeth. *Robot Navigation in Unseen Spaces Using an Abstract Map*. IEEE Transactions on Cognitive and Developmental Systems, Band 13, 2021. Zitiert auf Seite 98.

[Tateno16] K. Tateno, F. Tombari & N. Navab. *When 2.5D is not enough: Simultaneous reconstruction, segmentation and recognition on dense SLAM*. In 2016 IEEE International Conference on Robotics and Automation (ICRA), 2016. Zitiert auf Seite 96.

[Tellex14] S. Tellex, R. Knepper, A. Li, D. Rus & N. Roy. *Asking for Help Using Inverse Semantics*. Robotics: Science and Systems X, 2014. Zitiert auf Seite 98.

[Tenorth17] M. Tenorth & M. Beetz. *Representations for robot knowledge in the KnowRob framework*. Artificial Intelligence, Band 247, 2017. Zitiert auf Seite 95.

[Terven23] J. Terven, D.-M. Córdova-Esparza & J.-A. Romero-González. *A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS*. Machine Learning and Knowledge Extraction, Band 5, 2023. Zitiert auf Seite 38.

[Theologou15] P. Theologou, I. Pratikakis & T. Theoharis. *A comprehensive overview of methodologies and performance evaluation frameworks in 3D mesh segmentation*. Computer Vision and Image Understanding, Band 135, 2015. Zitiert auf Seite 27.

[Thielscher97] M. Thielscher. *Ramification and causality*. Artificial Intelligence, Band 89, 1997. Zitiert auf Seite 97.

[Thielscher98] M. Thielscher. *Introduction to the Fluent Calculus*. Electronic Transactions on Artificial Intelligence, Band 2, 1998. Zitiert auf Seite 97.

[Thippur15] A. Thippur, C. Burbridge, L. Kunze, M. Alberti, J. Folkesson, P. Jensfelt & N. Hawes. *A Comparison of Qualitative and Metric Spatial Relation Models for Scene Understanding*. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI Press, 2015. Zitiert auf Seite 96.

[Thomas03] U. Thomas, M. Barrenscheen & F. Wahl. *Efficient assembly sequence planning using stereographical projections of C-Space obstacles*. Proceedings of the IEEE International Symposium on Assembly and Task Planning, 2003. Zitiert auf Seite 112 und 113.

[Thomason22] J. Thomason, M. Shridhar, Y. Bisk, C. Paxton & L. Zettlemoyer. *Language grounding with 3d objects*. In Conference on robot learning. PMLR, 2022. Zitiert auf Seite 95.

[Tistarelli94] M. Tistarelli. *Recognition by using an active/space-variant sensor*. 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1994. Zitiert auf Seite 57.

[Tombari10] F. Tombari, S. Salti & L. Di Stefano. *Unique Signatures of Histograms for Local Surface Description*. Proceedings ECCV, Band 6313, 2010. Zitiert auf Seite 38.

[Tremblay18] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox & S. Birchfield. *Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects*. arxiv:1809.10790, 2018. Zitiert auf Seite 96.

[Tuomi20] A. Tuomi, I. Tussyadiah & J. Stienmetz. *Service Robots and the Changing Roles of Employees in Restaurants: A Cross Cultural Study*. e-Review of Tourism Research, 2020. Zitiert auf Seite 98.

[Valada17] A. Valada, G. L. Oliveira, T. Brox & W. Burgard. *Deep Multispectral Semantic Scene Understanding of Forested Environments Using Multimodal Fusion*. In 2016 International Symposium on Experimental Robotics. Springer International Publishing, 2017. Zitiert auf Seite 38.

[Vänni17] K. J. Vänni & S. E. Salin. *A Need for Service Robots Among Health Care Professionals in Hospitals and Housing Services*. In Social Robotics. Springer International Publishing, 2017. Zitiert auf Seite 98.

[Varady08] T. Varady. *Automatic Procedures to Create CAD Models from Measured Data*. Computer-Aided Design and Applications, Band 5, 2008. Zitiert auf Seite 27.

[Vasquez-Gomez17] J. Vasquez-Gomez, L. Sucar & R. Murrieta-Cid. *View/state planning for three-dimensional object reconstruction under uncertainty*. Autonomous Robots, Band 41, 2017. Zitiert auf Seite 57.

[Vaswani17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser & I. Polosukhin. *Attention is all you need*. Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017. Zitiert auf Seite 24.

[Venkataraman19] A. Venkataraman, B. Griffin & J. J. Corso. *Kinematically-Informed Interactive Perception: Robot-Generated 3D Models for Classification*. arxiv:1901.05580, 2019. Zitiert auf Seite 116.

[Várady07] T. Várady, M. A. Facello & Z. Terék. *Automatic extraction of surface structures in digital shape reconstruction*. Computer-Aided Design, Band 39, 2007. Zitiert auf Seite 27.

[Wang15] W. Wang, L. Chen, Z. Liu, K. Kühnlenz & D. Burschka. *Textured/textureless object recognition and pose estimation using RGB-D image*. Journal of Real-Time Image Processing, Band 10, 2015. Zitiert auf Seite 116.

[Wang18] P. Wang, L. Sun, A. F. Smeaton, C. Gurrin & S. Yang. *Chapter 9 - Computer Vision for Lifelogging: Characterizing Everyday Activities Based on Visual Semantics*. In Computer Vision for Assistive Healthcare, Computer Vision and Pattern Recognition. Academic Press, 2018. Zitiert auf Seite 96.

[Wang23] W. Wang, X. Li, Y. Dong, J. Xie, D. Guo & H. Liu. *Natural Language Instruction Understanding for Robotic Manipulation: a Multisensory Perception Approach*. In 2023 IEEE International Conference on Robotics and Automation (ICRA), 2023. Zitiert auf Seite 98.

[Welke13] K. Welke, D. Schiebener, T. Asfour & R. Dillmann. *Gaze selection during manipulation tasks*. In 2013 IEEE International Conference on Robotics and Automation, 2013. Zitiert auf Seite 58.

[Wilkes92] D. Wilkes & J. Tsotsos. *Active object recognition*. In Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1992. Zitiert auf Seite 57.

[Wilson94] R. H. Wilson & J.-C. Latombe. *Geometric reasoning about mechanical assembly*. Artificial Intelligence, Band 71, 1994. Zitiert auf Seite 112.

[Wölfel21] K. Wölfel. *SpIRo Sprachbasierte Instruktion kraftbasierter Roboterbewegungen*. Dissertation, Bayreuth, 2021. Zitiert auf Seite 97.

[Wolter92] J. Wolter, S. Chakrabarty & J. Tsao. *Mating constraint languages for assembly sequence planning*. In Proceedings 1992 IEEE International Conference on Robotics and Automation, 1992. Zitiert auf Seite 112.

[Wu09] K. Wu, E. Otoo & K. Suzuki. *Optimizing two-pass connected-component labeling algorithms*. Pattern Analysis and Applications, Band 12, 2009. Zitiert auf Seite 117.

[Wu14] C. Wu, I. Lenz & A. Saxena. *Hierarchical Semantic Labeling for Task-Relevant RGB-D Perception*. Robotics: Science and Systems, 2014. Zitiert auf Seite 38.

[Wu22] J. Wu, Z. Jin, A. Liu, L. Yu & F. Yang. *A survey Of learning-Based control of robotic visual servoing systems*. Journal of the Franklin Institute, Band 359, 2022. Zitiert auf Seite 79.

[Xiao00] J. Xiao & X. Ji. *A divide-and-merge approach to automatic generation of contact states and planning of contact motion*. In Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation, Band 1, 2000. Zitiert auf Seite 112.

[Xiao15] W. Xiao, B. Vallet, M. Brédif & N. Paparoditis. *Street environment change detection from mobile laser scanning point clouds*. ISPRS Journal of Photogrammetry and Remote Sensing, Band 107, 2015. Zitiert auf Seite 79.

[Xiong12] B. Xiong & X. Ding. *A Master–Slave System for Intelligent Visual Surveillance*. Lecture Notes in Electrical Engineering, Band 135, 2012. Zitiert auf Seite 57.

[Yang19] H. Yang, Y. Guo, X. Wang, J. Song, C. Sun & C. Yang. *Three-Dimensional Point Cloud Feature Change Detection Algorithm of Open-pit Mine Based on Discrete Curvature Analysis*. In 14th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2019. Zitiert auf Seite 79.

[Yazdi18] M. Yazdi & T. Bouwmans. *New trends on moving object detection in video images captured by a moving camera: A survey*. Computer Science Review, Band 28, 2018. Zitiert auf Seite 98.

[Yilmaz06] A. Yilmaz, O. Javed & M. Shah. *Object tracking: A survey*. ACM Computing Surveys, Band 38, 2006. Zitiert auf Seite 79.

[Yoon18] J. S. Yoon, Z. Li & H. S. Park. *3D Semantic Trajectory Reconstruction from 3D Pixel Continuum*. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018. Zitiert auf Seite 96.

[Younes17] G. Younes, D. Asmar, E. Shammas & J. Zelek. *Keyframe-based monocular SLAM: design, survey, and future directions*. Robotics and Autonomous Systems, Band 98, 2017. Zitiert auf Seite 27.

[Yu04] Y. Yu & K. Gupta. *C-space Entropy: A Measure for View Planning and Exploration for General Robot-Sensor Systems in Unknown Environments*. The International Journal of Robotics Research, Band 23, 2004. Zitiert auf Seite 57.

[Yu09] G. Yu & J.-M. Morel. *A fully affine invariant image comparison method*. 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. Zitiert auf Seite 38.

[Yu15] F. Yu & V. Koltun. *Multi-Scale Context Aggregation by Dilated Convolutions*. arxiv:1511.07122, 2015. Zitiert auf Seite 38.

[Zellers21] R. Zellers, A. Holtzman, M. Peters, R. Mottaghi, A. Kembhavi, A. Farhadi & Y. Choi. *PIGLeT: Language Grounding Through Neuro-Symbolic Interaction in a 3D World*. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Association for Computational Linguistics, 2021. Zitiert auf Seite 96.

[Zender08] H. Zender, O. Martínez Mozos, P. Jensfelt, G.-J. Kruijff & W. Burgard. *Conceptual spatial representations for indoor mobile robots*. Robotics and Autonomous Systems, Band 56, 2008. Zitiert auf Seite 98.

[Zeng18] Z. Zeng, Z. Zhou, Z. Sui & O. C. Jenkins. *Semantic Robot Programming for Goal-Directed Manipulation in Cluttered Scenes*. In 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018. Zitiert auf Seite 96.

[Zhang02] Y. Zhang, J. Ni, Z. Lin & X. Lai. *Automated sequencing and sub-assembly detection in automobile body assembly planning*. Journal of Materials Processing Technology, Band 129, 2002. Zitiert auf Seite 112.

[Zhang25] W. Zhang, X. Li, X. Liu, S. Lu & H. Tang. *Facing challenges: A survey of object tracking*. Digital Signal Processing, Band 161, 2025. Zitiert auf Seite 79.

Eigene Publikationen

[Bogner18] C. Bogner, B. Seo, D. Rohner & B. Reineking. *Classification of rare land cover types: Distinguishing annual and perennial crops in an agricultural catchment in South Korea*. PLOS ONE, Band 13, 2018. Nicht zitiert.

[Rohner19a] D. Rohner, M. Fichtner & D. Henrich. *Vision-based Generation of Precedence Graphs*. Annals of Scientific Society for Assembly, Handling and Industrial Robotics, 2019. Zitiert auf Seite 112, 114 und 115.

[Rohner19b] D. Rohner & D. Henrich. *Object Recognition for Robotics based on Planar Reconstructed B-Rep Models*. The Third IEEE International Conference on Robotic Computing, 2019. Zitiert auf Seite 36.

[Rohner20a] D. Rohner & D. Henrich. *Using Active Vision for Enhancing an Surface-based Object Recognition Approach*. The Fourth IEEE International Conference on Robotic Computing, 2020. Zitiert auf Seite 56, 60, 62, 64, 65, 71 und 72.

[Rohner20b] D. Rohner, D. Henrich & M. Sand. *User Guidance and Automatic Completion for Generating Planar B-Rep Models*. Annals of Scientific Society for Assembly, Handling and Industrial Robotics, 2020. Zitiert auf Seite 30, 109 und 118.

[Rohner22] D. Rohner, J. Hartwig & D. Henrich. *Detection and Handling of Dynamic Scenes during an Active Vision Process for Object Recognition using a Boundary Representation*. Annals of Scientific Society for Assembly, Handling and Industrial Robotics, 2022. Zitiert auf Seite 78, 85, 89 und 90.

[Singer21] D. Singer, D. Henrich & D. Rohner. *Robot-Based Creation of Complete 3D Workpiece Models*. Annals of Scientific Society for Assembly, Handling and Industrial Robotics, 2021. Zitiert auf Seite 116, 117, 118 und 119.

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet habe.

Weiterhin erkläre ich, dass ich die Hilfe von gewerblichen Promotionsberatern bzw. –vermittlern oder ähnlichen Dienstleistern weder bisher in Anspruch genommen habe, noch künftig in Anspruch nehmen werde.

Zusätzlich erkläre ich hiermit, dass ich keinerlei frühere Promotionsversuche unternommen habe.

Bayreuth, den

Dorian Rohner