

Applikation von Lerndialogen auf Objektbezeichner mit multimodaler Datenverarbeitung

Studienarbeit am Institut Interactive Systems Labs

Prof. Dr. Alex Waibel

Fakultät für Informatik

Universität Karlsruhe (TH)

von

Cand. Inform.

Daniel Neubig

Betreuer:

Prof. Dr. Alex Waibel

Dipl. Inform. Hartwig Holzapfel

Tag der Anmeldung: 1. Januar 2007

Tag der Abgabe: 10. Juli 2007

Abstract

The following paper describes methods of a dialog system for object learning in a changing environment. An object model is introduced, which enables the system to describe objects, using ontology. The dialog held by the robot to interact with the user and react on calculated dialog situations is implemented in a strategy. This strategy is spread into dialogue modules, which try to achieve the dialog goal, given by the user. To be able to adapt the system, functionality to detect and learn new words online is developed. Furthermore an approach to ground the correct ontology entry is given. Using a generated barkeeper scenario and an implemented prototype a set of tests was performed to evaluate the system. The success in learning and relocating the new objects during the testes shows, that the system is able to adapt its models to the scenery.

Danksagung

Ich möchte mich bei meinem Betreuer, Hartwig Holzapfel, der mich die ganze Zeit über gefördert und gefordert hat und bei Pedram Azad, der mir mit Rat, Tat und Programmteilen zur Seite stand, bedanken. Außerdem gebührt mein Dank meinem Kommilitonen Stefan Ziesemer, dessen weiterentwickelten Spracherkenner ich verwenden konnte und meinen ganzen Probanden: Jana Stachetzki, Tsvetelin Chehtov, Dennis Knoller, Michael Demel, Andreas Kreuzer, Anne Brehm, Dominik Englert, Svetlana Baglyk, David Brehm und Lilyana Shehtova.

Inhalt

1. Einleitung.....	7
1.1. Zielsetzung der Arbeit.....	8
1.2. Gliederung der Arbeit.....	8
2. Grundlagen	9
2.1. Natürlich sprachliche Dialogsysteme mit neuen Objekten	9
2.1.1. Spracherkenner	10
2.1.2. Sprachverstehen	11
2.1.3. Dialogsystem.....	13
2.2. Objekterkennung	14
2.2.1. SIFT Features.....	14
2.2.2. Objekte wiedererkennen	15
3. Bestehende Arbeiten.....	16
3.1. Erkennen und Lernen neuer Wörter	16
3.2. Barkeeperszenario	18
3.3. Lernen neuer Wörter im Dialog	19
3.4. Lernen von Objektbenennungen mit visuellen Prozessen.....	20
3.5. Zusammenfassung.....	21
4. Entwurf	23
4.1. Gesamtsystem.....	23
4.2. Modelle.....	24
4.2.1. Zustandsmodell	24
4.2.2. Slot Model.....	25
4.2.3. Aktionsmodell.....	25
4.2.4. Objektmodell.....	25
4.3. Dialogablauf	27

4.4. Fehlerarten im Dialog.....	32
5. Implementierung.....	33
5.1. Szenario.....	33
5.2. Experimente	35
6. Ergebnisse und Diskussion.....	37
6.1. Ergebnisse	37
6.1.1. Gesamtsystem	37
6.1.2. Objekterkennung.....	38
6.1.3. Buchstabiererkennung.....	39
6.1.4. Spracherkennung.....	41
6.2. Analyse der Ergebnisse	42
7. Zusammenfassung und Ausblick.....	44
8. Literaturverzeichnis	46

Abbildungsverzeichnis

Abbildung 2-1: Zyklus eines natürlich sprachlichen Dialogsystems.....	9
Abbildung 2-2: Das semiotische Dreieck von Ogden und Richards	11
Abbildung 2-3: Erkennung eines Objektes mit Anzeige der Merkmale.....	15
Abbildung 3-1: Barkeeperszenario	18
Abbildung 3-2: Ablaufdiagramm des OOV Dialogs	20
Abbildung 3-3: Architektur des Systems zum Objekte lernen	20
Abbildung 3-4: Dialogablauf	21
Abbildung 4-1: Allgemeiner Aufbau der Applikation.....	23
Abbildung 4-2: Ontologie Beispiel.....	26
Abbildung 4-3: Dialogablauf	27
Abbildung 4-4: Dialogbeispiel Objekt reichen	30
Abbildung 4-5: Dialogbeispiel Objekt lernen.....	31
Abbildung 5-1: Aufnahme des Roboters während dem Dialog.....	36
Abbildung 6-1: Lernen eines unbekanntes Objektes.....	38
Abbildung 6-2: Merkmale beim Lernen eines unbekanntes Objektes	39

1. Einleitung

Bisher arbeiten Roboter hauptsächlich dort, wo Arbeiten besonders gefährlich, schwer oder eintönig sind. Dass Roboter unsere Autos bauen und zum Mars fliegen ist inzwischen für viele Menschen normal. Doch eine Vision beschäftigt die Roboterforscher bereits seit Jahrzehnten: Serviceroboter, die in direktem Kontakt mit dem Menschen Aufgaben verrichten und dabei intelligent zu Werke gehen. In Forschungseinrichtungen weltweit werden erste Szenarien durchlaufen, in denen Roboter Gegenstände holen, aufräumen oder sonstige Aufgaben erledigen. Ein Ansatz verwendet dabei gesprochene Anweisungen und es wird ein Dialog mit dem Benutzer geführt. Diese Dialogsysteme erleichtern den Umgang mit dem Roboter, da der Nutzer nicht spezielle Hardware erlernen und herumtragen muss. Das hat den Vorteil, dass ohne Tastatur, Maus und Bildschirm mit dem System interagiert werden kann. Die Einsatzmöglichkeiten sind vielfältig: Im Auto, wenn man keine Hände frei hat oder vom Krankenbett aus, wenn man keine Möglichkeit hat, die nötige Hardware zu bedienen. Informationsdienste sind nicht notwendigerweise an unser heute übliches Bild von Computern gebunden, denn die gewünschte Auskunft kann auch, wie unter Menschen üblich, sprachlich ausgegeben werden.

Damit Roboter in unserem alltäglichen Umfeld eingesetzt werden können, ist es nötig, dass sie sich an die aktuelle Umgebung und den aktuellen Benutzer anpassen. Während der Herstellung eines Serviceroboters ist es nicht möglich, das System auf alle Eventualitäten vorzubereiten, die der Nutzer von dem System abrufen wird. Da ein Roboter im Gesundheitsgewerbe oder im Haushalt die Arbeit erleichtern und nicht unnötig verkomplizieren soll, sind lange Einarbeitungszeiten oder ein dickes Nutzerhandbuch nicht denkbar, in denen der genaue Funktionsumfang beschrieben wird. Ein freier und normaler Dialog soll ermöglicht werden. Ein dialoggesteuerter Roboter wird also früher oder später mit neuen Objekten und Aufgaben konfrontiert, die er bisher noch nicht kannte und lösen konnte, während sich die Nutzungsumgebung ändert, weil neue Objekte hinzukommen oder neue Benutzer mit dem System interagieren. Um darauf reagieren zu können, sollte das System in der Lage sein, neue Aufgaben und Objekte zu erkennen und diese zu lernen.

1.1. Zielsetzung der Arbeit

Im Rahmen des Sonderforschungsbereichs SFB588 „Humanoide Roboter – Lernende und kooperierende multimodale Roboter“ (Deutsche Forschungsgemeinschaft, 2001) hat die Deutsche Forschungsgemeinschaft (DFG) das Ziel ausgegeben, dass ein Robotersystem entsteht, das in der Lage ist, dem Menschen in alltäglichen Situationen zu helfen, mit den Menschen kooperieren kann und sich dabei an die Umgebung anpasst. Dabei entstand bereits eine Vielzahl an wissenschaftlichen Arbeiten, die sich mit der Kooperation auseinandersetzen. Bisher ist es aber notwendig, dass die Objekte, mit denen der Roboter interagiert im Vorfeld bekannt sind. Ziel dieser Arbeit ist es, dass diese Einschränkung überwunden wird.

Es soll ein Konzept erstellt werden, das es ermöglicht, ein Dialogsystem zu verfassen, in dem mit Methoden der Objekterkennung, der Spracherkennung und mit variablen Datenbasen Objekte gelernt werden können. Dazu sollen Merkmale des Objektes gefunden und zusammen mit einem Bezeichner und der Beschreibung des Objektes gelernt werden. Außerdem soll es möglich sein, mit Hilfe eines Dialoges die gelernten Objekte wieder zu beschreiben und aufzufinden. Dieses Konzept soll in einem Prototyp umgesetzt werden.

1.2. Gliederung der Arbeit

Kapitel 2 beschreibt die Grundlagen dieser Arbeit. Es wird auf Dialogsysteme eingegangen, die es ermöglichen, die Intention des Benutzers zu ermitteln und zu bearbeiten. Spracherkennung, Sprachverstehen und Dialogführung werden vorgestellt. Zusätzlich werden die Methoden der Objekterkennung erläutert.

Kapitel 3 zeigt einen kurzen Überblick und eine Zusammenfassung der vorhergehenden Arbeiten. Darin werden die notwendigen Änderungen für das Erkennen von unbekanntem Wörtern aufgezeigt, ein System, das Objektbezeichner lernt, vorgestellt. Zudem wird ein System vorgestellt, das ebenfalls Objekte mit Hilfe von einem Dialog lernen kann.

Kapitel 4 beschreibt den Entwurf des Gesamtsystems, wie die Modelle aufgebaut sind, der Dialog abläuft und welche Fehlerarten auftreten können. Die Modellierung von Objekten mit Hilfe einer Ontologie wird erklärt und der Ablauf des Dialogs in seinen Einzelheiten erläutert. Außerdem werden mögliche Fehler im Dialogablauf und der Erkennung aufgezeigt.

Kapitel 5 geht auf die Umsetzung des Gesamtsystems in einem Prototyp und auf die Einzelheiten der erstellten Implementierung ein. Die Durchführung der Tests mit mehreren Benutzern wird beschrieben.

Kapitel 6 befasst sich mit der Auswertung der durchgeführten Tests und die Ergebnisse dieser Arbeit werden diskutiert. Die Evaluation des Prototyps wird dargestellt.

Kapitel 7 fasst alles zusammen und gewährt einen Ausblick, wo noch Verbesserungen gemacht werden können.

2. Grundlagen

In diesem Kapitel werden die nötigen Grundlagen erläutert, die nötig sind, wenn man in einem Robotersystem neue Objekte lernen lassen möchte. Erst wird darauf eingegangen, welche Veränderungen am Dialogsystem nötig sind. Dann werden die verwendeten Sprachmodelle und Objektmodelle vorgestellt und am Ende wird die Erkennung von texturierten Objekten erläutert.

2.1. Natürlich sprachliche Dialogsysteme mit neuen Objekten

Um mit einem Roboter zu interagieren sollte der Benutzer nicht an eine Tastatur oder eine Maus gebunden sein. Um ein freies Handeln zu ermöglichen, werden Dialogsysteme verwendet, bei denen der Benutzer unvorbereitet und natürlich mit dem System sprechen kann. Dementsprechend formuliert der Benutzer seine Wünsche und Anweisungen sprachlich an den Roboter. Eine spezielle Kommandosprache ist nicht notwendig weil der Benutzer frei formuliert, ohne sich lange an das System anpassen zu müssen. Ein Spracherkenner erfasst die aufgenommene Sprache, interpretiert diese und das System reagiert mit Nachfragen bis die genaue Intention erkannt wurde. Dann führt das System den nötigen Schritt aus und erfüllt den Wunsch des Benutzers. Man kann also wie gewohnt mit seinem Gegenüber sprechen. Ein natürlich sprachliches Dialogsystem besteht, wie in Abbildung 2-1 dargestellt, aus einem Spracherkenner, einer Komponente zum Sprachverstehen und einem Dialogsystem. In jeder der Komponenten ist es nötig, dass neue Wörter, die das System bisher nicht kannte, so genannte Out Of Vocabulary Words (OOV) gefunden und gelernt werden können.

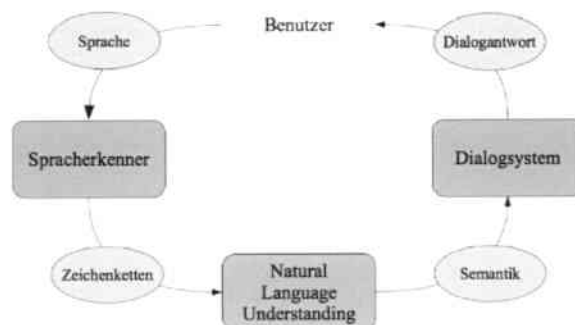


Abbildung 2-1: Zyklus eines natürlich sprachlichen Dialogsystems

Diese Einzelkomponenten werden vom Dialogmanager Tapas (Holzapfel H. , 2005) verwaltet, der eine Entwicklung von multilingualen und multimodalen Dialogsystemen ermöglicht.

2.1.1. Spracherkennung

Der Spracherkennung wandelt die aufgenommenen Signale mit Hilfe eines Vokabulars, eines akustischen und eines sprachlichen Modells in eine Hypothese dessen um, was der Benutzer gesagt hat. Dieser Ansatz ist rein statistisch und nutzt trainierte Modelle, die aus einer großen Menge von Trainingsdaten erstellt wurden. Wörter, die nicht in dem Vokabular enthalten sind, also unbekannte Wörter, kann der Spracherkennung nicht ohne Weiteres detektieren und erkennen, da er immer versucht das aufgenommene mit den trainierten Mustern zu vereinbaren. Wenn es keine Trainingsbeispiele für das Wort gibt, kann der Spracherkennung dieses eigentlich nicht verstehen. Nach der Erkennung wird nicht nur das wahrscheinlichste erkannte Wort ausgegeben, sondern eine Reihe an Möglichkeiten für dieses Wort.

Das Wörterbuch umfasst für jedes bekannte Wort eine Darstellung der Aussprache. Zu jedem Wort wird eine Reihe von Phonemen gespeichert, die die akustischen Einheiten darstellen, die das Wort zusammensetzen. Phoneme sind Einheiten von Sprache, also Laute, die sich untereinander alle unterscheiden.

Aktuelle Spracherkennung arbeiten mit einer statistischen Mustererkennung, indem sie die wahrscheinlichste Hypothese zu einer aufgenommenen Äußerung suchen. Mit Hilfe des akustischen Modells werden die Phoneme oder Subphoneme mit der Äußerung verglichen. Es gibt nur eine endliche Zahl von akustischen Modellen für jede Sprache, die die Aussprache vorschreibt. Zudem enthält das akustische Modell Einträge für Stottern, Atmen und andere Geräusche, die im Hintergrund auftreten können. Normalerweise werden hier Phoneme mit Phonemen oder gar Subphoneme mit Subphonemen verglichen. Außerdem gibt es noch ein Sprachmodell, das die möglichen Wörter zu sinnvollen Sätzen zusammenfasst. Dieses kann zum Beispiel mit Hilfe von kontextfreien Grammatiken oder statistischen Modellen, wie Bi- und Tri-Gram-Modellen geschehen. Die Wahrscheinlichkeit aus dem akustischen Modell wird mit der Wahrscheinlichkeit aus dem Sprachmodell kombiniert und mit Hilfe des Wörterbuches eine Liste der wahrscheinlichsten Wörter gebildet. Eine ausführliche Einführung in die Spracherkennung liefern (Waibel, 1990) und (Talamazzini, 2001).

2.1.2. Sprachverstehen

Die gewonnene Hypothese wird vom Sprachverstehen mit Hilfe einer Ontologie und einer semantischen Grammatik interpretiert. Eine Ontologie (Gruber, 1993) beschreibt Begriffe und deren Zusammenspiel, stellt also dar, wie die Begriffe miteinander in Verbindung gebracht werden können. Die Begriffe sind als Knoten dargestellt und das Zusammenspiel kann über logische Relationen, also Kanten realisiert werden. Jeder Knoten in einer Ontologie beschreibt den darin enthaltenen Begriff durch die Kanten mit den anderen Knoten in der Nachbarschaft. In einer Ontologie gibt es Begriffe, die Konzepte beschreiben, Instanzen, die Objekte eines Konzepts sind, Relationen, die die Eigenschaften eines Konzepts darstellen und Axiome, die direkt Informationen darstellen, die nicht anderweitig beschrieben werden können. Konzepte können auch von anderen Konzepten erben. In dem Fall, dass Objekte mit Hilfe einer Ontologie beschrieben werden, kann man mit den Relationen die Beziehungen zwischen Objekten darstellen oder Klassen von Objekten in Konzepte zusammenfassen. Die Ontologie beschreibt das Kontextwissen, um den Begriff und damit auch das Objekt, das durch den Begriff beschrieben wird, einzuordnen und zu interpretieren. Da nicht sichergestellt ist, dass der Benutzer beim Beschreiben den gleichen Begriff verwendet wie der gelernte Begriff, muss so eine Brücke geschlagen werden zwischen den beiden verschiedenen Beschreibungen und einem tatsächlich existierenden Objekt.

In der Linguistik wird dieser Zusammenhang mit dem semiotischen Dreieck (Ogden & Richards, 1923) beschrieben, in dem das Symbol einen Bezug ausdrückt und dieser wiederum ein bestimmtes Bezugsobjekt bestimmt. Eine direkte Verbindung zwischen Symbol und Bezugsobjekt gibt es nicht, diese ist nur gestrichelt angedeutet, da das Symbol nur von jemandem benutzt wird, um den Referenten zu vertreten. Abbildung 2-2 zeigt das Dreieck in der Definition von Ogden und Richards.

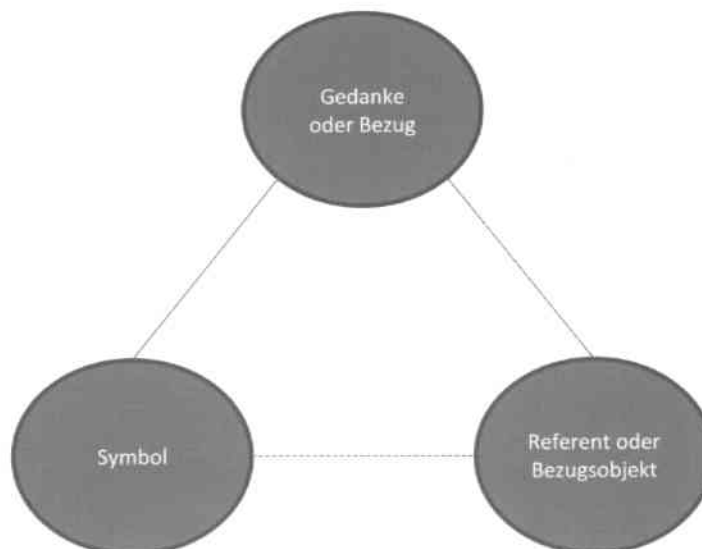


Abbildung 2-2: Das semiotische Dreieck von Ogden und Richards

Die Ontologie beschreibt also für das tatsächlich gesprochene Wort, das Symbol, in dem der Bezeichner fällt, einen Begriff, also die Semantik oder den Bezug. Damit wird ein tatsächlich existierendes Bezugsobjekt, das auf dem Tisch steht, beschrieben. Menschen können die drei Ecken des semiotischen Dreiecks normalerweise leicht miteinander verschmelzen, wenn sie von einem Objekt reden. Die Informationen über das Objekt und der direkte Bezug bilden eine Einheit. Das ist aber nur durch eine einheitliche Definition eines Begriffs unter allen am Gespräch beteiligten Personen möglich. Der Roboter, der ein reales Objekt finden soll, muss in der Lage sein, die Aussagen des Benutzers so zu interpretieren, dass er das gewünschte Objekt findet.

Aus dem Symbol, der Hypothese dessen, was der Benutzer gesagt hat, wird der Bezug, also eine semantische Beschreibung des Inhalts der Benutzereingabe erstellt, die in einer normierten Form, einer Typed Feature Structure (TFS), für das Dialogsystem verwendet wird. Diese Merkmalsstruktur wurde von (Carpenter, 1992) eingeführt. Außerdem wird eine an die Begriffsstruktur angepasste Grammatik verwendet, die mögliche Sätze umfasst, die der Benutzer verwenden kann. Aus dem wahrscheinlichsten Satz werden die semantisch wichtigen Inhalte extrahiert. So wird aus den verwendeten Verben die Intention gelesen und aus den jeweiligen Subjekten die Objekte. Die Sätze „give me a red cup“ und „bring me the cup in red“ werden in dem Zusammenspiel von Ontologie und Grammatik in die gleiche semantische Form umgewandelt:

```
[object_learn:act_bring
  ITEM [object_learn:obj_item
    TYPE_SLOT [„cup“]
    PROP [object_learn:prp_object
      PROP_SLOT [“red”]
    ]
  ]
]
```

Der Objekterkenner ordnet den erkannten Objekten ebenfalls eine semantische Beschreibung zu. Diese wird zusammen mit einer Repräsentation des Objektes im Objektmodell abgespeichert, damit das Objekt gefunden werden kann. Der genaue Aufbau des Objektmodells wird in Kapitel 4.2.4 beschrieben.

2.1.3. Dialogsystem

Das Dialogsystem verwendet eine Strategie, nach der das System auf den aktuellen Zustand reagiert. Erst wird aus der TFS der Diskurs ausgelesen und zusammen mit den Dialogzielen in einen Zustand des Systems umgewandelt. Dann entscheidet die programmierte oder gelernte Strategie, welcher Schritt notwendig ist, damit mehr Informationen angesammelt oder bestehende Informationen abgesichert werden oder die gewünschte Aktion ausgeführt wird. Die Strategie versucht erst zu ermitteln, was der Benutzer möchte, also was sein Ziel ist und führt den nächsten Schritt aus. Das kann eine Antwort auf eine Frage des Benutzers sein oder eine Frage an den Benutzer. Diese Strategie ist unabhängig von der verwendeten Sprache oder Domäne und entscheidet alleine auf dem Dialogzustand, welcher Dialogschritt ausgeführt wird.

Da dieses System aber neue Objekte lernen soll, kann es sein, dass der Benutzer Wörter verwendet, die nicht in dem Wortschatz des Erkenners vorhanden sind und somit keine semantische Repräsentation für die Weiterverarbeitung vorliegt. Es ist also nötig, dass der Spracherkennung diese Wörter bereits erkennt oder zumindest detektiert. Die unbekanntenen Wörter und Objekte müssen modelliert werden, damit der Dialog darauf reagieren kann.

2.2. Objekterkennung

Um die Umgebung des Roboters wahrzunehmen wird eine Bumblebee (Point Grey, 2005) Kamera von Point Grey verwendet. Es wird eine Firewire Schnittstelle verwendet und die Weiterverarbeitung der Daten erfolgt mit Hilfe des Integrated Vision Toolkit (IVT), einer Open Source Bildverarbeitungsbibliothek von Azad Pedram (Azad P. , 2007). Das IVT ermöglicht einen plattformunabhängigen und sehr komfortablen Zugriff auf OpenCV Funktionalitäten. Die Kameras werden für die Segmentierung von Vordergrund und Hintergrund mit Hilfe eines Schachbrettmusters kalibriert.

2.2.1. SIFT Features

Für die Erkennung von Objekten werden sogenannte SIFT Features detektiert und gespeichert. SIFT Features sind Scale Invariant Feature Transforms, die verwendeten Merkmale der Objekte sind weitgehend unabhängig von der Skalierung, dem Blickwinkel, der Rotation, Lichtverhältnissen und der Position im Bild. In (Lowe, 2004) werden SIFT Features vorgestellt.

Dabei werden zuerst mit allen Skalierungen und an allen Positionen im Bild Schlüsselpunkte gesucht, die sich zu Nachbarpunkten am meisten unterscheiden. Diese sollen sich auch bei verschiedener Rotation und Skalierung immer noch detektieren lassen. Das wird gewährleistet, indem man das Bild mit einer Gaußfunktion faltet und somit einen Skalenraum (scale space) erhält. Auf einem DoG (Difference of Gaussian) mit einem konstant verschobenen scale space werden dann alle Punkte mit ihren Nachbarn verglichen. Wenn dieser Punkt ein Extremwert ist, wird er als mögliches Merkmal gespeichert.

Im zweiten Schritt werden die Schlüsselpunkte aussortiert, die nicht genügend Kontrast zu ihren Nachbarn haben. Deshalb wird für jedes Merkmal ein Deskriptor berechnet, der zum Vergleich mit den anderen dient. Damit man Rotationsinvarianz erreicht, wird jedem Schlüsselpunkt eine Richtung zugewiesen und der Gradient wird dann relativ zu dieser Richtung aufgebaut. Zusammen mit der Richtung ergibt der Gradient den Deskriptor. Die Merkmale werden auf verschiedene Stabilitätskriterien getestet. Sie sind zum Beispiel nur stabil, wenn sie nicht auf Kanten liegen und genügend Kontrast zu ihren Nachbarn haben. Nur stabile Merkmale werden verwendet.

Beim Lernen von Objekten werden diese Merkmale gesucht und falls genügend gefunden wurden, mit dem Bezeichner für ein neues Objekt in der Datenbank abgelegt.

2.2.2. Objekte wiedererkennen

Um Objekte wiederzuerkennen, werden wieder Merkmale berechnet, die mit den Features der Objekte in der Datenbasis verglichen werden. Die gefundenen Objekte werden mit dem Bezeichner und der Klasse, unter der sie in der Datenbank abgelegt wurden in das Objektmodell geschrieben und zusätzlich wird festgehalten, an welcher Stelle das Objekt gefunden wurde. Falls keine bekannten Objekte gefunden wurden, wird überprüft ob sich überhaupt ein Objekt im Sichtbereich befindet. Falls genügend Merkmale gefunden werden, wird das Objekt gelernt. Abbildung 2-3 zeigt ein Objekt mit den Merkmalen, die erkannt wurden. Dabei werden Vordergrund und Hintergrund segmentiert und Merkmale detektiert. Die grünen und roten Punkte stellen die gelernten Merkmale dar.



Abbildung 2-3: Erkennung eines Objektes mit Anzeige der Merkmale

Durch die große Variabilität der Skalierung, Drehung und der Lichtverhältnisse wird eine robuste Wiedererkennung in zufälligen Umgebungen, in denen die Objekte nicht speziell ausgerichtet werden gewährleistet.

3. Bestehende Arbeiten

Um neue Objekte zu lernen sind viele verschiedene Probleme zu lösen. Einzelne Aufgaben wurden bereits mehrfach und sehr tiefgreifend untersucht und werden für die vorliegende Arbeit verwendet und dafür angepasst. Gerade das Lernen von neuen Wörtern für den Spracherkennung ist dringend nötig, um den Namen oder die Art des Objektes zu lernen.

3.1. Erkennen und Lernen neuer Wörter

Thomas Schaaf erforschte in (Schaaf, 2004) das Auftreten, Detektion und Lernen von unbekanntem Wörtern in Dialogen. Diese Arbeit stellt den Ursprung der Umsetzung neue Objekte zu lernen dar. Da es möglich wurde, im laufenden Betrieb durch Nachfragen neue Wörter zu lernen wurden die vorher starren Spracherkennung weiterentwickelt und sind nun in der Lage, sich an den Benutzer und die Umgebung anzupassen.

Dafür werden sogenannte Head-Tail Modelle entwickelt, die es ermöglichen, dass unbekannte Wörter ebenfalls modelliert werden. Im Kopfteil (Head) des Modells für ein Wort befindet sich ein exaktes phonetisches Modell, aber im Schwanzteil (Tail) wird nur noch ein weniger präzises akustisches Modell verwendet. Da die meisten bekannten Wörter bereits nach wenigen Phonemen eindeutig erkannt werden können, reicht es aus, unbekannte Wörter so lange genau zu beschreiben, bis klar ist, dass es ein unbekanntes Wort ist. Durch die genaue Modellierung am Anfang, wird der Suchraum stark eingeschränkt, während die ungenauere Modellierung des Restes immerhin noch die Länge und eine ungefähre Erkennung des Restes liefert und Rechenzeit spart.

Schaaf teilt die unbekanntem Worte im Vorfeld unüberwacht in Gruppen ein. Das so entstandene Sprachmodell hat mehrere verschiedene Klassen von unbekanntem Wörtern, die durch nur zufällig vorkommende Vertreter ihrer Klasse trainiert wurden. Manche Firmennamen kommen beispielsweise häufig in dem Trainingstext vor, aber manch andere tauchen gar nicht auf. Die Firmennamen haben ähnliche Eigenschaften und werden durch das Clustern zusammengefasst. Durch die Ermittlung der Eigenschaften bei der Erkennung ist eine Einordnung des unbekanntem Begriffs möglich.

Weiterhin wurde die Möglichkeit untersucht, den neu gelernten Wörtern eine sinnvolle Wahrscheinlichkeit für die Erkennung zuzuweisen. Damit wurde ein erster Prototyp entwickelt, bei dem es möglich ist, online neue Wörter zu erkennen und zu lernen. Während

dem Dialog mit dem Benutzer kann die Klasse des neuen Wortes erkannt werden, also ob es sich zum Beispiel um einen Ortsnamen oder einen Personennamen handelt. Im weiteren Verlauf des Gesprächs wird mit der Hilfe eines Phonemerkenner das Wort in seine Einzelphoneme aufgeteilt. Dieser Erkennen liefert eine Liste mit wahrscheinlichen Phonemfolgen für dieses Wort zurück. Ein Buchstabiererkennen versucht die Schreibweise des Wortes zu lernen. Wenn all diese Punkte erfolgreich durchgeführt wurden, endet der Lerndialog.

3.2. Barkeeperszenario

In (Holzapfel & Prommer, 2006) wird ein System vorgestellt, in dem es möglich ist, einem Robotersystem zu sagen und zu zeigen, welches Objekt man wünscht. Dafür sind verschiedene Objekte auf einem Tisch platziert und der Benutzer bestellt eines dieser Objekte bei dem Roboter. Dieser findet durch Nachfragen heraus, welches Objekt gemeint ist und serviert dieses. Abbildung 3-1 zeigt das System während der Ausführung.



Abbildung 3-1: Barkeeperszenario

Die Strategie, nach der das System vorgeht wurde durch Methoden des Reinforcement Learnings erstellt und erfolgt automatisch. Dabei kann das System Fragen über fehlende Informationen über das Objekt stellen, bereits erhaltene Informationen bestätigen lassen oder nach Zeigegesten fragen, um herauszufinden welches Objekt gemeint ist. Der Aufbau der Objekte auf dem Tisch ist dabei statisch und muss vorher ausgemessen werden.

Das System fragt so lange nach mehr Informationen über das gewünschte Objekt, bis nur ein vorhandenes Objekt damit übereinstimmt. Dieses wird dann an den Benutzer ausgegeben. Die Objekte in diesem Szenario werden nach ihrer Farbe und Form unterschieden. Um ein Objekt eindeutig zu referenzieren, müssen die Art und die Farbe von dem Benutzer angegeben werden. Um keine falsch erkannten Informationen zu verarbeiten, müssen diese Informationen noch einmal von dem Benutzer bestätigt werden. Wenn es mehrere Instanzen eines Typs mit der gewünschten Farbe gibt, wird außerdem noch gefragt, welche Instanz gewünscht wird. Durch Zeigegesten oder direktes Nachfragen, ob eines der Objekte gewünscht wird, werden die möglichen Instanzen selektiert.

3.3. Lernen neuer Wörter im Dialog

In (Schulz, 2005) wird das Auftreten von unbekanntem Wörtern in natürlich sprachlichen Dialogen zwischen Menschen und einem automatischen System untersucht. Diese Arbeit basiert auf der von Thomas Schaaf (Schaaf, 2004) und führt die Arbeiten fort. Dabei wird vorgestellt, welche Mechanismen notwendig sind, damit neue Wörter bereits im Spracherkenner erfasst werden können. Außerdem wird ein Dialog entworfen, der es ermöglicht, das neue Wort in den Wortschatz mit aufzunehmen. Vor allem der Entwurf der Grammatik und der Ontologie wird untersucht, damit der gesamte Dialog weniger kompliziert gestaltet werden kann.

Der Spracherkenner Janus von (Finke, 1997) und (Soltau, 2001) wird an das Auftreten von OOVs angepasst. In das akustische Modell werden OOVs mit alle möglichen Head-Tail-Modellen aus Kapitel 3.1 eingefügt und trainiert. Um zu gewährleisten, dass OOVs nur erkannt werden, wenn tatsächlich kein anderes Wort passt, werden Strafterme verwendet, die die Wahrscheinlichkeit eines OOVs vermindern.

In der Ontologie werden Konzepte und Unterkonzepte modelliert, die alle bisher bekannten Objekte abdecken. Ein Konzept beschreibt die Eigenschaften eines Objektes, indem es zum Beispiel alle Objekte umfasst, die essbar sind. Dieses Konzept kann weiter unterteilt werden, in kalte und warme Lebensmittel. So wird in der Ontologie immer genauer beschrieben, um was für ein Objekt es sich handelt. Es werden Untersuchungen durchgeführt, wie man die Domäne von Objekten in der Küche sinnvoll unterteilen kann.

In der Grammatik muss das Auftreten von unbekanntem Wörtern ebenfalls modelliert werden. In den Unterkonzepten werden OOVs hinzugefügt, die es ermöglichen, aus der Position, in der das neue Wort auftritt direkt die semantische Information auszulesen. Zudem können OOVs als Oberbegriffe erscheinen.

Ein Dialog zur Klärung der genauen Schreibweise eines OOVs wird vorgestellt. Dieser Dialog ist in Abbildung 3-2 dargestellt. Er beginnt mit der Detektion eines OOVs. Falls es fälschlicherweise detektiert wurde, geht der Dialog in den Fehlerzustand (F) über und wird nicht weiter verfolgt. Falls das OOV korrekt erkannt wurde (B), beginnt der OOV Dialog mit der Information, dass nun geklärt wird, was gesagt wurde. Dabei werden Informationen, die bereits aus der Grammatik abgeleitet wurden bestätigt. So lange noch nicht genügend Informationen gesammelt wurden (U) fragt das System weiter nach, welchem Konzept das Objekt angehört. Falls ein Konzept gefunden wurde, dieses aber Unterkonzepte enthält (G1) werden diese vorgelesen und die Einordnung wird weiter verbessert, bis es keine weitere Verzweigung mehr gibt (G2). Nun wird der Benutzer aufgefordert, den Namen in ein Textfeld einzugeben und der Dialog endet erfolgreich. Um den Namen in der aktuellen Stelle in die Konzepthierarchie einzutragen, kann der Benutzer jederzeit in den Zustand N übergehen und den Dialog dann beenden (Endzustand X). Der Dialog kann jederzeit abgebrochen werden (abort).

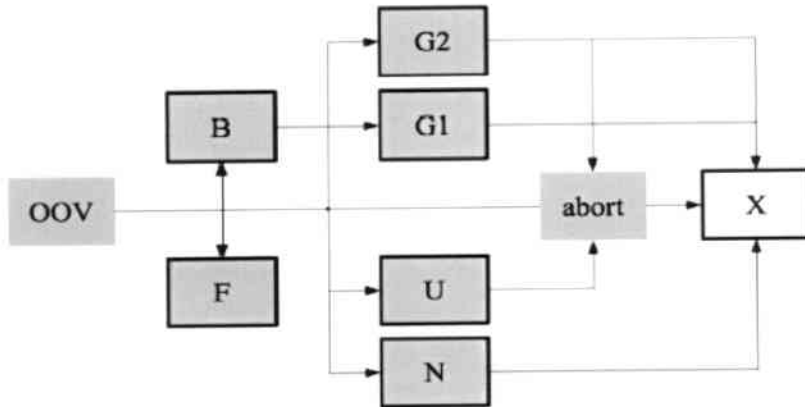


Abbildung 3-2: Ablaufdiagramm des OOV Dialogs

Durch die Vielzahl an Änderungen werden der natürliche Dialog mit einem System und die Erweiterung des Vokabulars mit Angabe der semantischen Bedeutung des Wortes ermöglicht. Das Wort wird in die Ontologie eingetragen und an den notwendigen Stellen im Spracherkennung gespeichert.

3.4. Lernen von Objektbenennungen mit visuellen Prozessen

In (Lömker, 2004) wird eine der wenigen vergleichbaren Arbeiten vorgestellt, in der Objekte in einem Dialog gelernt werden. Es werden sprachliche, visuelle und aktorische Informationen verarbeitet. Ein Dialog führt den Benutzer durch das Lernen und Wiederfinden der Objekte. Dieser Dialog wählt mit Hilfe der Informationen, die aus der Spracherkennung, dem Objekterkennung und dem Gestenerkennung Systemreaktionen aus, damit unbekannte Objekte gelernt werden können. Abbildung 3-3 zeigt den Aufbau des Gesamtsystems.

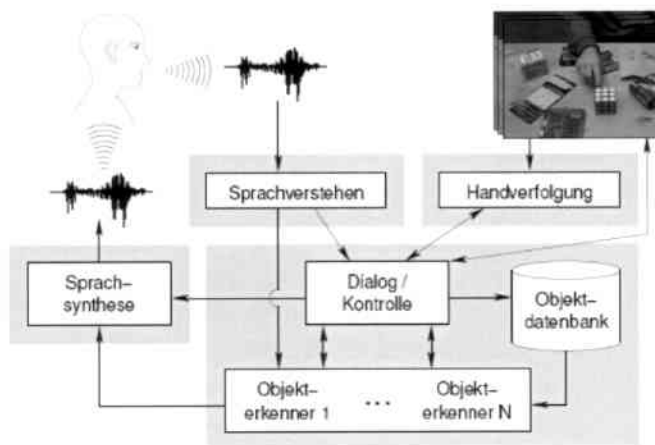


Abbildung 3-3: Architektur des Systems zum Objekte lernen

Das System verwendet ein Nahbesprechungsmikrofon, mehrere Farbkameras, von denen eine mit einem größeren Bildausschnitt die Gesten aufnimmt und ein bis zwei weitere die Objekte beobachtet. Jeder Weg in dem Graphen stellt eine Mögliche Kommunikation dar. Die Übergänge zwischen verschiedenen Zuständen können durch neu gewonnene Informationen der Einzelkomponenten entstehen. Das System kann entweder mit Hilfe von erkannten Handbewegungen gesteuert werden oder der Benutzer kann direkt Objekte referenzieren. Es ist nötig und beabsichtigt, dass der Dialog bei falschen Annahmen von dem Benutzer korrigiert wird. Abbildung 3-4 zeigt den Dialogaufbau.

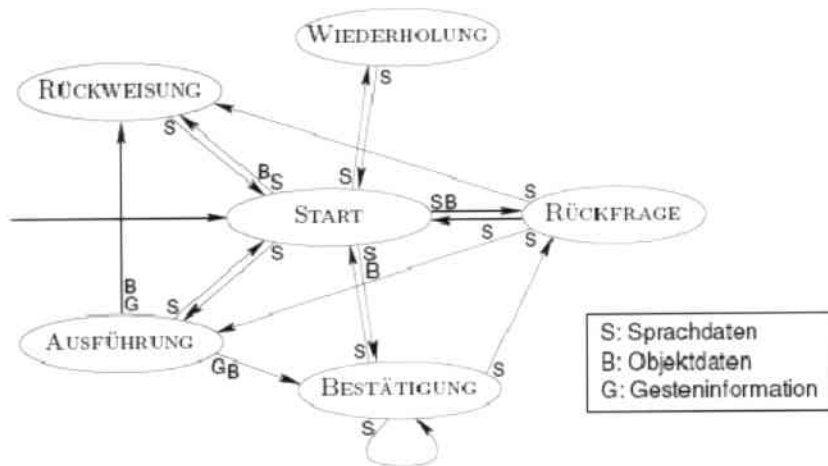


Abbildung 3-4: Dialogablauf

Das System ist in der Lage, Objekte zu lernen und zu lokalisieren, wenn es dazu aufgefordert wird. So kann zum Beispiel „Gib mir den Würfel“ von dem Benutzer gefordert werden. Dabei können nur dem Spracherkner bekannte Bezeichner verwendet werden, das System kann keine neuen Bezeichner lernen. Außerdem wurde die Gestenerkennung intensiv getestet. Dabei wurde festgestellt, dass die zusätzlichen Informationen zur Unterscheidung von gleichen Objekten erfolgreich integriert werden konnten. Um Objekte robuster zu referenzieren, werden Informationen aus der Gestenerkennung, der Objekterkennung und der Spracherkennung fusioniert.

3.5. Zusammenfassung

Im Rahmen des SFB588 (Deutsche Forschungsgemeinschaft, 2001) sind bereits viele Grundlagen erarbeitet worden, auf die in dieser Arbeit zurückgegriffen wird. Teile der Anwendung sind bereits erfolgreich getestet worden und müssen nur an die neue Aufgabe angepasst werden.

Es wurden bereits Richtlinien verfasst, die den Aufbau eines Systems, das in der Lage ist, neue Wörter permanent und korrekt in den Wortschatz hinzuzufügen, ermöglichen. Außerdem besteht ein passendes Szenario, das erweitert werden kann. Durch die

Kombination der einzelnen Projekte kann der größte Kritikpunkt an dem Szenario, das statische Einlesen des Versuchsaufbaus, vermieden werden.

Um das zu erreichen, müssen die Modelle, die den Dialog, den Zustand der Applikation und die Objekte beschreiben angepasst und erweitert werden. Es gibt nur sehr wenige vergleichbare Arbeiten, da es nötig ist sehr gute Einzelkomponenten, die Objekte visuell lernen und robust wiedererkennen können, Sprache erfolgreich zu erkennen und zu verarbeiten und den Dialog zu berechnen und zu führen, zu verwenden. Da diese Kombination bisher noch fast nie erfolgreich durchgeführt wurde, konnte auch kein System, das Objekte lernt vorgestellt werden.

4. Entwurf

Aufbauend auf den vorhandenen Teilprojekten aus den in Kapitel 3 vorgestellten Arbeiten wird das Barkeeperszenario um die Möglichkeit erweitert, unbekannte Wörter und Objekte durch einen Klärungsprozess zu lernen. Dafür werden die Ontologie und die Modelle um die Behandlung neuer Objekte erweitert.

4.1. Gesamtsystem

Es soll ein multimediales System erstellt werden, mit dem Objekte gelernt werden können. Dafür wird der bestehende Aufbau der Dialogsysteme mit Tapas erweitert, wie es Abbildung 4-1 zeigt.

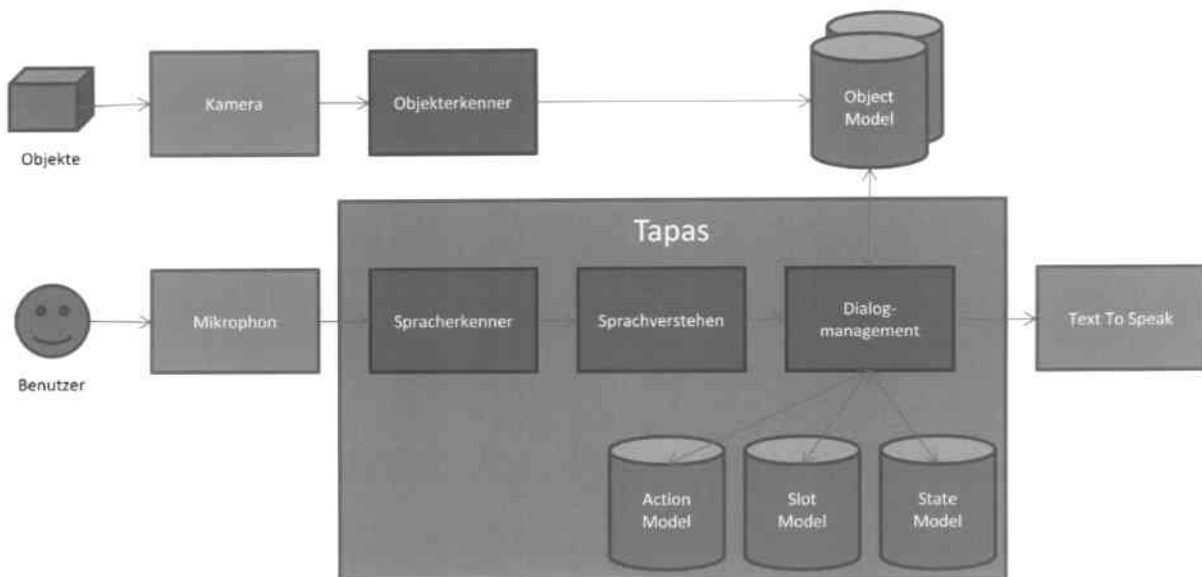


Abbildung 4-1: Allgemeiner Aufbau der Applikation

Die gesamte Situation wird mit Hilfe eines Mikrophones und einer Stereokamera aufgezeichnet. Bei dem Mikrophon handelt es sich um ein Nahbesprechungsmikrophon, das die Hintergrundgeräusche weitgehend außer Acht lässt. Die Kamera nimmt die Objekte auf dem Tisch auf, blickt zu einzelnen Objekten, falls dieses gerade Fokus der Unterhaltung ist und hält Augenkontakt mit dem Benutzer, wenn der Roboter mit diesem spricht. Damit die

Kamera bewegt werden kann, ist sie auf einer Pan-Tilt-Unit (PTU) festgemacht, die mittels einer seriellen Schnittstelle auf 1/10 Grad genau angesteuert werden kann. So kann der Roboter den Kopf drehen und senken, um seine Umwelt wahrzunehmen. Die Sprachausgabe (Text To Speak) erfolgt mit Hilfe von Stereolautsprechern.

Um die gestellte Aufgabe zu lösen, wird ein Spracherkenner verwendet, der mit Hilfe der in Kapitel 3.3 vorgestellten Methoden Hypothesen liefert, die direkt ein „oov“ enthalten, wenn ein unbekanntes Wort auftritt. Die Erkennung kann auf kontinuierlich gesprochenen Sätzen oder auf buchstabierten Wörtern geschehen. Die erwartete Eingabe wird von dem Dialogmanagement bestimmt. Wenn in einem Dialogschritt nach der Buchstabierung eines unbekanntes Wortes gefragt wird, werden in dem Spracherkenner Modelle verwendet, die buchstabierte Eingabe verarbeiten können.

Der Objekterkenner, der wie in Kapitel 2.2 beschrieben, die ankommenden Bilder der Stereokamera verarbeitet, erstellt ein Objektmodell, in das alle gefundenen Objekte mit den Informationen aus der Datenbank über die Art und die Eigenschaft des Objektes und die aktuell gefundene Position geschrieben werden.

Das Sprachverstehen parst mit Hilfe der Grammatik die Eingabe in eine semantische Form. Dabei wird die Angabe von Eigenschaften und Klassen des gewünschten Objektes aufgelöst und gespeichert.

Der Dialogmanager verfolgt das Ziel, dem Benutzer das gewünschte Objekt zu geben. Dafür gibt es verschiedene Teildialoge, die je nach Zustand des Dialoges ausgeführt werden. Der stattfindende Dialogablauf wird in Kapitel 4.3 beschrieben. Die Auswahl der Dialogmodule geschieht durch die Strategie. Diese wählt aufgrund des aktuellen Zustandsmodells das Modul aus, das den nächsten Schritt durchführen soll.

Wenn der Name eines Objektes aus der Buchstabierung heraus richtig erkannt wurde, wird der Bezeichner gelernt. Zudem wird geklärt, wie das neue Objekt in die Ontologie eingetragen werden muss. Das genaue Vorgehen beim Lernen wird in Kapitel 4.3 beschrieben.

4.2. Modelle

Die verwendeten Modelle, die den aktuellen Zustand und die Aktionen des Dialogs, die bereits gesammelten Informationen aus den Äußerungen des Benutzers und die Objekte beschreiben, müssen für dieses Szenario angepasst werden. Hier werden der Aufbau und die Eigenschaften dieser Modelle beschrieben.

4.2.1. Zustandsmodell

Das Zustandsmodell beschreibt die Situation des Dialogs. Dabei wird festgehalten, welche Aktionen bereits wie oft ausgeführt wurden, welche Dialogteile gerade aktiviert sind und

welche Aktion als letztes ausgeführt wurde. Zudem enthält es Zähler für die Gesamtlänge des Dialoges.

Die Strategie nutzt das Zustandsmodell um zu entscheiden, welcher Dialogteil als nächstes ausgeführt wird. In den einzelnen Dialogmodulen wird dann, je nach weiterem Vorgehen, die Situation verändert und im Zustandsmodell gespeichert.

4.2.2. Slot Model

Das Slot Model speichert die Informationen, die bereits über das Objekt der Begierde angesammelt wurde für den Dialogmanager. Da die semantische Eingabe nach jedem Benutzerschnitt neu ist und vorher gegebene Information darin nicht mehr enthalten ist, wird diese für die Applikation gespeichert. So wird die Information, die bereits von dem Benutzer gegeben wurde gespeichert. Für jedes Dialogmodul gibt es Slots, in denen die erfasste Information angesammelt wird.

4.2.3. Aktionsmodell

Im Aktionsmodell werden, sortiert für die Dialogmodule die möglichen Aktionen gespeichert und für die Weiterverarbeitung durch den Dialogmanager in Aktionen von Tapas umgesetzt.

4.2.4. Objektmodell

Das Objektmodell beschreibt ein Objekt. Es wird von dem Objekterkenner erstellt und während dem Dialog von dem Dialogmanager verwendet um zu bestimmen, welches Objekt gewünscht wird. Jedes Objekt wird durch eine Klasse und eine Eigenschaft beschrieben. Somit entsteht wie in Kapitel 2.1.2 vorgestellt eine Ontologie.

Die oberste Klasse an Objekten in dieser Ontologie stellen tragbare Objekte dar. In dem vorgestellten Szenario geht es darum, dem Benutzer ein gewünschtes Objekt zu reichen, also werden andere Objekte, wie ein Tisch oder eine Tür nicht modelliert. Um diese Einschränkung zu umgehen wäre aber auch eine generellere Klasse möglich, die alle Objekte umfasst.

Durch Vererbung und Unterteilung der Klassen werden die Objekte immer besser eingeordnet und spezifiziert. Jede Unterteilung wird durch eine neue Eigenschaft des Objektes bestimmt. So kann zum Beispiel die Klasse aller Getränke in warme oder kalte Getränke unterteilt werden. Die kalten Getränke werden weiter unterteilt in Softgetränke und Säfte. Dadurch entsteht ein Baum von Objekten, durch den die Objekte immer genauer beschrieben werden. In den Blattknoten gibt es dann noch einzelne Ausprägungen, in denen einzelne Orangensäfte auch noch nach dem Firmennamen des Herstellers unterschieden werden kann. Ein Objekt, das abgespeichert wurde, hat einen Bezeichner für die Klasse und einen für die beschreibende Eigenschaft. Damit kann das Objekt in diese Begriffsstruktur

eingordnet werden. Abbildung 4-2 zeigt ein Beispiel für den Aufbau einer solchen Ontologie.

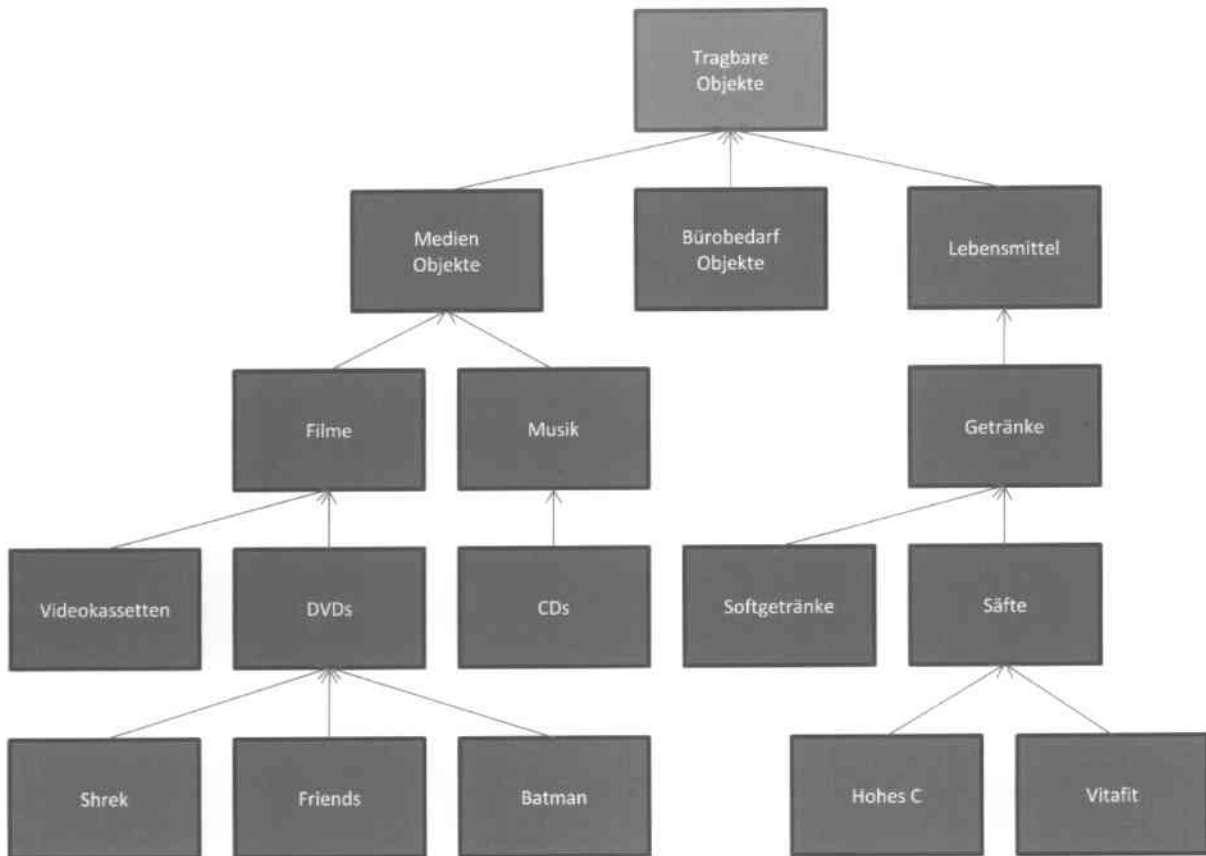


Abbildung 4-2: Ontologie Beispiel

Ein gelerntes Objekt wird von der Objekterkennung als bestimmtes Objekt einer bestimmten Klasse erkannt. Das kann zum Beispiel ein Orangensaft (der Klasse Saft) oder eine DVD des Typs Batman sein. In diesem Fall, könnte man also die DVD durch den Bezeichner „Batman“ und die Klasse „DVD“, die eine Unterklasse von „Film“ und „Medienobjekt“ und „tragbares Objekt“ ist beschreiben. Der Saft wäre zum Beispiel „Hohes C“, also ein Objekt der Klasse „Saft“, die Unterklasse von „Getränk“ und „Lebensmittel“ und „tragbares Objekt“ ist. Wenn im Dialog die Beschreibung „object“ auftritt, kann das noch jedes der beiden Objekte beschreiben. Wenn allerdings als Bezeichnung „movie“ auftaucht, kann diese nur noch die DVD beschreiben.

Zusätzlich enthält das Objektmodell die genaue Position, an der das Objekt in diesem Versuchsaufbau gefunden wurde. Diese Information stammt direkt aus dem Objekterkennung.

Im Laufe eines Klärungsdialogs, welches Objekt gewünscht wird, wird nach und nach immer mehr Information angesammelt, welcher Gegenstand gemeint ist. Im Vergleich zwischen allen gefundenen und dem beschriebenen Objekt wird festgestellt, welche Objekte in der engeren Auswahl des Systems sind. Wenn nach einem Gegenstand direkt gefragt wurde, kann dieser einzeln ausgewählt oder abgewählt werden.

Die Sammlung an Objektmodellen für jeden Gegenstand ergibt alle Objekte, die in diesem Versuchsaufbau vorhanden sind.

4.3. Dialogablauf

In Abbildung 4-3 wird der Ablauf eines Dialogs dargestellt. Die grünen Zustände, sind Dialogmodule, die von dem Dialogmanager ausgewählt werden. Die lila Dreiecke stellen Entscheidungsfragen dar, die je nach Setup und Ablauf des Dialogs unterschiedlich ausfallen können. Der blaue Zustand „Bring an object“ stellt ein Dialogziel dar. Es ist denkbar, dass in einem allgemeinen Versuchsaufbau verschiedene Aufgaben an das System gestellt werden. In diesem Fall gäbe es hier noch eine weitere Aufspaltung in mehrere Dialogziele.

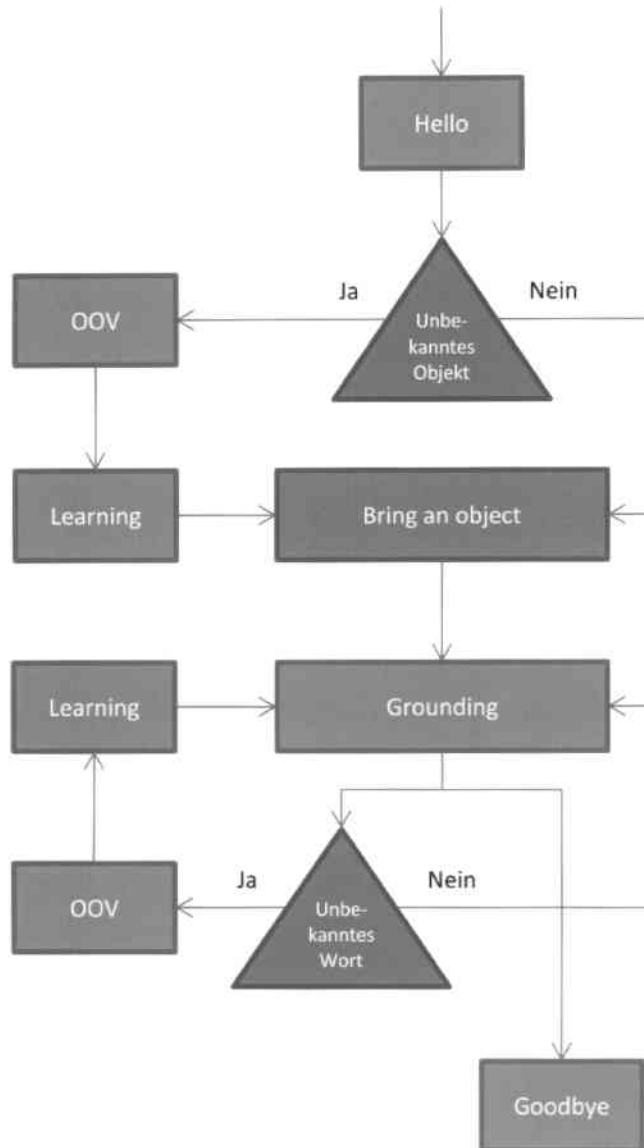


Abbildung 4-3: Dialogablauf

Der Gesamtdialog wird von der Strategie ausgeführt. Dabei wird mit Hilfe des Zustandsmodells entschieden, welches Dialogmodul zur Ausführung kommt. Durch die Trennung der einzelnen Teilaufgaben in Module, können diese einfach ausgewechselt werden. Zudem bleibt der Status eines Moduls erhalten, wenn ein anderer Teildialog eingeschoben werden muss. Außerdem verbessert es die Übersichtlichkeit des Ablaufs und die einzelnen Teile können getrennt voneinander entwickelt werden.

- Der Dialog beginnt bei jedem neuen Durchlauf in dem Dialogmodul „Hello“. Darin wird der Benutzer höflich begrüßt, über die Möglichkeiten des Systems unterrichtet und dann wird der Versuchsaufbau untersucht. Das bedeutet, dass die Kamera auf den Tisch ausgerichtet wird und die Objekterkennung die darauf aufgebauten Gegenstände erkennt. Falls alle Objekte bekannt sind, fragt das System nach, was es für den Benutzer machen kann und bearbeitet weitere Eingaben. Falls ein unbekanntes Objekt detektiert wird, wird dieses erst als neues Objekt abgespeichert, der Benutzer informiert, dass das Objekt nicht bekannt ist und anschließend in den „OOV“ Dialogmodul weiter bestimmt.
- In dem „OOV“ Dialogmodul wird der Benutzer darauf hingewiesen, dass der Name des Objektes nicht erkannt wurde. Daraufhin erfolgt eine Bestimmung der Schreibweise durch einen Buchstabiererkenner. Der Roboter bittet den Benutzer, den Namen des Objektes zu buchstabieren. Dafür wird der Spracherkennung auf Buchstabiererkennung umgeschaltet und eine Hypothese für die Schreibweise wird erstellt. Diese Hypothese wird dem Benutzer mittels des Text To Speak (TTS) Moduls vorgelesen und es wird gefragt, ob das der Name des Objektes ist. Durch die generierte Aussprache des TTS können sich auch leicht falsch erkannte Wörter richtig anhören. So kann zum Beispiel ein als „Batteman“ erkanntes Wort trotzdem als „Batman“ angenommen werden.
- Das Lernen geschieht in dem Dialogmodul „Learning“. Darin wird bestimmt, zu welcher Klasse das Objekt gehört und wo in der in Kapitel 4.2.4 beschriebenen Ontologie das Objekt eingefügt werden soll. In der erfassten Semantik kann eventuell die Objektklasse abgelesen werden. Dies kann dann ein Einstiegspunkt für den Dialog sein, um herauszufinden, wo das Objekt in der Ontologie eingetragen werden soll. Der nötige Dialog wird in Kapitel 0 beschrieben. Der gelernte Objektbezeichner und die Klasse des Objektes werden in der Ontologie, dem aktuellen Objektmodell und der Datenbasis für den Objekterkennung gespeichert. Zudem wird in der Grammatik des Spracherkenners an der Stelle für Objektbezeichner beziehungsweise der Objektklasse die Möglichkeit des neuen Wortes hinzugefügt und ein Wörterbucheintrag erstellt. Das gelernte Wort wird mit der von der TTS generierten Aussprache ins Wörterbuch des Spracherkenners aufgenommen, damit dieser das Wort später auch erkennen kann. Damit sind der Spracherkennung, die Modelle für den Dialogmanager und für die Objektverwaltung für die Verwendung des neuen Objektes vorbereitet. Wenn alle Objekte bekannt sind, wird mit dem normalen Dialog weitergemacht.
- Das „Grounding“ Modul umfasst den ganzen Dialog um das gewünschte Objekt zu erfassen. Dabei erklärt der Benutzer das gewünschte Objekt oder gibt Antworten auf

direkte Fragen des Systems. Auch während dem Klärungsdialog kann ein unbekanntes Wort auftauchen. In diesem Fall kann dieses aber nur ein unbekannter Bezeichner für ein bekanntes Objekt sein. Das Lernen beschränkt sich also auf das Abspeichern von einem Synonym, das ein bekanntes Objekt beschreibt. Trotzdem ist es nötig, die Klasse und die Eigenschaften des Objektes mit Hilfe des „Learning“ Dialoges zu bestimmen

- Wenn ein Objekt eindeutig bestimmt wurde, geht das System in den „Goodbye“ Dialog über, gibt den Gegenstand aus und verabschiedet sich von dem Benutzer. Falls der Dialog zu lange dauert, oder von dem Benutzer gestoppt wurde, wird dieser Fehler kommuniziert und das System bricht ab.

Der Dialog kann mit verschiedenen Voraussetzungen starten. Je nachdem wird die Strategie einen anderen Ablauf durchlaufen. Nach der Begrüßung und Erfassung der Objekte wird immer nach dem gewünschten Objekt gefragt. Dabei kann ein neuer Objektbezeichner auftreten und es kann zu folgenden Fällen kommen:

- Ein unbekanntes Objekt
In diesem Fall wird das Objekt direkt bei der Erkennung als unbekanntes Objekt neu in die Datenbank des Objekterkenners eingetragen. Das System versucht sofort, das Objekt zu lernen und fragt den Benutzer, wie dieses Objekt heißt und was es ist. Dadurch soll sichergestellt werden, dass das System alle Objekte kennt. Der Objekterkenner speichert den Gegenstand auf jeden Fall als neues Objekt ab und wird den Benutzer auch wieder über den Gegenstand ausfragen, wenn er es wieder sieht und der Lerndialog bisher nicht erfolgreich durchgeführt werden konnte.
- Eine bekannte Bezeichnung
Dies stellt den einfachsten Fall dar. Hier wird der Grounding Dialog so lange durchgeführt, bis klar ist, welches Objekt gewünscht wird. Dieses wird serviert und der Dialog beendet.
- Eine unbekannte Bezeichnung
Hier wird sich das System damit konfrontiert sehen, dass ein Synonym zu einem bisher bereits bekannten Objekt verwendet wurde, um es zu beschreiben. Die Aufgabe ist nun, herauszufinden, welches Objekt damit gemeint ist und den neuen Namen dafür zu lernen. Als erstes wird der Bezeichner geklärt. Dies geschieht mit einem OOV Dialog. Daraufhin wird das Objekt in die Ontologie einsortiert. Dafür wird die Hierarchie schrittweise nach unten durchlaufen um zu verfeinern, welcher Klasse das Objekt angehört. Wenn das geklärt ist, kann dieses Objekt direkt serviert werden und der Dialog endet.

Im Folgenden werden zwei Beispieldialoge erläutert. Es handelt sich um einen Dialog, der das Ziel hat, ein Objekt zu reichen und einen, der ein neues Objekt lernen soll. In dem ersten Dialog wird der Benutzer erst begrüßt und dann die vorhandenen Objekte eingelesen. Da dem System alle Objekte bekannt sind, fragt es direkt nach, welches Objekt gereicht werden soll. Der Bezeichner „movie“ trifft noch auf mehrere Objekte zu und deshalb wird genauer nachgefragt, welche DVD gewünscht wird. Durch die genauere Frage wird klar, dass eine

weitere Beschreibung, zum Beispiel durch den Filmtiteln benötigt wird. Nach Auswertung dieser Information wird das Objekt gereicht. Der Ablauf wird in Abbildung 4-4 dargestellt.

```
System: Hello my name is Robbi. How are you?

User: Hi I am Daniel, I am fine!

System: Before I can help you, I will have a look at the table to
recognize the objects on it.

User: Ok.

System: I found a Batman DVD and a Spiderman DVD on the table.

User: Great.

System: Which item can I serve you?

User: Please give me the movie.

System: Is your item a DVD?

User: Yes.

System: What kind of DVD is the item? Does it have a color or a
special name?

User: It is a Batman DVD.

System: Thank you for your patience. I will serve you the Batman DVD
now.

User: Thank you very much.

System: Have a nice day.
```

Abbildung 4-4: Dialogbeispiel Objekt reichen

Im zweiten Dialog findet der Erkenner ein Objekt, das er nicht kennt. Als Grundlage wird die Ontologie, die in Abbildung 4-2 dargestellt ist, verwendet. In diesem Beispieldialog wird ein unbekanntes Objekt auf dem Tisch gefunden. Der Roboter meldet dieses Ereignis an den Benutzer und fragt höflich, was dieses Objekt ist. Erst wird der Name gelernt, indem der Benutzer diesen buchstabiert. Der erkannte Name wird wiederholt, bis der Benutzer diesen Namen annimmt. Daraufhin wird die Art des Objektes erfragt, indem das System in der Ontologie absteigend fragt, in welche Klasse das Objekt eingeordnet werden kann. Nach erfolgreicher Klassifizierung wird das Objekt gelernt und der Benutzer wieder gefragt, welches Objekt er haben möchte. Dieser Dialog ist in Abbildung 4-5 dargestellt.

```
System: Hello my name is Robbi. How are you?

User: Hi. I am fine!

System: Before I can help you, I will have a look at the table to
recognize the objects on it.

System: I don't know this object. (system looks at the unknown DVD of
the movie "Troy")

System: Please spell the name of this object.

User: T R O Y

System: Is Troy the correct name?

User: Yes.

System: What kind of object is this? Is it a media object, an object
from the bureau or food?

User: It is a media object.

System: Is it a movie or music?

User: It is a movie.

System: Is it a DVD or a videotape?

User: A DVD.

System: Thank you. I have learned this object.

System: I found a Troy DVD and a Spiderman DVD on the table.

System: Which item can I serve you?

User: Please give me the Troy DVD.

System: Thank you for your patience. I will serve you the Troy DVD
now.

User: Thanks!

System: Have a nice day.
```

Abbildung 4-5: Dialogbeispiel Objekt lernen

In dem ganzen Dialog wird darauf geachtet, dass das System höflich mit dem Benutzer spricht und Augenkontakt mit dem Benutzer hält, wenn es mit ihm spricht. Wenn das System ein bestimmtes Objekt benennt, dreht es seinen Kopf zu dem Objekt um zu verdeutlichen, was der Fokus der Unterhaltung ist. Außerdem wiederholt das System gesammelte Informationen und verwendet immer Objektbezeichner und die Klasse, um einen Gegenstand zu referenzieren.

4.4. Fehlerarten im Dialog

Während dem in Kapitel 4.3 dargestellten Dialog können verschiedene Fehler auftreten. Die genaue Verteilung und Häufigkeit der Fehler wurde in Anbetracht der Kürze der Arbeit nicht untersucht.

Der gesamte Ablauf der Versuche stützt sich auf eine korrekte visuelle Erkennung aller Objekte auf dem Tisch. Der Roboter gibt nach dem Erkennen per Sprache aus, was er alles auf dem Tisch gefunden hat. Dadurch wird für den Benutzer ersichtlich, ob der Roboter alle Objekte richtig erkannt hat. Dadurch kann dieser den Roboter darauf hinweisen und die Erkennung kann wiederholt werden. Wenn ein unbekanntes Objekt detektiert wird, wird dieses sofort neu gelernt. Dadurch geht der Dialog immer davon aus, dass alle vorhandenen Objekte im Objektmodell beschrieben sind.

In dem Fall, dass nur bekannte Objekte auf dem Tisch gefunden wurden, können zwei Fehler auftreten:

- Ein Objekt wurde fälschlicherweise als ein anderes erkannt.
Der Roboter wird ein Objekt falsch benennen.
- Ein Objekt steht zwar auf dem Tisch, wurde aber gar nicht erkannt.
Der Roboter übergeht dieses Objekt und es fehlt in der Liste der erkannten Objekte.

In dem Fall, dass mindestens ein Objekt gefunden wurde, das nicht erkannt wurde, aber als unbekannter Gegenstand detektiert wurde, wird das Objekt gelernt. Dabei kann dieser Fehler auftreten:

- Ein bekanntes Objekt wurde nicht erkannt, sondern als unbekanntes Objekt erkannt.
In diesem Fall wird das falsch erkannte Objekt zum Fokus des Klärungsdialoges. Das System lernt einen neuen Namen für ein bisher schon bekanntes Objekt und lernt somit ein Synonym.

Falls aber der Klärungsdialog abgebrochen wird, kann es sein, dass ein unbekanntes Objekt weiter vorhanden ist ohne dass das System gleich lernt, was dieses ist. In diesem Fall wird das noch nicht fertig gelernte Objekt beim nächsten Mal, wenn es gefunden wird wieder Fokus des Dialogs und der Roboter wird wieder nachfragen, was dieses ist, da der Dialog zum Finden eines gewünschten Objekts erst starten kann, wenn alle Objekte bekannt sind.

5. Implementierung

Die bestehenden Arbeiten aus Kapitel 3 wurden alle auf verschiedenen, zum Teil veralteten Versionen von Tapas durchgeführt. Dadurch entstand eine Vielzahl an nötigen Anpassungen und Veränderungen dieser eigentlich fertigen Teilprojekte, die einen Großteil der Implementierungszeit einnehmen sollten. In Kapitel 4 wird das erstellte System vorgestellt, hier werden Details der Umsetzung erläutert.

Die entwickelte Applikation läuft als Wizard of Oz (WOZ) Studie ab. Dabei werden Teile des Dialogablaufs durch einen Administrator gesteuert. Aus der Sicht des Benutzers läuft der komplette Dialog automatisch ab, damit eine unverzerrte Aufnahme der Aktionen und Reaktionen des Probanden möglich ist. Dieser soll ausschließlich mit dem System interagieren.

5.1. Szenario

Um ein vergleichbares System zu implementieren, wurde der Ansatz eines Barkeeper Roboters weiter entwickelt. In der Kürze der Zeit konnten aber nicht alle Ansätze bis zur Ausführung integriert werden.

- Der Objekterkenner liefert eine Hypothese davon, was alles auf dem Tisch steht. Er erkennt beliebig viele bekannte Objekte wieder. Falls er aber ein unbekanntes Objekt lernen soll, wird dieses nur detektiert, wenn keine anderen Objekte im Bild vorhanden sind. Diese Einschränkung kann leicht durch eine Selektion der einzelnen gefundenen Merkmale in Merkmale, die durch bekannte Objekte gebunden sind und freie Merkmale aufgelöst werden. In Anbetracht der Zeit wurde das vorliegende System aber so gestaltet, dass es bisher nur einzelne unbekannte Objekte lernt.
- Das System kann nur ein Objekt nach dem anderen lernen. Das System wird nachfragen, um welches Objekt es sich bei einem unbekanntem Objekt handelt. Dieses wird mit einem Namen und einer Klasse versehen, wenn der Klärungsdialog erfolgreich abgeschlossen wird. Wenn er das nicht wird, wird das System bei der nächsten Detektion wieder nach dem Objekt fragen und dieses versuchen zu lernen. Falls es zu diesem Zeitpunkt zwei verschiedene unbekannte Objekte hat, würden beide Objekte diesen Namen und diese Klasse zugewiesen bekommen. Um das zu

umgehen ist eine genauere, variable Zuweisung von Bezeichnern für neue Objekte notwendig.

- Die Bestimmung der Klasse eines neuen Objektes wird gerade soweit verkürzt, dass Objekte nur als tragbare Objekte gespeichert werden. Diese sind in der Hierarchie immer genereller als die vorhandenen Objekte und werden somit auch wieder gefunden, wenn ein Unterkonzept beschrieben wird. Der gelernte Bezeichner wird als Eigenschaft dieser Klasse abgespeichert. Um diesen Mangel zu beheben ist es nötig, das Dialogmodul „Learning“ weiter zu verfeinern, damit dort auch mehrere Schritte ausgeführt werden, in denen die Klasse bestimmt wird. Die Vorgehensweise, wie das geschehen kann wird in Kapitel 4.3 beschrieben.
- In der WOZ Studie, die zur Gewinnung der Daten führt, wird der Dialogteil „OOV“ durch einen Menschen simuliert. Dessen Aufgabe ist es, den Benutzer dazu aufzufordern, den Namen zu buchstabieren. Wenn eine Buchstabierung erkannt wurde, wird der so gewonnene Name vorgelesen und nachgefragt, ob dieser richtig erkannt wurde. Wenn der Benutzer diesen bestätigt, endet der OOV Dialog und der automatische Dialog wird weitergeführt.

Das realisierte System kann somit die Grundzüge eines Dialoges, in dem es möglich ist, Objekte zu lernen und wiederzufinden durchlaufen. Eine verfeinerte und weiterführende Erkennung wurde nur erdacht, konnte aber in der vorhandenen Zeit nicht umgesetzt werden. So wurden zum Beispiel Fehler, die in bereits entwickelten Einzelkomponenten auftreten können nicht speziell abgefangen. Wenn also der Objekterkenner ein Objekt auf dem Tisch nicht erkennt, wird dieser Dialogdurchlauf mit einem lückenhaften Objektmodell durchgeführt.

Der Roboter führt mit seinen Fragen und Anweisungen durch den Dialog. Er reagiert selber auf die Wahrnehmung seiner Umgebung und versucht, Wissenslücken selbstständig zu schließen, indem er direkt nach unbekanntem Objekten nachfragt. Wie ein neugieriges Kind fragt er nach allem was er nicht kennt. Dadurch soll vermieden werden, dass der Benutzer mit einem System interagiert, das nur einen Teil seiner Umgebung gelernt hat. Durch diese Systeminitiative führt der Roboter höflich durch die Unterhaltung. Im Gegensatz dazu wäre eine Benutzerinitiative denkbar gewesen, bei der der Benutzer die Führung des Dialoges übernimmt. Das erfordert aber eine weitaus längere Einarbeitung des Benutzers, da dieser dann genau wissen muss, wie er sein Ziel erreichen kann. Durch die Führung des Roboters wird dem Benutzer in jedem Schritt gesagt, was er machen sollte um das Ziel zu erreichen.

5.2. Experimente

Um zu testen, ob das System in der Lage ist, neue Objekte zu lernen, wurden mit verschiedenen Testpersonen mehrere Dialoge in dem erweiterten Barkeeperszenario, das in Kapitel 4 beschrieben wird, durchgeführt. Dabei wurde die Auswahl an Objekten von Durchlauf zu Durchlauf verändert, um das System und den Benutzer vor neue Aufgaben zu stellen.

Der Benutzer bekam die Hinweise, dass er sich ein Objekt von dem Tisch aussuchen möchte. Dieses soll er im Laufe des Gesprächs von dem Roboter einfordern. Der Roboter stellt als Dienstgeber also eine Funktionalität zur Verfügung und kann, wie ein Barkeeper Objekte servieren.

Allen Benutzern wurde diese Information gegeben:

Versuchsaufbau

Der hier stattfindende Versuch beschäftigt sich mit dem Dialog zwischen Menschen und Robotern. Dafür wird auf dem kleinen Tisch vor dem Roboter eine Reihe von Gegenständen aufgebaut. Ihre Aufgabe ist es, sich eines der Objekte auszusuchen und den Roboter darum zu bitten.

Die Aufgabe des Roboters ist es, das gewünschte Objekt zu finden und ihnen zu geben. Dafür wird er einen Dialog auf Englisch mit Ihnen führen und Ihnen verschiedene Fragen stellen. Bitte benutzen Sie das Handmikrofon vor Ihnen um mit dem Roboter zu sprechen.

Wenn der Roboter ein Objekt auf dem Tisch nicht kennt, wird er danach fragen und versuchen den Namen zu lernen. Später kann der Roboter dieses Objekt auch wieder erkennen. Als erstes wird ein Beispieldialog durchgeführt, in dem der Roboter bereits alle Objekt auf dem Tisch kennt. Suchen Sie sich also bitte eines der Objekte aus und es kann schon losgehen.

Viel Erfolg!

Als erster Dialog wurde bei jedem Probanden ein Setup gewählt, in dem alle Objekte bekannt waren. Dadurch konnten sich die Benutzer an den Ablauf gewöhnen und einmal durchlaufen ohne spezielle Aufgaben zu erfüllen. Danach wurden verschiedene Setups durchlaufen, um verschiedene Objekt zu lernen und diese im weiteren Verlauf des Tests wieder zu erkennen.

Abbildung 5-1 zeigt eine Aufnahme, die der Roboter von dem Tisch macht, auf dem er die Objekte erwartet. Mit Hilfe des hier gezeigten linken und des rechten Bildes erkennt er die Objekte auf dem Tisch, wie in Kapitel 2.2 beschrieben. Mit blauen Rahmen werden die Objekte bei der Erkennung umrandet und mit einem roten Punkt markiert, der die Koordinaten des Objektes angibt, an der der Roboter das Zentrum des Objektes erkannt hat.



Abbildung 5-1: Aufnahme des Roboters während dem Dialog

Insgesamt wurden 56 Dialoge mit acht Personen durchlaufen. Dabei wurden Logs erstellt, die aufzeichnen, was gesagt und erkannt wurde und welche Objekte aufgestellt wurden. Der genaue Ablauf dessen, was das System erkannt hat und wie es reagiert hat wurde aufgenommen.

6. Ergebnisse und Diskussion

Mit Hilfe der Experimente aus Kapitel 5.2 wird die Fähigkeit des Systems, neue Objekte zu lernen und sich an eine Umgebung anzupassen bewertet.

6.1. Ergebnisse

In den 56 Durchläufen mit acht verschiedenen Probanden wurden 106 Objekte verwendet. In jedem Durchlauf wurden verschiedene Objekte verwendet, um den Roboter und den Probanden mit einer unterschiedlichen Ausgangssituation zu konfrontieren.

6.1.1. Gesamtsystem

Der gesamte Dialog wurde 21 Mal durchgeführt, damit ein Objekt gelernt wird. 35 Mal wurde ein Objekt von dem Roboter eingefordert. Die Ergebnisse sind in Tabelle 6-1 dargestellt.

Das Lernen eines Objektes wird als erfolgreich angenommen, wenn während dem Dialog das System erfolgreich visuelle Merkmale extrahiert und speichert um das Objekt wieder zu finden und anschließend der Name durch Buchstabierung so gelernt wird, dass der Benutzer diesen annimmt. Die Misserfolge beim Lernen entstanden alle, weil das System nicht in der Lage war, den buchstabierten Namen des Objektes erfolgreich zu lernen. Deshalb wurde der Buchstabiererkenner im Laufe der Experimente verbessert.

Ein Objekt zu reichen erfordert, dass das gewünschte Objekt erfolgreich von der Objekterkennung gefunden wird und im darauf folgenden Dialog geklärt wird, welches Objekt gewünscht wird. Objekte konnten nicht gereicht werden, wenn das Objekt nicht gefunden wurde oder der Bezeichner nicht verstanden wurde.

Gesamtsystem	Anzahl	erfolgreich		nicht erfolgreich	
Objekt lernen	21	15	71%	6	29%
Objekt reichen	35	27	77%	8	23%
Gesamt	56	42	75%	14	25%

Tabelle 6-1: Ergebnisse Gesamtsystem

6.1.2. Objekterkennung

Der Objekterkennner liefert eine Hypothese der Objekte, die auf dem Tisch vorhanden sind und die von dem Roboter erreicht werden können. Dabei hatte das System einige bereits trainierte Objekte in der Datenbank. Bei diesen Objekten wurden in einem Bild die Objektgrenzen von Hand angegeben und auf diesem Bereich wurden Merkmale gelernt. Bei Objekten, die während den Dialogen gelernt werden, wird zuerst durch Vordergrundsegmentierung ein mögliches Objekt bestimmt und dann werden die Merkmale in diesen Grenzen berechnet und gelernt. Abbildung 6-1 zeigt ein Objekt, das während einem Dialog gelernt wurde mit den Objektgrenzen und den Punkten, die als Merkmale gespeichert wurden.



Abbildung 6-1: Lernen eines unbekanntes Objektes

Beim Wiederfinden wird im aktuellen Bild wieder eine Vordergrundsegmentierung durchgeführt und Merkmale berechnet. Diese Merkmale werden mit den Objekten aus der Datenbank verglichen um eine Aussage über die Objekte auf dem Tisch machen zu können.

Um unbekannte Objekte zu finden, werden, falls keine Objekte gefunden werden, aber trotzdem viele Merkmale detektiert werden diese Merkmale einem neuen Objekt zugeschrieben. Während den Tests wurden alle unbekanntes Objekte im Bild gefunden und konnten gelernt werden.

Beim Wiederfinden von automatisch gelernten Objekten traten mehr Fehler auf, da auch Merkmale auf anderen Objekten im Vordergrund, wie zum Beispiel die Türklinke am rechten Bildrand gelernt wurden oder weil nur wenige Merkmale gelernt wurden, weil das Objekt beim Zeitpunkt des Lernens zum Teil durch die Beschichtung und die Sonneneinstrahlung sehr helle Flächen aufwies. Dadurch werden also auch falsche oder nur wenige Merkmale diesen Objekten zugewiesen und das Wiedererkennen wird erschwert. Abbildung 6-2 zeigt ein Beispiel für dieses Fehlverhalten.



Abbildung 6-2: Merkmale beim Lernen eines unbekanntes Objektes

Während allen Tests kam es nie vor, dass ein Objekt fälschlicherweise als ein anderes erkannt wurde. Die Ergebnisse sind in Tabelle 6-2 dargestellt.

Objekterkenner	Anzahl	richtig erkannt		nicht erkannt		falsch erkannt
		Anzahl	Prozent	Anzahl	Prozent	
alle Objekte	106	97	92%	9	8%	0%
trainierte Objekte	46	45	98%	1	2%	0%
unbekannte Objekte	15	15	100%	0	0%	0%
gelernte Objekte	45	37	82%	8	18%	0%

Tabelle 6-2: Ergebnisse Objekterkennung

6.1.3. Buchstabiererkennung

Um die unbekanntes Objektbezeichner zu lernen wurde Buchstabiererkennung eingesetzt. Um ein Wort erfolgreich zu lernen, fordert der Roboter den Benutzer auf, dieses zu buchstabieren. Daraufhin fragt der Roboter den Benutzer, ob das Wort, so wie es verstanden wurde korrekt sei. Wenn der Benutzer das bejaht, wird der Dialog weitergeführt und das Objekt unter diesem Namen gelernt. Hier fällt auf, dass sich die Anzahl der tatsächlich korrekten von der Anzahl der angenommenen Wörter stark unterscheidet. Da der Roboter das verstandene Wort vorliest wird es für den Benutzer schwer zu unterscheiden, ob das Wort wirklich korrekt war oder nicht. Die Erkennung wurde als erfolgreich bewertet, wenn innerhalb von drei Mal Buchstabieren das richtige Wort erkannt wurde. Wenn mehr Buchstabierungen nötig waren oder der Erkennen nie das richtig geschriebene Wort erkannt hat, wurde das Ergebnis negativ bewertet. Für alle befragten Nutzer waren drei Wiederholungen vertretbar. Da am Anfang der Tests ein Sprachmodell verwendet wurde, das für die Erkennung von deutschen Namen mit einem deutschen Text trainiert wurde, waren die Ergebnisse nicht sehr gut und die Dialoge sehr lang. Daraufhin wurde das Sprachmodell noch

einmal mit einem englischen Text trainiert um die englischen Objektbezeichner besser zu verstehen. Die Erkennung wurde dadurch deutlich besser. Die Ergebnisse mit dem neuen Sprachmodell des Erkenners werden in Tabelle 6-3 aufgezeigt.

Buchstabiererkenner mit bis zu drei Versuchen	Anzahl	erfolgreich		nicht erfolgreich	
		Anzahl	Prozent	Anzahl	Prozent
Buchstabierung korrekt	7	5	71%	2	29%
Buchstabierung angenommen	7	6	86%	1	14%

Tabelle 6-3: Ergebnisse Buchstabiererkennung mit englischem Training

Dadurch verbesserte sich das Gesamtsystem erheblich. Die Ergebnisse, die nur mit dem neuen Buchstabiererkenner erzielt wurden, sind in Tabelle 6-4 dargestellt.

Gesamtsystem	Anzahl	erfolgreich		nicht erfolgreich	
		Anzahl	Prozent	Anzahl	Prozent
Objekt lernen	7	6	86%	1	14%
Objekt reichen	11	9	82%	2	18%
Gesamt	18	15	83%	3	17%

Tabelle 6-4: Ergebnisse Gesamtsystem mit englischem Training

Die Güte der Buchstabiererkennung kann wie die normale Spracherkennung mit Fehlerraten gemessen werden. Dabei wird die Wortfehlerrate, die angibt, wie viele Worte in dem Satz falsch verstanden wurden zur Buchstabenfehlerrate und die Satzakkuratheit wird zur Wortakkuratheit.

Die Buchstabenfehlerrate ergibt sich aus

I: Anzahl der Einfügungen von Buchstaben

D: Anzahl der Löschungen von Buchstaben

S: Anzahl der Ersetzungen von Buchstaben

N: Länge des ursprünglichen Wortes

$$\frac{I + D + S}{N}$$

Bei 13 Buchstabierungen wurden 19,7 % der Buchstaben falsch erkannt.

Die Wortakkuratheit gibt an, wie viele Worte in der Gesamtheit der Tests richtig erkannt wurden. 61,9 % der Worte wurden richtig erkannt. Dabei wurde die Erkennung der einzelnen verschiedenen Wörter berechnet und dann normiert auf die Wörter die Gesamterkennung berechnet. Dadurch wurden die Wörter, die schlechter erkannt wurden, nur genauso stark gewichtet, wie Wörter, die gut erkannt wurden, auch wenn diese mehrfach buchstabiert wurden.

6.1.4. Spracherkennung

Die Spracherkennung musste insgesamt 115 Objektbezeichner und Objektklassen verstehen. Meistens wurden Objektbezeichner und Objektklasse verwendet, um das Objekt zu referenzieren. Zum Teil wurde aber auch nur ein Teil verwendet und die fehlende Information musste später zusammengesammelt werden. Oft wählten die Benutzer die gleiche Formulierung, die auch der Roboter verwendete um die Gegenstände zu benennen. Erst nachdem die Benutzer ermutigt wurden, auch andere Bezeichnungen zu verwenden, wurde diese Möglichkeit genutzt. Diese Variabilität wurde dann sehr geschätzt und es sorgte für Erstaunen, dass der Roboter das Objekt tatsächlich auch mit anderen Bezeichnungen referenzieren kann. Die Ergebnisse sind in Tabelle 6-5 dargestellt.

Spracherkenner	Anzahl	richtig erkannt		nicht erkannt	
		Anzahl	Prozent	Anzahl	Prozent
Objektbezeichner	61	44	72%	17	28%
Objektklasse	54	44	81%	10	19%
Gesamt	115	88	77%	27	23%

Tabelle 6-5: Ergebnisse Spracherkennung

6.2. Analyse der Ergebnisse

Die am Anfang eingesetzte Buchstabiererkennung war zu schlecht, um einen flüssigen Dialog zu führen. Dadurch, dass der Erkenner mit deutschen Texten trainiert wurde, aber englische Wörter erkennen sollte wurden auch sehr deutlich buchstabierte Wörter zu sinnlosen Buchstabenreihen, da die korrekte Aneinanderreihung für einen deutschen Text zu unwahrscheinlich war. Dadurch wurden viele Dialoge deutlich länger als das nötig war und die Erfolgsrate beim Lernen neuer Objekte wurde sehr viel schlechter als dies eigentlich möglich wäre. Nach dem Training des Buchstabiererkenners mit einem englischen Text war die Erkennung der Bezeichner meistens innerhalb der ersten drei Buchstabierungen erfolgreich. Dadurch wurden die Dialoge deutlich kürzer und insgesamt erfolgreicher.

Alle Probanden hatten nach den Versuchen die Möglichkeit, sich kritisch über das System zu äußern. Es wurde kein spezieller Fragebogen ausgeteilt, sondern in einer Diskussion auf Stärken und Schwächen des Systems eingegangen. Dadurch kamen einige sehr persönliche Empfindungen zum Tragen.

Die gute und robuste Objekterkennung wurde mehrfach gelobt, da die meisten Objekte erfolgreich erkannt wurden. Hier ist es aber nötig, dass der Benutzer das System darauf hinweisen kann und der Roboter noch einmal auf den Tisch schaut, wenn ein Fehler auftritt. Gerade wenn eines der Objekte nicht detektiert wurde sollte man ihn dazu auffordern können. Bemerkenswert ist, dass der vorher angedachte Fehler, dass ein Objekt als ein anderes erkannt wird gar nicht aufgetreten ist.

Einer der Probanden hatte den ganzen Versuch über das Gefühl, dass er sehr offen und sehr frei mit dem System interagieren kann. Immer wieder schweifte er von den Fragen und dem erdachten Ablauf ab, stellte Zwischenfragen über andere Objekte und wechselte während dem Gespräch seine Objektauswahl. Dadurch musste der Dialog mehrmals erfolglos abgebrochen werden. In der Diskussion kam auf, dass er dachte, dass ein System, das die Einzelschritte, wie Objekte erkennen und Objekte lernen beherrscht, auch diese Funktionalitäten einzeln zur Verfügung stellen muss. Da in diesem Szenario allerdings nur die Möglichkeit, eines der erkannten Objekte zu lernen oder zu reichen gegeben ist, war das System damit überfordert. Um das anzubieten, kann man verschiedene Dialogziele definieren, die dann einzelne Funktionalitäten erfüllen.

Ein anderer Proband bemängelte, dass der Roboter „Robbi“ nicht höflich genug sei. Mit einem netteren Roboter sei es angenehmer zu reden und es würde mehr Spaß machen, ihm Sachen beizubringen. Daraufhin wurde die Phrase „How are you?“ an die Begrüßung angehängt. Interessanterweise bemängelten spätere Tester, dass ein Roboter überhaupt nach ihrem Wohlbefinden fragt. Es entstand eine weitere Diskussion, ob das überhaupt sinnvoll ist, wenn ein Gerät fragt, wie es einem gehe. Obwohl alle Probanden stets höflich mit dem System redeten, jeder „please“ verwendete und keiner im Befehlston etwas verlangte, war der Schritt, mit einem Roboter Smalltalk zu führen zu weit. Zum Teil mag dass daran liegen, dass das deutsche „Wie geht's dir?“ eine ehrlichere Antwort erwartet als das englische „how are you?“, das rein als Floskel gebraucht und verstanden wird. Bei den ersten Dialogen die

Probanden führten, waren viele von dieser Frage des Roboters irritiert und verunsichert. Nach mehreren Dialogen wurde diese Begrüßungsfloskel oft nur noch abgewartet oder mit einem kurzen „I'm fine“ übergangen, um möglichst schnell zu der eigentlichen Aufgabe zu kommen.

Alle Probanden wurden während den ersten Dialogen deutlich sicherer im Umgang mit dem System. Die Varianz in den Äußerungen wurde immer klein gehalten und wenn der Proband einmal wusste, wie ein Dialog ablaufen kann, wurde dieser Weg immer wieder gewählt. So kam kaum einer auf die Idee, einmal mit einer fehlerhaften oder unzureichenden Aussage das System zu fordern. Vielmehr wurden sogar meistens die Bezeichnungen gewählt, die das System selbst verwendet um auszusagen, was alles gefunden wurde. Der stattfindende Dialog wurde generell als sehr klar strukturiert und deutlich bewertet. Vor allem die Wiederholung des Erkannten Objektes und dessen was der Benutzer gesagt hat, wurde als sehr positiv herausgestellt, weil man eine direkte Rückmeldung über Erfolg oder Misserfolg des letzten Dialogschrittes bekommt.

Bei Entscheidungsfragen des Systems haben alle Probanden im ersten Moment unsicher reagiert, wenn sie die Frage ablehnen mussten. Wenn das System den korrekten Namen vorgelesen hat, haben alle schnell und sicher das Wort angenommen, aber wenn es ein falscher, zum größten Teil sinnloser Name war, haben alle irritiert abgewartet und wussten nicht, was sie weiter tun sollten. Sogar nach mehrmaligen Fehlversuchen, also wenn die Benutzer bereits wussten, dass sie falsche Wörter ablehnen können, wurde die Entscheidung gegen das falsche Wort langsam und zögernd getroffen. In Diskussionen danach stellte sich heraus, dass die Wörter nicht richtig verstanden wurden und am Besten noch einmal wiederholt werden sollen.

Trotz des sehr rudimentären Aufbaus, der letzten Endes nur aus einer Stereokamera besteht, die auf einem Stativ angebracht ist, vermenschlichten manche Probanden den Roboter. Durch die sprachliche Kommunikation mit dem System entsteht direkt der Eindruck, dass der Roboter menschenähnlicher und intelligenter ist als ein Computer, weil eine freie Unterhaltung möglich ist, die normalerweise nur unter Menschen üblich ist. Dadurch dass der Roboter den Benutzer angeschaut hat, wenn er mit diesem redet und dass der Kopf gedreht wird, wenn er auf den Tisch oder auf ein Objekt darauf schaut entsteht der Eindruck von einem tatsächlichen Gegenüber, mit dem alle Spaß hatten zu interagieren.

Allerdings wurde bemängelt, dass der Proband nach einer Äußerung des Roboters eine kurze Zeit warten muss, bis er antwortet, da der Roboter das Gesprochene sonst ignoriert. Diese Pausen wirken sich so negativ auf den Dialog aus, dass er unnatürlich auf die Probanden wirkte. Auch dass die Kommunikation nur durch einzelne Sätze geführt wird und weitere Informationen nicht weiterverarbeitet werden störte bei der freien Kommunikation sehr. Eine sehr gute Segmentierung der Sprache ist also wichtig für eine flüssige Unterhaltung, die dadurch noch natürlicher und freier wirkt.

7. Zusammenfassung und Ausblick

In dieser Studienarbeit wird ein System vorgestellt, das es ermöglicht, neue Objekte, die in der Umgebung eines Roboters auftauchen, zu lernen und später wieder zu finden. Es wird ein Objektmodell vorgestellt, das eine Klassifizierung der Objekte in einer Ontologie erlaubt. Im Verlauf eines Dialogs mit dem Benutzer sammelt das System Informationen über das neue Objekt oder das Objekt, das der Benutzer wünscht und reagiert so darauf, dass das Dialogziel erreicht werden kann.

Als erstes werden die Grundlagen der natürlich sprachlichen Dialogführung mit einem automatischen System erklärt. Dabei werden Spracherkennung, Sprachverstehen und der Ablauf des Dialogsystems erläutert. Dann wird eine Möglichkeit der Objekterkennung vorgestellt, die in der Lage ist, ein bekanntes Objekt in einer natürlichen und zufälligen Anordnung auf einem Tisch wieder zu erkennen.

Die vorhergehenden Arbeiten werden vorgestellt: Die Grundlagen des Detektierens und Lernens neuer Wörter und ein Basisszenario, in dem der Benutzer mit dem System interagiert um ein Objekt gereicht zu bekommen.

Im weiteren Verlauf wird der Entwurf des Gesamtsystems dargestellt. Durch Kombination der einzelnen Prototypen und Anpassung der verwendeten Modelle wird ein System entwickelt, in dem zusammen mit der entwickelten Objektmodellierung und dem Dialogablauf das Erkennen und Lernen neuer Objekte möglich ist. Das Hauptaugenmerk wird dabei auf die Modellierung der Objekte und den Dialogablauf gelegt. Die Modelle umfassen Merkmale, die die Objekte visuell beschreiben, damit der Objekterkenner diese wiederfinden kann. Eine Ontologie wird erläutert, in der die Objekte beschrieben werden und mögliche Objektbezeichner mit dem tatsächlichen Objekt verknüpft werden. Dieser Ansatz wird in einem Prototyp umgesetzt.

Bei der Evaluation mit Hilfe des Prototyps wird herausgestellt, dass es möglich ist, online neue Objekte zu lernen und somit dauerhaft zu modellieren. In einer Reihe von Tests mit verschiedenen Benutzern wird die Einsatzmöglichkeit erfolgreich gezeigt.

Um die Arbeit weiter zu entwickeln, sind folgende Ansatzpunkte denkbar:

- Der Dialogablauf kann, wie in anderen Projekten am ISL bereits üblich, von einem Reinforcement Learning Algorithmus automatisch gelernt werden. Dadurch kann die Strategie leicht an andere Umgebungen und Aufgaben angepasst werden.
- Die verwendete Ontologie kann von der bisher verwendeten Taxonomie, einer rein hierarchischen Ordnung in Klassen und Unterklassen weiter entwickelt werden zu einem semantischen Web, indem die Relationen freier und komplexer sein können.
- Die in Kapitel 4.4 beschriebenen Fehler sollten noch an geeigneten Stellen abgefangen werden, damit der Dialog robuster wird und besser auf falsche Vermutungen über die Situation reagieren kann.

Während der Arbeit an dieser Studienarbeit sind dem Autor keine ähnlichen Systeme bekannt, die mit Hilfe eines Dialogs Objekte lernen und wiederfinden können. Um ein solches System zu erschaffen, ist es nötig, dass fundierte Kenntnisse in der Spracherkennung, Dialogführung und Objekterkennung vorhanden sind. Am Interactive Systems Labs (ISL) werden diese verschiedenen Forschungsrichtungen erfolgreich in ein Gesamtsystem integriert, die das erlaubt. Die Einzelkomponenten müssen sehr robust und zuverlässig arbeiten, da der Ausfall eines Teiles den ganzen Dialog scheitern lässt.

8. Literaturverzeichnis

Asfour, T., K.Regenstein, Azad, P., Schröder, J., Bierbaum, A., Vahrenkamp, N., et al. (2006). ARMAR III: An integrated humanoid platform for sensory-motor control. *IEEE Institute of electrical and electronics engineers*. Genveva, Italien: IEEE Service Center, Piscataway, NJ, USA.

Azad, P. (2007). *Integration Vision Toolkit*. Von <http://ivt.sourceforge.net>. abgerufen

Azad, P., Asfour, T., & Dillmann, R. (2006). Combining Appearance based and Model-based Methods for Real-Time Object Recognition and 6D-Localization. *International Conference on Intelligent Robots and Systems (IROS)*.

Carpenter, B. (1992). *The Logic of Typed Feature Structures*. Cambridge, England: Cambridge University Press.

Deutsche Forschungsgemeinschaft. (Juni 2001). *SFB 588 Humanoide Roboter*. Von <http://www.sfb588.uni-karlsruhe.de/>. abgerufen

Finke, M. (1997). Recognition Of Conversational Telephone Speech Using The Janus Speech Engine. *ICASSP '97*.

Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. In *Knowledge Acquisition 5(2)* (S. 199-220).

Holzapfel, H. (2005). Towards Development of Multilingual Spoken Dialog Systems. *2nd Language and Technology Conference (L&T'05)*.

Holzapfel, H., & Prommer, T. (Mai 2006). Rapid Simulation-Driven Reinforcement Learning of Multimodal Dialog Strategies for Human-Robot Interaction. Technical Report. (T. Prommer, Hrsg.) Universität Karlsruhe (TH) / Carnegie Mellon University.

Levin, L., Lavie, A., Woszczyna, M., Gates, D., Gavalda, M., Koll, D., et al. (2000). The JANUS-III Translation System: Speech-to-Speech Translation in Multiple Domains. In *Machine Translation*.

Lömker, F. (30. 07 2004). Lernen von Objektbenennungen mit visuellen Prozessen. Bielefeld: Universität Bielefeld.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60, 2 , 91-110.

Ogden, C. K., & Richards, I. A. (1923). *The meaning of meaning. A study of the influence of language upon thought and of the science of symbolism. With an introduction by J.P. Postgate and supplementary essays by B. Malinowski and F. G. Crookshank* . London: Kegan Paul, Trench, Trubner & Co.

Point Grey. (14. February 2005). Technical Specification, Bumblebee Digital Stereo Vision Camera.

Schaaf, T. (2004). *Erkennen und Lernen neuer Wörter*. Universität Karlsruhe.

Schulz, B. (2005). *Lernen neuer Wörter im Dialog*.

Soltau, H. (2001). *A One-Pass Decoder Based on Polymorphic Linguistic Context Assignment*.

Talamazzini, E.-G. S. (2001). *Automatische Spracherkennung. Grundlagen, statistische Modelle und effiziente Algorithmen*. Vieweg.

Waibel, A. (1990). *Readings in Speech Recognition*. Morgan Kaufmann.