

EMERGENCE OF VERB AND OBJECT CONCEPTS THROUGH LEARNING
AFFORDANCES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

NİLGÜN DAĞ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2010

Approval of the thesis:

**EMERGENCE OF VERB AND OBJECT CONCEPTS THROUGH
LEARNING AFFORDANCES**

submitted by **NILGÜN DAĞ** in partial fulfillment of the requirements for the degree
of **Master of Science in Computer Engineering Department, Middle East
Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Adnan Yazıcı _____
Head of Department, **Computer Engineering**

Asst. Prof. Dr. Sinan Kalkan _____
Supervisor, **Computer Engineering Department, METU**

Asst. Prof. Dr. Erol Şahin _____
Co-supervisor, **Computer Engineering Department, METU**

Examining Committee Members:

Prof. Dr. Fatoş Yarman Vural _____
Computer Engineering, METU

Asst. Prof. Dr. Sinan Kalkan _____
Computer Engineering, METU

Asst. Prof. Dr. Erol Şahin _____
Computer Engineering, METU

Prof. Dr. Göktürk Üçoluk _____
Computer Engineering, METU

Asst. Prof. Dr. Didem Gökçay _____
Informatics Institute, METU

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: NİLGÜN DAĞ

Signature :

ABSTRACT

EMERGENCE OF VERB AND OBJECT CONCEPTS THROUGH LEARNING AFFORDANCES

Dağ, Nilgün

M.Sc., Department of Computer Engineering

Supervisor : Asst. Prof. Dr. Sinan Kalkan

Co-Supervisor : Asst. Prof. Dr. Erol Şahin

September 2010, 56 pages

Researchers are still far from thoroughly understanding and building accurate computational models of the mechanisms in human mind that give rise to cognitive processes such as emergence of concepts and language acquisition. As a new attempt to give an insight into this issue, in this thesis, we are concerned about developing a computational model that leads to the emergence of concepts. Specifically, we investigate how a robot can acquire verb and object concepts through learning affordances, a notion first proposed by J. J. Gibson in 1986. Using the affordance formalization framework of Şahin et al. in 2007, a humanoid robot acquires concepts through interactions in an embodied environment.

For the acquisition of verb concepts, we take an alternative approach to the literature, which generally links verbs to specific behaviors of the robot, by linking them to specific effects that different behaviors may generate. We show how our robot can learn effect prototypes, represented in terms of feature changes in the perception vector of the robot, through demonstrations made by a human supervisor. As for the object

concepts, we use the affordance relations of objects to create object concepts based on their functional relevance. Additionally, we show that the extracted effect prototypes corresponding to verb concepts can also be utilized to discover stable and variable properties of objects which can be associated to stable and variable affordances.

Moreover, we show that the acquired concepts provide a suitable basis for communication with humans or other agents, for example to understand and imitate others' behaviors or for goal specification tasks. These capabilities are demonstrated in simple interaction games on the iCub humanoid robot platform.

Keywords: concepts, object categorization, affordances, language embodiment and grounding, language acquisition

ÖZ

NESNE VE FİİL KAVRAMLARININ SAĞLARLIKLARI ÖĞRENME YOLUYLA ORTAYA ÇIKMASI

Dağ, Nilgün

Yüksek Lisans, Bilgisayar Mühendisliği Bölümü

Tez Yöneticisi : Yrd. Doç. Dr. Sinan Kalkan

Ortak Tez Yöneticisi : Yrd. Doç. Dr. Erol Şahin

Eylül 2010, 56 sayfa

Bilim adamları insan zihnindeki kavramların oluşması veya dil edinimi gibi bilişsel süreçlere yol açan mekanizmaları iyice anlayabilmekten ve bunlara uyan modeller yapabilmekten hala çok uzaklar. Bu konuda yeni bir çalışma olarak, bu tezde, kavramların oluşmasını inceleyen hesaplamalı bir model geliştirmekle ilgileniyoruz. Spesifik olarak, bir robotun fiil ve nesne kavramlarını 1989'da J. J. Gibson tarafından öne sürülen bir kavram olan sağlıkları öğrenme yoluyla nasıl edinebileceğini araştırıyoruz. Sahin ve arkadaşları tarafından 2007'de öne sürülen sağlık formalizasyon sistemini kullanarak, insansı bir robot çevresiyle etkileşime geçerek kavramlar oluşturuyor. Fiil kavramlarını oluştururken, filleri robot hareketlerinin doğurabileceği farklı etkilerle ilişkilendirerek, filleri robotun farklı hareketlerine bağlayan genel literatürden farklı bir yol izliyoruz. Robotumuzun etki prototiplerini, algı vektöründeki niteliklerin değişimleri cinsinden, bir insan yönetiminde nasıl öğrenebileceğini gösteriyoruz. Nesne kavramları için ise, nesnelerin fonksiyonel benzerliklerine dayalı kavramlar oluşturmak için objelerin sağlık bağıntılarını kullanıyoruz. Bunlara ek olarak, fiil kavramlarına karşılık gelen prototiplerimizin objelerin değişen ve sabit kalan özelliklerini keşfetmeye

de yarayabileceğini gösteriyoruz. Son olarak, oluşturulan kavramların başkalarının hareketlerini anlama ve taklit etme veya hedef belirleme görevleri gibi işlerde, insanlar ve diğer ajanlar arasındaki iletişim için uygun bir temel sağlayabileceğini gösteriyoruz. Bu yetenekleri iCub insansı robot platformunda basit etkileşim oyunları ile gösteriyoruz.

Anahtar Kelimeler: kavramlar, nesnelerin sınıflandırılması, sağlıklar, dili cisimleştirme (*ing.* embodiment) ve temellendirme (*ing.* grounding), dil edinimi

To my family

ACKNOWLEDGMENTS

I would like to thank my supervisors Sinan Kalkan and Erol Şahin for their extensive contribution in this thesis. Sinan Kalkan was a wonderful supervisor, always available in order for me to turn ‘the hard’ to ‘the easiest’ with his helpful comments, ideas and solutions. And beside all the supervision he has given me, he was a great friend, a thoughtful and a funny man. Erol Şahin put a lot into this thesis with his ideas and comments but more importantly he created our admirable atmosphere with robots, sensors and greatest friends. I also want to thank Fatoş Yarman Vural for her being a wonderful teacher; I owe my knowledge of pattern recognition and image processing to her.

My dear friends, Hande Çelikkanat, Ömer Nebil Yaveroğlu, Selma Süloğlu and Burçin Sapaz also deserve a big thank you. I can’t find words to express how thankful I am to Hande for being the one to talk to whenever I need and making me feel all better. Ömer, my friend long ago, has bring the joy into every single day. He was also a great partner as a teaching assistant by undertaking all the hard work. Selma was always full of fun, worthy of all the long times she made us wait for her. And Burcin, my rival of all times, was always helpful and supportive. Thank you all.

Barış Akgün and Fatih Gökçe, with their secret meetings at nights, İlkay Atıl, with his funny stories, Emre Uğur, with his great hospitality, Doruk Tunaoglu, with his valuable partnership, Kadir Fırat Uyanık, with stormy squash matches and Güven İşcan, with the nice conversations we had together, made Kovan Lab. a beautiful place for me to spend all these days. Also, despite the little time we spent together, it was very nice to know our new members Onur Yürüten, Mustafa Parlaktuna, Yiğit Çalışkan, Çiğdem Avcı and Asil Kaan Bozcuoğlu. I wish good luck to them. And finally, I would like to thank our summer interns İsmail Can Coşkuner, Meryem Sağcan, Gonca Gül Yıldırım and Üstün Yıldırım for bringing a color into the hardest days writing this thesis.

As a last word, I want to thank Kübra Kültür for being the perfect housemate, supporter at the hardest times and enduring all the hardships of me.

This work is partially funded by the European Commission through projects ROSSI (FP7-ICT-216125) and RobotCub (FP6-ICT-004370), and by TÜBİTAK (Turkish Scientific and Technical Council) through project no 109E033.

TABLE OF CONTENTS

ABSTRACT	iv
ÖZ	vi
ACKNOWLEDGMENTS	ix
TABLE OF CONTENTS	xi
LIST OF TABLES	xiii
LIST OF FIGURES	xiv
CHAPTERS	
1 INTRODUCTION	1
1.1 Contributions of the Thesis	4
1.2 Outline of the Thesis	5
2 BACKGROUND AND LITERATURE SURVEY	6
2.1 Concept Formation, Perception, Action, Language	6
2.2 Affordances	10
2.2.1 Stable and Variable Affordances	13
2.3 Literature on Concepts	14
2.3.1 Verb Concepts	14
2.3.2 Object Concepts	15
2.4 Feature Extraction Methods	17
2.4.1 Curvature Based Feature Extraction Methods	17
2.4.1.1 Normal and Principle Curvatures	17
2.4.1.2 Mean and Gaussian Curvatures	18
2.4.1.3 Surface Type	18
2.4.1.4 Shape Index	18
2.4.1.5 Degree of Curvedness	19

3	EXPERIMENTAL SETUP AND METHODS	20
3.1	The iCub Humanoid Platform	20
3.2	SwissRanger SR4000 Camera	20
3.3	Data and Features	21
3.4	Affordance Learning Model	25
3.5	Effect Prototype Extraction	26
3.6	Object Categorization	27
4	ACQUIRED CONCEPTS	30
4.1	Verb Concepts	30
4.1.1	Interaction Games	32
4.2	Object Concepts	33
4.2.1	Stable and Variable Affordances	34
5	DISCUSSION	41
	REFERENCES	43
	APPENDICES	
A	Feature Extraction Methods	48
A.1	Surface normal based methods	48
A.2	Principal Component Analysis (PCA) based methods	49
A.3	Image descriptor based methods	49
A.3.1	Scale Invariant Feature Transform (SIFT)	49
A.3.2	2.5 SIFT	50
A.4	Image Moments	50
A.5	Methods inspired from texture representation	50
B	Support Vector Machine (SVM) Classifier	53
C	Relief and ReliefF Algorithms for Feature Selection	55

LIST OF TABLES

TABLES

Table 2.1	Eight surface types, based on the signs of H and K curvatures [1]. . .	19
Table 4.1	Effect prototype strings that are extracted using the method introduced in section 3.5. Note that each feature element has an associated mean and variance of the change. θ stands for orientation.	31
Table 4.2	The detected categories of novel and non-symmetrical objects. . . .	35

LIST OF FIGURES

FIGURES

Figure 1.1	(a) Different behaviors through which one can push an object. (b) Chairs that are perceptually different yet functionally equivalent ¹	2
Figure 2.1	Principle curvatures ²	18
Figure 2.2	Surfaces represented by different values of the shape index.	19
Figure 3.1	The iCub humanoid robot platform.	21
Figure 3.2	SwissRanger SR4000 time-of-flight range camera.	21
Figure 3.3	On the left, the objects used in the experiments are shown. On the right, range images corresponding to different objects are given. A portable handle is used to assign orientation to cylinders and boxes.	22
Figure 3.4	Shape index histograms corresponding to objects with different shapes.	23
Figure 3.5	Sub-figures (a) and (b) show the data points from the view point of the camera corresponding to a box and a sphere object, respectively. We transform these points to obtain a top view image of the objects; (c) and (d). The sizes of the objects along eight different directions, shown with different colors in (e) and (f), are used as inputs to a learning method that predicts one specific orientation.	24
Figure 3.6	Labeled clusters in the effect space.	25
Figure 3.7	Clusters in the effect space are used for training an SVM, which allows predicting the effects of an behavior on a new object.	26

Figure 3.8 The distribution of effect features for different effect clusters. In (a), we see the effect corresponding to no change. As the sub-figure (b) shows, *rotated 90°* effect causes a consistent positive change on the fourth feature, namely the orientation of the objects, whereas other features are not affected significantly. Likewise, *pushed left* effect, sub-figure (c), can be mainly characterized by a positive change in the first feature (corresponding to x position) and *rolled forward* effect, sub-figure (d), by a negative change in the second and third features (corresponding to the y and z positions of the objects). 28

Figure 3.9 Grouping of feature elements into four clusters in the mean-variance space. This grouping is used for assigning ‘+’, ‘-’, ‘0’ and ‘*’, labels to feature elements. 29

Figure 4.1 A robot can utilize *verb concepts* in order to communicate with humans. We show that the robot can understand and imitate others’ behavior or accomplish a specified goal using *verb concepts* he has developed by simple interaction games³. 31

Figure 4.2 The distances between the extracted effect prototypes. We calculate distances by taking mean of one cluster and comparing it with the distribution of the other using equation 4.1. 37

Figure 4.3 (a) In the first line, iCub is demonstrated the *pushed right* effect, which it can successfully match with the corresponding *verb concept*, i.e., +0000000000000000. Then, it is introduced a new object on which it is required to create the same effect. The second line shows iCub executing the *push right* behavior which it successfully chooses among the behaviors in its repertoire. (b) Similarly, iCub is first demonstrated the *rolled left* effect (corresponding to a verb concept, i.e., -000000000*000000) and a new object is put in front of it. It chooses to execute the *push left* behavior to produce the same effect. 38

Figure 4.4 (a) iCub is given a goal task *--*****, meaning that the goal is to produce a decrease in the second and third dimensions and the change in the other dimensions can be ignored. iCub matches this goal with the <i>pushed forward</i> verb concept and chooses to execute the <i>push forward</i> behavior accomplishing the specified goal. (b) Similarly, iCub is given a goal +*****, meaning that iCub is to produce an increase in the first dimension of the perceptual representation of the object and the change in the other dimensions can be ignored. iCub matches this goal with the <i>pushed right</i> verb concept and executes <i>push right</i> behavior to accomplish the goal.	39
Figure 4.5 Categories of the objects, i.e., object concepts. We see that the interactions of iCub lead to three object concepts O.C.1, O.C.2 and O.C.3, which respectively correspond to balls, big objects (including medium and big cylinders and cubes) and small objects (both cylinders and cubes). . .	40
Figure A.1 Surface normals ⁴	48
Figure B.1 Support vector machines use nearest instances on each side, called support vectors, to find a maximum margin hyper-plane in the feature space ⁵ . 53	

CHAPTER 1

INTRODUCTION

Despite the big effort put by researchers from different disciplines (psychologists, neuroscientists, cognitivists, roboticists and philosophers), it is still a big challenge to fully understand the mechanisms that allow humans to form *concepts* and associate them with words and in general, language. In this thesis, we construct a computational model for investigating how a humanoid robot can form *verb* (i.e., the concepts corresponding to verbs in a language) and *object concepts* (i.e., the concepts corresponding to nouns in a language) by interacting with the objects in the environment.

From the point of psychology and neuroscience, concepts should be represented in sensorimotor space of the agents, i.e., they should be grounded in perception and action [2, 3, 4]. From the view point of robotics [5, 6, 7, 8, 9], the notion of affordances is a suitable basis to investigate the complex and dynamic nature of the real world. To put these together, we propose a framework in which a robot utilizes affordances to develop *verb* and *object concepts* through interactions in an embodied environment. Furthermore, we propose to associate *verb concepts* to specific effects generated by different behaviors, as opposed to the common literature in which verbs are linked to specific behaviors of the robot. Additionally, we show how *object concepts* can be linked to categories that are formed based on objects' functional similarities instead of solely relying on their appearances.

The term *concept* is defined by psychologists [2] as the information associated with its referent and what the referrer knows about it. For example, the concept of a car is all the information that we know about cars. This concept includes not only how a car looks like (i.e., that a car has a set of wheels, doors and windows with a certain

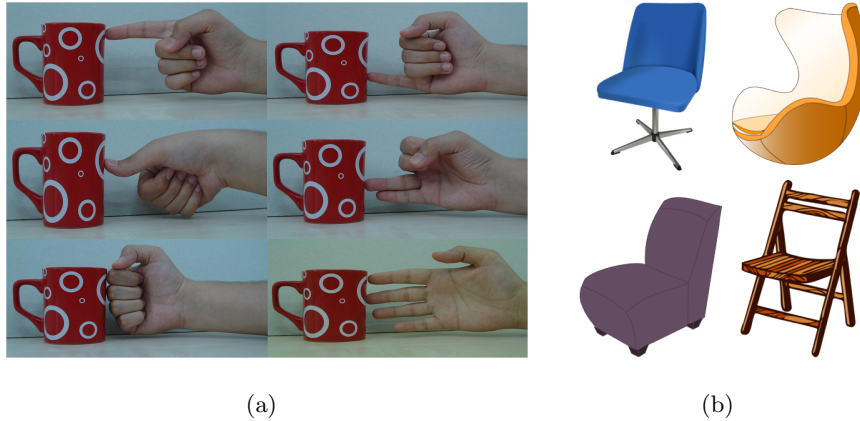


Figure 1.1: **(a)** Different behaviors through which one can push an object. **(b)** Chairs that are perceptually different yet functionally equivalent².

spatial configuration) but also how it sounds, how it feels when we touch its various parts and what we can do with a car and how we can do what we can do with a car.

Based on this definition, we can call concepts as abstractions of similar experiences; or, in more developmental terms, we can say that concepts are categories, or generalizations, in the sensorimotor space that is shaped by the experiences of the agent. Such abstractions not only (1) allow agents to make generalizations and transfer acquired abilities based on these generalizations, but also (2) give labels (or, names, words) to the abstractions to be able to communicate about them. Giving names or labels to abstractions, or concepts, means basically the emergence of language.

One objective of this thesis is the development of concepts represented by verbs and nouns in language in a robot to enable simple forms of communication with humans. In order to carry out a command such as “pick an apple”, the robot needs to understand what it means to “pick” an object, as well as what “an apple” is. We aim to develop a computational model for the sensorimotor grounding of concepts in robots and relate them to verbs and nouns in language to create a shared world model over which communication can occur on the iCub humanoid robot platform.

It is tempting to associate the concept of a verb with a category that covers all the interactions that are generated by the execution of a particular behavior. If we want the robot to lift a particular object, the verb lift can trigger the lift behavior in the robot to accomplish our goal.

²Images are taken with permission from <http://www.clker.com>. Last access date: 24.09.2010.

However, such an association provides a limited coverage for all the meanings that the verb lift should convey. First, the robot can probably lift an object with different behaviors, such as lift-with-right-arm and lift-with-left-arm on a humanoid robot (for example, Figure 1.1(a) shows six different behaviors that can be used by humans to push an object towards right). Second, the execution of the particular behavior may fail on some objects, such as the failure to lift a heavy object. Third, in certain cases, a seemingly contradictory behavior such as pressing, may also lift an object that is placed on a lever to accomplish lifting.

An alternative is to associate verbs with effect categories. Such an association imposes that the concept being conveyed by the verb is the request for a certain effect be generated through the use of an appropriate behavior. In this sense, when we ask the robot to lift an object, we specify the goal as the elevation of object position in the vertical axis and leave the choice of the particular behavior to the robot itself. This is referred to as *goal emulation* in the literature as a form of imitation characterized by the replication of the observed end effect [10], and is observed in infants after 12 months [11].

The issue of what a noun concept such as *strawberry* represents is also subject to debate. Studies in neuroscience and psychology suggest that objects are processed through two different pathways, one involving the Object Recognition (OR) system categorizing an object based on its visual (or distal) appearance, the other involving their functional properties, or affordances. The categorization of objects based on their visual appearances is a well-studied and hot topic in computer vision. However, such approaches often fail to capture the essence of a concept such as chair which may appear in very different forms (see Figure 1.1(b) for some examples) and are beyond the focus of our study.

We argue that an object can be defined by its affordances, i.e. all the affordance relations that it is part of. For instance, the concept of strawberry consists of all affordance relations that can be obtained from it such as: (1) eating (behavior) it would create a sweet sensation (effect) on the tongue, (2) power grasping (behavior) it would create a wet and squashed sensation on the hand (effect), etc. Such groupings can be acquired by trying out the robot's full behavioral repertoire on an object.

1.1 Contributions of the Thesis

This thesis has the following contributions:

- We use the notion of affordances to investigate how a robot can develop *object* and *verb concepts* through interactions in an embodied environment.
- We argue that a verb is linked not to a specific behavior of the robot, but to a specific effect that different behaviors may generate.
- We show how demonstrations made by a human (can be viewed as an example of social learning) allow the robot to learn an effect prototype represented in terms of changes in the perception vector of the robot.
- We investigate how the affordance relations of objects can be used to create *object concepts* based on objects' functional relevance.
- The acquired concepts are represented in the robots own sensorimotor space which means they are grounded in perception and action.
- We discuss that acquired concepts can lead the robot to discover stable and variable properties of objects which can be associated to stable and variable affordances.
- We demonstrate that acquired *verb concepts* can be used by the robot as a basis for understanding and imitating others' behaviors or for goal specification tasks through simple interaction games on iCub humanoid robot platform.

The work presented in this thesis has appeared in the following publications:

- Akgün B., Dağ N., Bilal T., Atıl I., Şahin E. (2009). Unsupervised Learning of Affordance Relations on a Humanoid Robot. International Symposium on Computer and Information Sciences.
- Dağ, N., Atıl, I., Kalkan, S., and Şahin, E. (2010). Learning affordances for categorizing objects and their properties. International Conference on Pattern Recognition.

- Atıl, I., Dağ, N., Kalkan, S., and Şahin, E. Affordances and Emergence of Concepts. Tenth International Conference on Epigenetic Robotics. Örenäs Slott, Sweden, November, 5-7, 2010
- Dağ, N., Atıl, I., Kalkan, S., and Şahin, E. Emergence of Object and Verb Concepts through Affordances. Cognitive Processing, Special Issue on “ Cognitive Robotics - Perception-Action-Interaction: Systems and Architectures ” (submitted)

1.2 Outline of the Thesis

In chapter 2, called “background and literature survey”, we first present the literature on formation of concepts and the relation between perception, action and language. Next, we examine the notion of affordances, affordance formalization and how this formalization can be used for deriving verb and object concepts. Then, recent work on formation of verb and object concepts is summarized. Finally, we discuss background on curvature based feature extraction methods from which we benefit to extract shape related information from objects in the experiments.

In chapter 3, called “experimental setup and methods”, we describe our experimental setup i.e., the iCub humanoid robot platform and SwissRanger SR4000 camera. Next, we introduce the objects used in the experiments and the features extracted from the range data. Finally, we discuss the methods used for deriving *verb* and *object concepts*, specifically for learning affordance relations, extracting effect prototypes and categorizing objects.

In chapter 4, called “acquired concepts”, we present the *verb* and *object concepts* acquired using methods introduced in chapter 3. We show that verb concepts can be used as a basis for understanding and imitating others’ behaviors or for goal specification tasks on the iCub robot platform. Next, object concepts acquired through categorization based on objects’ affordances are presented. As a final remark, we show our results can lead the robot to discover stable and variable affordances of objects.

In the final chapter, called “discussion”, we discuss the results of the thesis and describe several ways the work in this thesis can be improved.

CHAPTER 2

BACKGROUND AND LITERATURE SURVEY

In this chapter, we first present the literature on formation of concepts and the relation between perception, action and language. Next, we examine the notion of affordances, affordance formalization and how this formalization can be used for deriving verb and object concepts. Then, recent work on formation of verb and object concepts is summarized. Finally, we discuss background on curvature based feature extraction methods from which we benefit to extract shape related information from objects in the experiments.

2.1 Concept Formation, Perception, Action, Language

In 1950s, Alan Turing claimed [12] that if a human engaging in non-verbal communication with a computer behind a curtain cannot decide whether the responder behind the curtain is a computer or not, then we can claim that we have built a computer (program) that is artificially intelligent. This test, which is now known as the Turing test, has set the vision for much of the artificial intelligence research and was also subject to philosophical debates. In 1980s, J. Searle criticized the test through a thought experiment, known as the Chinese Room [13]. He argued that the ability to merely deceive a human being behind a curtain does not necessarily entail that the program is intelligent, or cognitive. Searle claimed that, as long as the program is not aware of the meanings of the symbols that it is manipulating, it is not reasonable to talk about intelligence.

Later, Harnad [14] argued that the gap between the symbols and their meanings cannot be closed by an external programmer and that this would be equivalent to "learning Chinese from the Chinese dictionary". Instead, he pointed out the *symbol grounding problem*, arguing that the symbols should be grounded in the *sensory projections* of the objects and the events in the environment. Harnad discusses three kinds of *symbolic representations* and their grounding as:

(1) "iconic representations" , which are analogs of the proximal sensory projections of distal objects and events, and (2) "categorical representations" , which are learned and innate feature-detectors that pick out the invariant features of object and event categories from their sensory projections. [...] Higher-order (3) "symbolic representations" , grounded in these [...] symbols, consist of symbol strings describing category membership relations (e.g., "An X is a Y that is Z").

The symbol grounding problem becomes more apparent for robots. Unlike computers, the robots have the means of physical interaction with the environment and are more likely to have similar sensory-motor experiences to humans. Nevertheless, the sensory-motor experiences that they experience are and will be different from ours, and the issue of how they can develop a shared set of symbols to represent the basic concepts of a language remains to be an open question.

The term *concept* is defined by the psychologists, as the information associated with its referent and what the referrer knows about it [2]. For example, the concept of a car is the information that we know about it. The issue of how concepts are acquired is still controversial among researchers.

As a solution to symbol grounding problem, an approach called *embodied cognition*, has been proposed by researchers [2, 3, 4]. This view argues that cognitive processes can not be abstract and should be deeply rooted in the sensorimotor experiences of an agent. These sensorimotor experiences are distributed over different modalities such as auditory, visual, and tactile information. According to this view, basic properties of cognition can be outlined as follows [2]:

- Cognition is embodied: Mental processes can not be thought independently of the hardware on which they are implemented; i.e., human brain and body [3].

Therefore, cognition is determined by the experiences of a body with particular physics and a sensorimotor system.

- Cognition is situated: Cognitive activity takes place in the context of a real world environment.
- Cognition is grounded in sensorimotor activities: Concepts develop along with perception and action. Observation unaccompanied by action is not sufficient for cognitive processes.
- Cognition is flexible and variable: Concepts should be dynamically represented in order to adapt to the present situation and context.
- Cognition automatically activates motor information: Seeing or hearing words about a particular concept automatically activates corresponding motor processes [2, 15].
- Cognition is for action [16]: The purpose of perceptual and cognitive processes is to guide for appropriate behaviors given context.

There are also neuroscientific evidences which support the *embodied* view of cognition. Rizzolatti et al. [17] discovered a group of neurons, called mirror neurons, which fire when a monkey performs an behavior as well as when it observes someone else executing the same behavior. This finding can be interpreted as that the monkey's ability to perform a behavior plays a role in his understanding of the meaning of the same behavior. This "behavior understanding" interpretation of mirror neurons (i.e. the brain internally reproduces/simulates the observed behaviors) goes hand in hand with the "object understanding" view of psychologists who suggest that the concept of an object should activate an online simulation of the past interactions with that object [2].

In the next paragraphs, we will first present studies which experimentally and theoretically examine the relationship between language, action and perception. Afterwards, a number of computational models for grounding language and meaning in the sensorimotor capabilities of robots will be summarized.

From a theoretical perspective, [4, 18] propose a system called Indexical Hypothesis (IH) which supports the view that sentence comprehension is achieved by a sensorimotor simulation of actions or situations referred by the sentences. Instead of the arbitrary, amodal and abstract symbols [19], perceived words are first mapped to *perceptual symbols* (first proposed by [20]) which are modal and not arbitrarily related to their referents. These symbols are then used to derive affordances. Since perceptual symbols are not arbitrarily related to their referents, new affordances can be derived from them. If the affordances derived from these symbols are mesh-able (can be smoothly integrated or combined as action plans) then sentence comprehension is successful.

Also, Glenberg et al.[4] experimentally validate the existence of an action-sentence compatibility effect (ACE) which relate sentence comprehension to bodily action. When participants were required to make an action towards the opposite of the direction referred in the sentences they are told, the response times were slower. This was explained by the inference of sentence comprehension mechanisms with the ones responsible for making movements to support the view that language is grounded in bodily actions.

Roy et al. [21] propose a computational model called CELL (Cross-channel Early Lexical Learning) which addresses the problem of speech segmentation (i.e. discovering words from fluent speech), jointly with the problem of associating words to co-acquired semantic categories. The model takes spoken utterances together with their corresponding visual contexts as input and returns the speech-to-category mappings by statistically modelling the consistent structure across sensory data. Whereas this work regards the agent as a passive observer, the work by Sagita et al. [22] also take into account the action capabilities of the agent. Their model explores how *compositionality of semantics*¹ can be acquired through interactions between the linguistic and behavioral processes.

From an evolutionary perspective, [23] investigates how language, a living, complex adaptive system, can originate, evolve and adapt among artificial agents. Within an

¹ Here, “compositionality” refers to the ability of humans to understand sentences from the meanings of their components and the way in which these components are put together. Compositional semantics can be used to derive meaning of unknown sentences using already sentences whose meanings are already known[22].

experimental setup called “Talking Heads”, they show that agents are able to originate language-like communication while playing grounded language games.

Along with sensorimotor grounding, i.e., the ability to link perceptual representations to symbols, *grounding transfer*, i.e., the ability to acquire new higher-order grounded symbols from grounded ones, is also a crucial mechanism for successfully grounding language in agents. Cangelosi et al. [24] propose a model that addresses both sensorimotor grounding and grounding transfer. Their hypothesis propose that the agents can acquire new sensorimotor capabilities by using basic words (already grounded in their sensorimotor system) to express new categories.

In [25], Steels et al. emphasize the role of social learning for concept formation. They designed three experiments in which SONY AIBO robot acquires visual data while interacting with a human in a complex real world environment. The human supplies words to the robot in order to pair words with the view of the objects. The experiments differ in the amount of information the mediator supplies to the robot. The results show the crucial role of social learning on category formation when results are evaluated in terms of their relevance to human categories.

[26] gives a review of different approaches used by various computational models for the problem of language acquisition. The models are discussed in terms of their approach to learning from the standpoint of generative, statistical, embodied and social cognition, developmental and the cultural evolution stances. In conclusion, they propose to form a synthesis of discussed approaches using the notion of learning biases.

2.2 Affordances

The notion of affordances, first introduced by J.J. Gibson [27], is a suitable framework for investigating the co-development of action, perception and language.

Gibson defined affordances of the environment as [27]:

... what it offers the animal, what it provides or furnishes, either for good or ill. The verb to afford is found in the dictionary, but the noun affordance is not. I have made it up. I mean by it something that refers to both the

environment and the animal in a way that no existing term does. It implies the complementarity of the animal and the environment.

For example, a bottle affords graspability affordance for a human but not for a dog. Likewise, a chair affords sittability affordances for a human and hideability affordance for a cat.

Affordances are defined to be dependent both on the agent, the environment and the specific context[27]:

... an affordance is neither an objective property nor a subjective property; or both if you like. An affordance cuts across the dichotomy of subjective-objective and helps us to understand its inadequacy. It is equally a fact of the environment and a fact of behavior. It is both physical and psychical, yet neither. An affordance points both ways, to the environment and to the observer.

Moreover, affordances are invoked based on the online visual information and do not involve object recognition. For example, we do not need to recognize and label a “chair” in order to sit on it, rather it is a set of features of the object such as shape, height or carrying capacity that we need to consider.

There are some formalizations of affordances [28, 29, 30, 31] but the one proposed by Şahin et al.[32] is more suitable to be used in robot control. Therefore, in our work, we use their formalization, where an affordance is represented as a relation between an entity (e), a behavior (b) and an effect (f):

$$a = (e, b, f). \tag{2.1}$$

In this relation, entity e is the raw perceptual appearance of the scene. It encapsulates the perceptual representation of an agent at different complexity levels, ranging from raw sensory data to the features extracted from the environment. However, within the context of this thesis, we confine the use of entity to a single object for simplicity. *Behavior* represents the physical embodiment of the agents interaction encoding the internal representation that defines a unit of action that can often take parameters for the initiation and online control. Finally, *effect* is defined as the change generated in the environment due to the execution of the behavior.

For instance, a robot applying its *lift-with-right-hand* behavior on a *blue-pen* to generate a *lifted* effect can be represented with a relation instance as:

$$(blue-pen, lift-with-right-hand, lifted), \quad (2.2)$$

where the terms *blue-pen*, *lift-with-right-hand*, and *lifted* are merely placeholders for the related perceptual and proprioceptive representations. However, a single relation instance provides little predictive ability over the future experiments, such as whether the application of the same behavior on a *red-pen* or a *blue-desk* will generate the same effect or not. Only after interacting with other objects, such as a *green-pen*, one can join the relation instances together as:

$$\left(\begin{array}{c} blue-pen \\ green-pen \end{array} \right), lift, lifted). \quad (2.3)$$

This relation can then be compacted by a mechanism that operates on the class to produce the (perceptual) invariants of the entity equivalence class as:

$$(<*-pen>, lift, lifted), \quad (2.4)$$

where $<*-pen>$ denotes the derived invariants of the entity equivalence class. In this particular example, $<*-pen>$ means “pens of any color” that can be *lifted* upon the application of *lift* behavior. Such invariants create a general affordance relationship, enabling the robot to predict the *effect* of the *lift* behavior applied on a novel object, like a *red-pen*. Such a capability offers great flexibility to a robot. When needed, the robot can search and find objects that would provide support for a desired affordance. We argue that the creation of equivalence classes over the three components of the relation provides a mechanism for creating abstract *categories* that can be linked to concepts represented by verbs and nouns.

Recently, the concept of affordances has been used in [5, 6, 7, 8, 9], which develop systems to control agents which are capable of predicting outcomes of their behaviors, making plans to achieve a specific goal, recognizing others’ behaviors, interacting with other agents and learning by imitation. In this study, we also make use of affordances but for a different purpose, namely to investigate the emergence of *verb* and *object concepts*.

2.2.1 Stable and Variable Affordances

It is suggested that two neural pathways are involved in processing the information from the visual cortex [33]. The dorsal stream, commonly referred to as the “where/how” stream, is involved in guiding behaviors to spatial locations by communicating with regions that control eye and hand movements. The ventral stream, commonly referred as the “what” stream, is involved in recognition, identification and categorization of visual stimuli.

It is claimed that these two streams process two kinds of affordances of the objects [34]; i.e., the *stable* and *variable affordances*.

Here, variable affordances are defined as the affordances that emerge from temporary characteristics of objects, such as the current handle orientation or position of an object. These affordances depend on the context and therefore, proposed to be integrated to the cognitive processes online, when a behavior is intended.

On the other hand, stable affordances rather emerge from the invariant characteristics of objects, for example standard size or shape. They are assumed to encode information on the most frequent interactions between the human and the object, such as canonical orientation, the grasp type associated with the object or other functional properties. Therefore, it is likely that stable affordances may incorporate into object representation.

Given properties of stable and variable affordances, it is important for an agent to extract stable and variable properties of objects itself. The prototypes used for representing *verb concepts* in our study, can be utilized by the robot to determine what aspects of the object can a certain behaviour change. When such analysis is done for all the behaviours in the repertoire of the robot, then the union of these representations would allow the robot to figure out which aspects of a given object can be changed, and which aspects can't. We claim that the aspects that can be changed can be associated with the variable affordances, whereas the rest can be associated with the stable affordances.

2.3 Literature on Concepts

Humans' ability to categorize objects and events in the environment is a crucial property for creating an understanding of the world and communicating with each other. If we were to treat every one of our sensorimotor experiences uniquely, an infinite number of words would be needed to label them and communication would be impossible. Therefore, we need to develop mechanisms that can group similar experiences together. In the next two sections, we describe our view of what concepts represented by verbs and nouns in language should refer to and summarize recent work on formation of these concepts.

2.3.1 Verb Concepts

Although the literature mostly identifies verbs as words that define behaviors directly performed by the agent (e.g., [22, 24]), this view sometimes can be insufficient as one will have to associate a different verb for each push behavior in figure 1.1.

Alternatively, we take the approach of associating verbs to each effect that can be achieved on the entities. This association can be realized by generalizing over behaviors first to find the behaviors leading to similar effects and then linking verbs with these generalizations. We represent each verb with a prototype that reflects characteristics of the feature changes in the perception vector of the robot.

Similarly, [35] proposes a system to store behaviors as a representation of their effects and effects as a representation of feature changes. They designed an experiment in which a robot threw balls against a pyramid of ten cans. A Bayesian Network is trained to learn the correlation between the hit point and number of cans thrown off. The behavior, parameterized by the hit point coordinates in this context, is represented by effects i.e. number of thrown off cans. This work is parallel to our study in that they also represent effects as feature changes but generalization over behaviors or effects is not pursued for deriving *verb concepts*.

Kozima et al. [6] propose a model to examine the emergence of mirror system in humans. They argue that the role of mirror system is to perform effect-based imitation.

Their work also generalize over behaviors based on their effects but is only aimed for imitation purposes.

In [36, 37], Cohen et al. propose a maps-for-verbs framework to describe the semantics of action words, i.e., verbs. First, they experimentally prove that semantics of the words that humans choose to utter relies on the dynamical movement of objects in the situation. In their experiments, children were required to describe some movies. When the parameters that were used to generate dynamical movements in the movies and distributions of words that were mainly used by children to describe movies was compared, a strong dependence was observed. Using this finding, they choose to represent verbs based on the dynamics of before, during and after phases of interaction between whole bodies. Maps were constructed as a trajectory using relative velocity versus distance and energy transfer versus time characteristics of the interactions through the mentioned phases. These maps constitute for verbs. This study diverge from ours in that they emphasize the dynamic aspects of the motion generated by actions and associate verbs with action words rather than effects generated by them.

Similar to the view of [37], Marocco et al. [38] propose a model that associates action words with the dynamical properties of interactions with the objects. They made experiments with the iCub robot simulator in which the robot is controlled by a recurrent artificial neural network that runs Back-Propagation-Through-Time algorithm. When the neural network is trained by parameters from interactions with different objects, the neural network generates categories that maps to *the rolling one* instead of a *sphere*, *the sliding one* rather than a *cube* and *the fixed one* rather than a *cylinder*. This shows that words learned by the robot relies on the dynamical aspects of the interaction with the objects rather than appearance of interacted objects.

2.3.2 Object Concepts

One can take *object concepts* (i.e., concepts represented by nouns) as (1) perceptually different sets of features (i.e., appearance-based categorization) or (2) functionally different sets of features (i.e., function-based categorization) based on what *nouns* and *adjectives* refer to. It is known that adults utilize both kinds of mechanisms for categorizing objects [39]; yet, for the current study, we are interested in approach (2)

since it is developmentally more relevant.

In [39], Borghi et al. analyzed the role of perceptual and functional similarity of objects in category formation. Neural networks were used to group the stimuli by pressing a button in one of two categories which may have been formed by perceptually very similar, moderately similar or different objects depending on the task. In their setup, they used a 5 layer neural network to simulate the nervous system of an artificial organism. The input consisted of three main components: visual input units connected to the first of the three hidden layers, task demand units connected to the second hidden layer and the proprioceptive inputs connected directly to the output. The output layer encoded the actions performed by the organism. Their results can be summarized as follows:

- In the hidden layers closer to the sensory input, which do not have specific information on the action to perform, perceptual properties of objects are used as a cue for forming categories.
- In the upper hidden layers of the network, where action information incorporates, category formation is based on functional properties.
- If the task demand requires the organism to put together perceptually dissimilar objects, task information overrides perceptual similarity.
- Nonetheless, if there is a congruence between perceptual similarity and the task to perform, i.e. the objects in the same category based on the task are perceptually similar, the categorization task is facilitated.
- The results are consistent with the distinction between primary (based on perceptual similarity) and secondary categorisation (based on functional requirements) first proposed by Barsalou [40].

Nolfi et al. [41] used tactile information retrieved from touch sensors of a 3-segment arm for categorizing objects. They used evolutionary methods to classify sphere or cubic objects with a fitness function defined as the sum of number of phases in which individuals correctly classify the current object. The arm of successfully evolved agents that were able to follow the surface of the objects to decide whether it is curvilinear

or not. The results show that classification of objects does not rely on observable features of the environment but rather emerges from the overall interaction between the control system of the agent and the environment.

Sun et al. [42] used object categories to improve the scalability of affordance learning. They proposed a new probabilistic and graphical affordance learning model called Category-Affordance (CA) model. As opposed to *direct perception* phenomena, this approach first organizes visual inputs into categories and defines relationships between affordances and these categories. Then the categorization of a new input determines its affordances.

2.4 Feature Extraction Methods

In this section, we give a background on the curvature based feature extraction methods which are used in our experiments to extract shape related information from range images (see section 3.3). We also reviewed other methods based on surface normals, principle component analyses, point sets and texture representations. A brief summary of those methods can be found in appendix A.

2.4.1 Curvature Based Feature Extraction Methods

The term *curvature* is used as a measurement of the deviation of a geometric object from being flat². In this subsection we first describe the terms *normal* and *principle curvatures*, which are used to calculate *mean* and *gaussian curvatures* of a surface. Then these curvature values are used to define three measures of curvedness; namely *surface type*, *shape index* and *degree of curvedness*.

2.4.1.1 Normal and Principle Curvatures

The *normal curvature* of a point on a surface is defined as the curvature of any curve that is at the intersection of the surface and other planes which contain the normal vector at the given point. The *principle curvatures* (k_1 and k_2), the minimum and

² <http://en.wikipedia.org/wiki/Curvature>

maximum valued normal curvatures at that point, are useful indicators of shape as they can be interpreted as the measure of bending in different directions at the given point. (see figure 2.1)

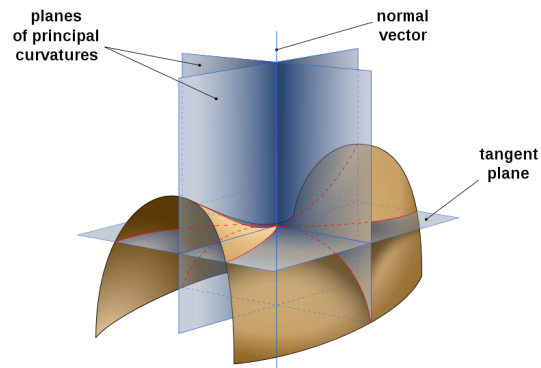


Figure 2.1: Principle curvatures³.

2.4.1.2 Mean and Gaussian Curvatures

Mean(H) and Gaussian(K) curvatures are calculated by geometric mean and cross product of principle curvatures, respectively.

$$H = \frac{k_1 + k_2}{2}, \quad (2.5)$$

$$K = k_1 \times k_2. \quad (2.6)$$

2.4.1.3 Surface Type

Based on the signs of Mean and Gaussian curvatures, [1] classifies surfaces into eight discrete categories as shown in table 2.1.

2.4.1.4 Shape Index

The shape index, is a continuous measure of shape information defined in the range $[-1; 1]$ and calculated using principle curvatures with the formula given in equation

³Image is taken with permission from http://commons.wikimedia.org/wiki/File:Minimal_surface_curvature_planes-en.svg. Last access date: 22.09.2010.

Table 2.1: Eight surface types, based on the signs of H and K curvatures [1].

	$K(i, j) > 0$	$K(i, j) = 0$	$K(i, j) < 0$
$H(i, j) < 0$	Peak	Ridge	Saddle Ridge
$H(i, j) = 0$	Undefined	Flat	Minimal Surface
$H(i, j) > 0$	Pit	Valley	Saddle Valley

2.7. Different shapes represented by different values of shape index can be seen in figure 2.2.

$$S = \frac{2}{\pi} \tan^{-1} \left(\frac{k_2 + k_1}{k_2 - k_1} \right). \quad (2.7)$$

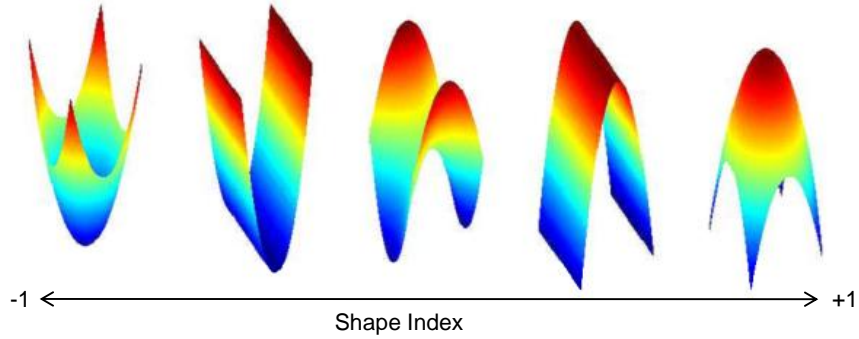


Figure 2.2: Surfaces represented by different values of the shape index.

2.4.1.5 Degree of Curvedness

Another measure of shape of a surface is the *degree of curvedness* and is given by equation 2.8.

$$R = \sqrt{\frac{k_1^2 + k_2^2}{2}}. \quad (2.8)$$

CHAPTER 3

EXPERIMENTAL SETUP AND METHODS

In this chapter, we describe our experimental setup i.e., the iCub humanoid robot platform and SwissRanger SR4000 camera. Next, we introduce the objects used in the experiments and the features extracted from the range data. Finally, we discuss the methods used for deriving *verb* and *object concepts*, specifically for learning affordance relations, extracting effect prototypes and categorizing objects.

3.1 The iCub Humanoid Platform

The iCub [43] is a fully open source humanoid robot designed for cognitive and developmental robotics research (see figure 3.1). It has the dimensions of a 3.5 years old child and has capability of actuating 53 degrees of freedom on its hands, arms, torso and legs.

Simple behaviors such as *push-left*, *push-right* and *push-forward* have been implemented on the iCub platform and used for testing the learned *verb concepts*.

3.2 SwissRanger SR4000 Camera

SwissRanger SR4000 (shown in figure 3.2) is a time-of-flight range camera capable of capturing depth of scenes with a resolution of 176×144 at 30fps. It provides three kinds of information (as three 176×144 images): the range data, the amplitude of the signal and the confidence of the signal. This information is used for segmenting background and extracting features from the object of interest in the scene.

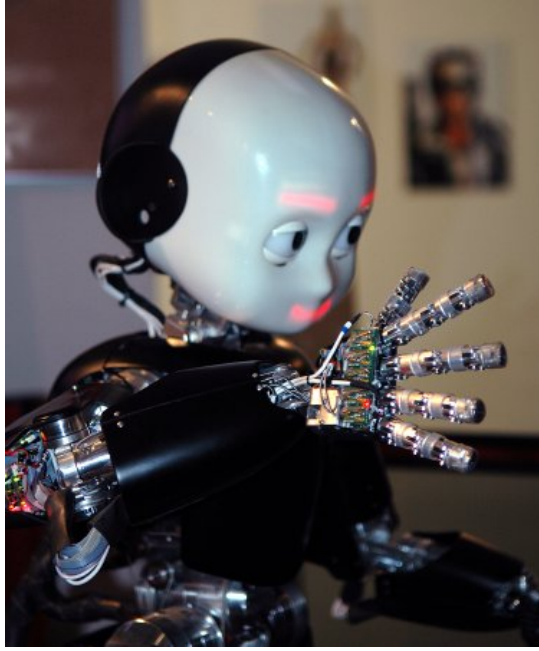


Figure 3.1: The iCub humanoid robot platform.



Figure 3.2: SwissRanger SR4000 time-of-flight range camera.

3.3 Data and Features

We acquired the range data using SwissRanger 4000 camera. Our setup had a low-amplitude background which allowed us to make a clean segmentation of the objects.

In the experiments, objects of different sizes (big, medium, small) and of different

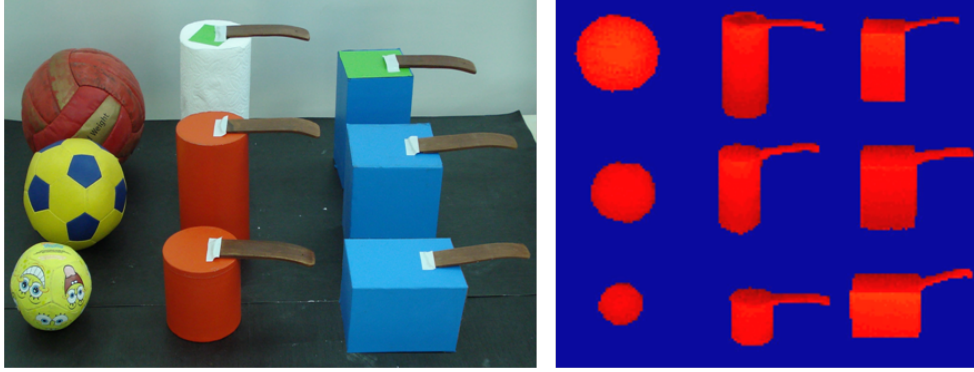


Figure 3.3: On the left, the objects used in the experiments are shown. On the right, range images corresponding to different objects are given. A portable handle is used to assign orientation to cylinders and boxes.

shapes (ball, cylinder, box) are used (shown in figure 3.3). We perform simple behaviors, namely *push-right*, *push-left*, *push-forward*, *rotate 45°*, *rotate 90°* and *lift* on these objects.

The behaviors are performed by a human similar to the ones applied by our robot. Note that these interactions could have been carried out using the iCub as well. We have preferred to have a human in place, due to two practical reasons. First, the development of interaction behaviors on the robot has to be done manually and was time consuming. Second, we wanted to minimize the wearing out of the robot. Moreover, for interactions that are as primitive as the ones used in this study, the use of the robot would not have changed our results. We used the robot only during the testing and the demonstration phase using only the three behaviors mentioned in the previous section.

From the segmented range data, we extract the following shape, size, position and orientation related features (17 elements in total):

- 3D position of the object (3 values).
- Shape of the object (10 values). Shape indexes, as described in section 2.4.1.4, are calculated at each data point excluding edge pixels. Then, a histogram of these values is created using linear interpolation. The number of bins of this histogram, 10, is experimentally determined based on the histogram's capability to represent shapes of objects used in the experiments. Figure 3.4 shows histograms

corresponding to objects with different shapes, namely a box, a cylinder and a sphere. We note that, in addition to their capability to discriminate between different shapes, shape indexes also decouple shape information from orientation, unlike histograms of surface normals (see section A.1).

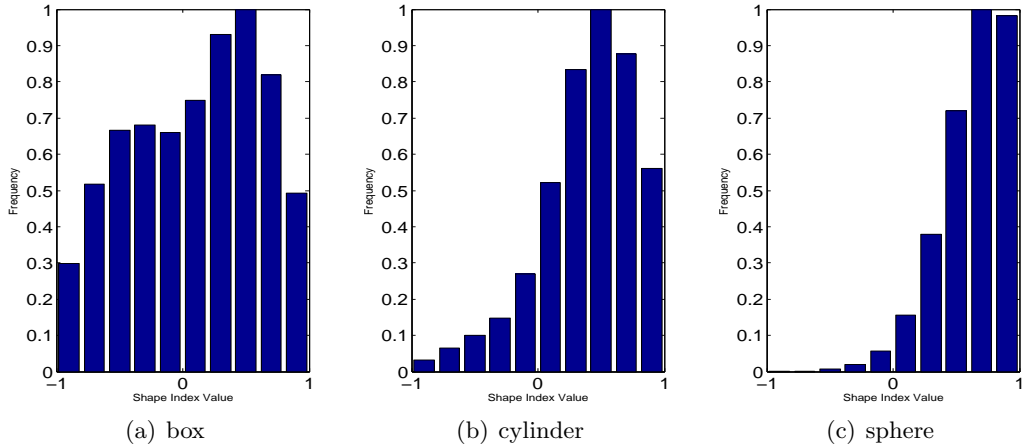


Figure 3.4: Shape index histograms corresponding to objects with different shapes.

- 3D orientation of the object (1 value). As pointed out by [44], the problem of predicting 3D orientation from images is not trivial. Therefore, we choose to train an SVM classifier (see section) which estimates the orientation of the objects. The input to the classifier is the distances between the extreme points of the top view of the object along the eight different orientations (0, 45, 90, 135, 180, 225, 270, 315, shown by different colors in figures 3.5(e) and 3.5(f)). The system is able to achieve an accuracy of over 97% to predict one of the discretized eight directions.
- Size of the object along X, Y and Z axes (3 values). We apply a transformation on objects in order to recover from the predicted orientation to calculate the sizes.

We want to note that, although we used the described feature detectors in our experiments, our methods to derive *verb* and *object concepts* would not be affected if another set of features were used because our methods are unsupervised. To put it another way, an agent with another set of feature detectors could still use our methods to derive concepts represented in its own sensorimotor space.

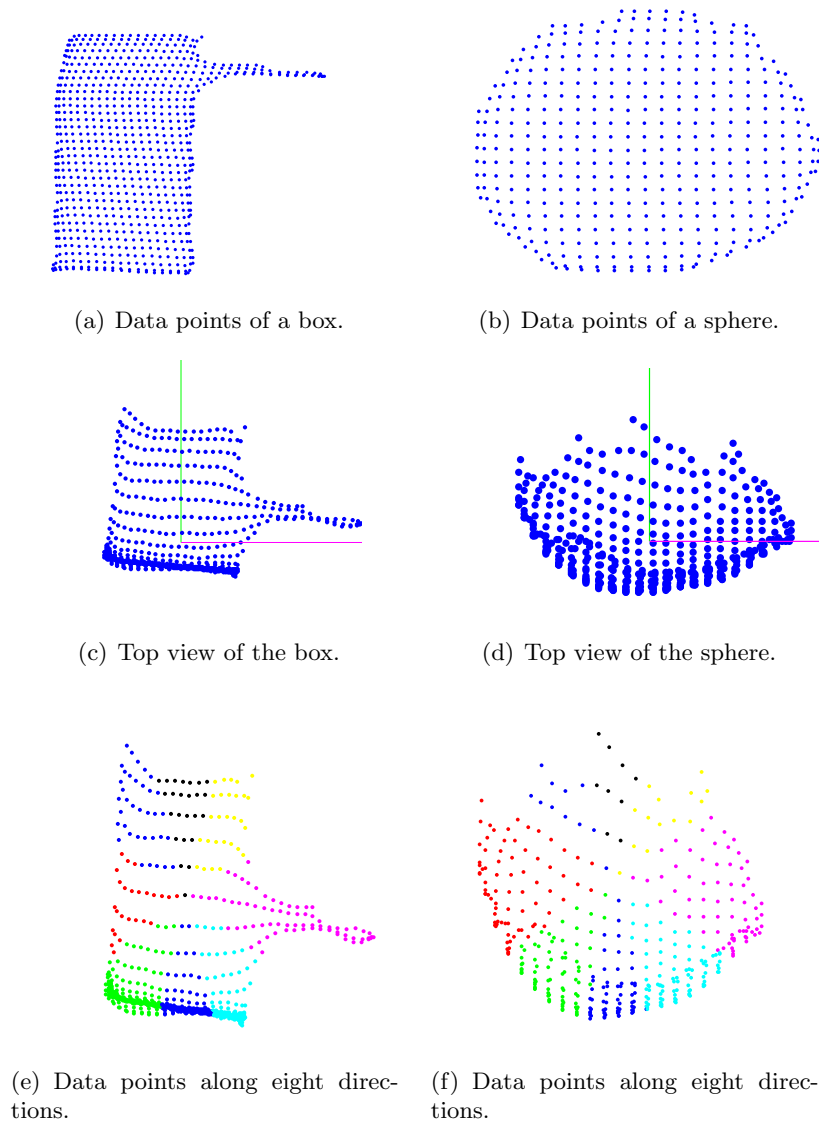


Figure 3.5: Sub-figures **(a)** and **(b)** show the data points from the view point of the camera corresponding to a box and a sphere object, respectively. We transform these points to obtain a top view image of the objects; **(c)** and **(d)**. The sizes of the objects along eight different directions, shown with different colors in **(e)** and **(f)**, are used as inputs to a learning method that predicts one specific orientation.

The features extracted from the object before the behavior are called the *initial features* whereas the features extracted after the behavior are called the *final features*. The difference between the final and the initial features defines the *effect features*. These initial and effect features correspond to entity (*e*) and effect (*f*) in our affordance formulation, respectively. (see equation 2.1).

3.4 Affordance Learning Model

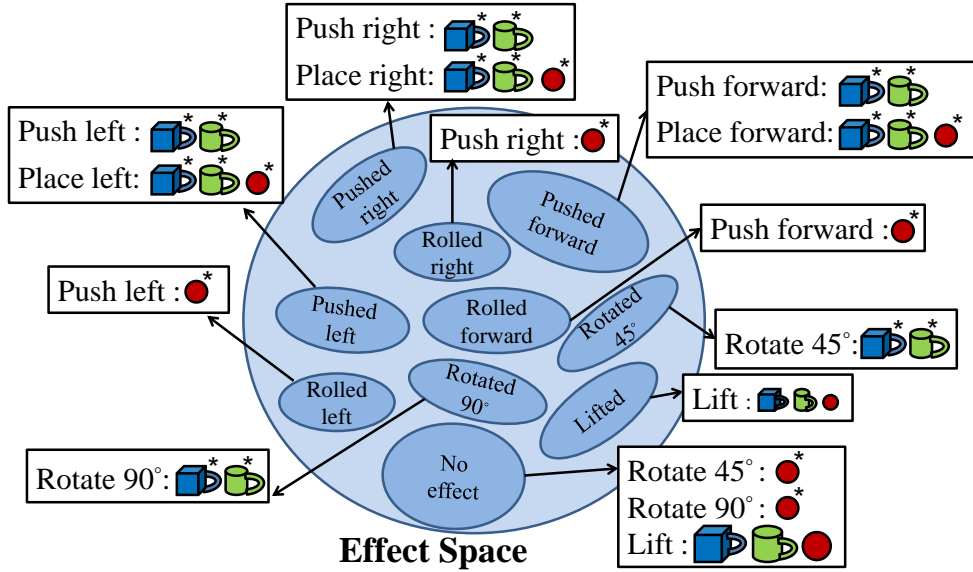


Figure 3.6: Labeled clusters in the effect space.

In order to learn an affordance relation, we assume that clusters in the effect space are labeled by a human supervisor. Figure 3.6 shows the effect clusters derived as a result of such a labeling for our data. Note that the effects arisen from different behaviors can be assigned to same clusters. Using these labels, we first find relevant features for each behavior using the ReliefF feature selection algorithm [45] presented in appendix C. Then, Support Vector Machine classifiers (see appendix B) are trained for each behavior to map the relevant initial features to the effect clusters. These SVM’s are then used to predict which effect cluster a new object can yield to for a given behavior to be applied on the object (see figure 3.7).

In our experiments, we used a threshold of 80% for the feature selection algorithm; i.e. features that are at least as 80% relevant as the most relevant feature to the behavior were selected to be input to the SVM Classifiers. Consequently, trained SVM’s were able to achieve an accuracy of 99% for predicting affordance classes of new samples. This result shows the robustness of our affordance learning schema to predict affordances of objects.

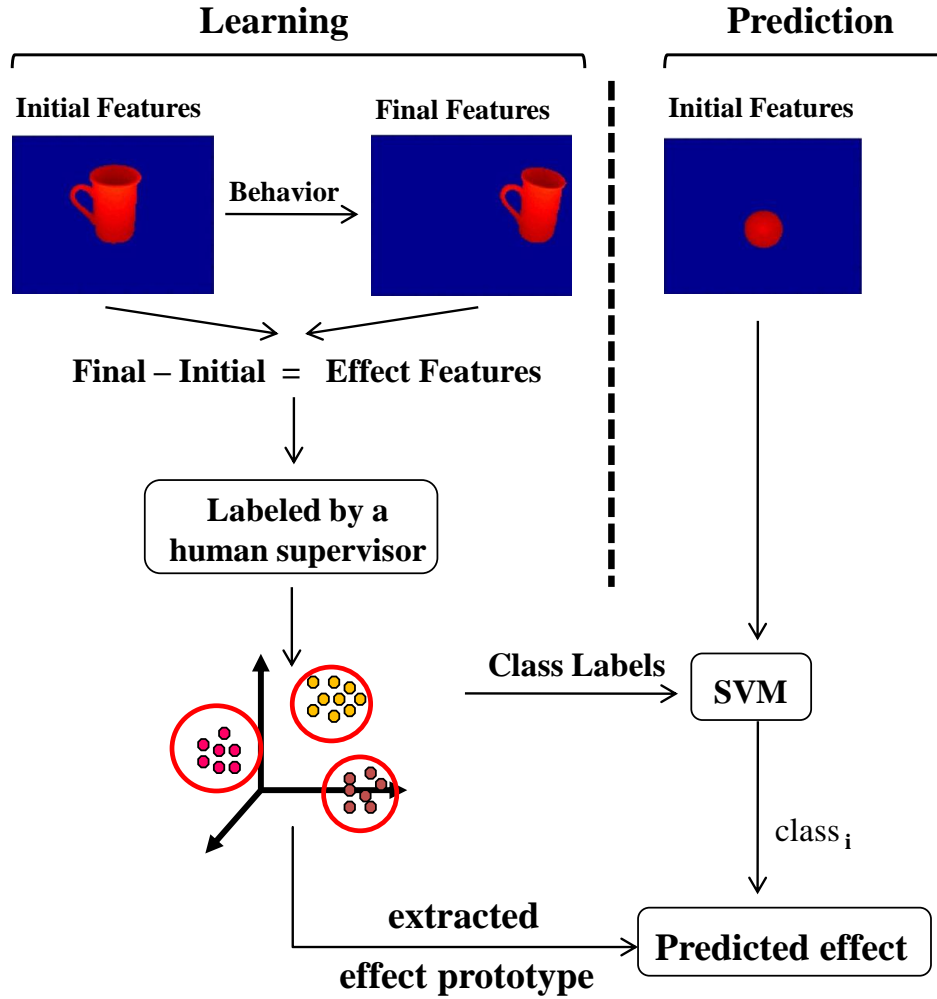


Figure 3.7: Clusters in the effect space are used for training an SVM, which allows predicting the effects of an behavior on a new object.

3.5 Effect Prototype Extraction

In this section, we describe how we derive the condensed representations of *verb concepts*, called *effect prototypes*, given labeled effect clusters in the effect space (figure 3.6).

Figure 3.8 shows distribution of effect features for different effect categories. We observe that, most of the time, an effect category has a consistent increase or decrease effect on some features of the object while leaving others unchanged. We utilize this information to prototypically represent different effects.

Examining the change of feature elements across all effect categories, we observe four different characteristics; feature elements that increase consistently, decrease consistently, stay constant or change in an unpredictable way. Therefore, we represent an effect prototype using labels ‘+’, ‘-’, ‘0’, ‘*’, corresponding to increase, decrease, no change and unpredictable change in the feature element, respectively. In addition to these labels, we also include mean and variance of the changes in the representation to quantify the changes. As a result, we define an *effect prototype* as a string consisting of labels ‘+’, ‘-’, ‘0’, ‘*’, together with two vectors corresponding to mean and variance of the observed changes.

In order to assign ‘+’, ‘-’, ‘0’ and ‘*’, labels to feature elements, we use unsupervised clustering (namely, Robust Growing Neural Gas [46]) in the space of mean and variance. This procedure can be summarized as follows:

1. A set of effects are collected.
2. The mean and variance of each element in this set of effects are computed.
3. Next, feature elements are grouped based on their mean and variance into four clusters using unsupervised clustering (see figure 3.9).
4. Finally, ‘+’, ‘-’, ‘0’ and ‘*’, labels are assigned to the clusters based on what they represent.

3.6 Object Categorization

We use affordances of objects, in order to categorize them by the fact that different objects afford different behaviors (or similar behaviors with different effects). This can be done by taking objects one by one and predicting the effects they can generate. If the effects produced by objects due to the same behaviors are similar, then these objects are also similar. Otherwise, i.e., if objects produce different effects due to the same behaviors, the objects are dissimilar. Based on this, we make a categorization of objects.

For this, we predict effect cluster of objects for each behavior in the repertoire of the

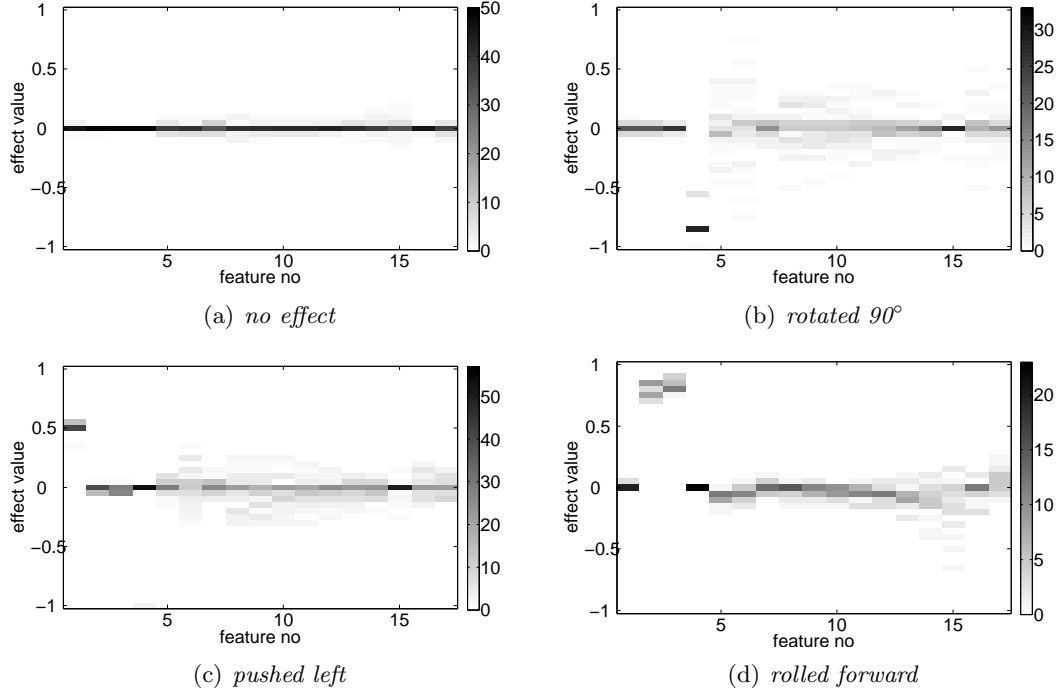


Figure 3.8: The distribution of effect features for different effect clusters. In (a), we see the effect corresponding to no change. As the sub-figure (b) shows, *rotated 90°* effect causes a consistent positive change on the fourth feature, namely the orientation of the objects, whereas other features are not affected significantly. Likewise, *pushed left* effect, sub-figure (c), can be mainly characterized by a positive change in the first feature (corresponding to x position) and *rolled forward* effect, sub-figure (d), by a negative change in the second and third features (corresponding to the y and z positions of the objects).

robot using our affordance learning model presented in section 3.4. Then, nominal k-means clustering is performed in the $\langle y_1, \dots, y_i, \dots, y_n \rangle$ space, where n is the number of behaviors and y_i is the predicted effect cluster for the i_{th} behavior.

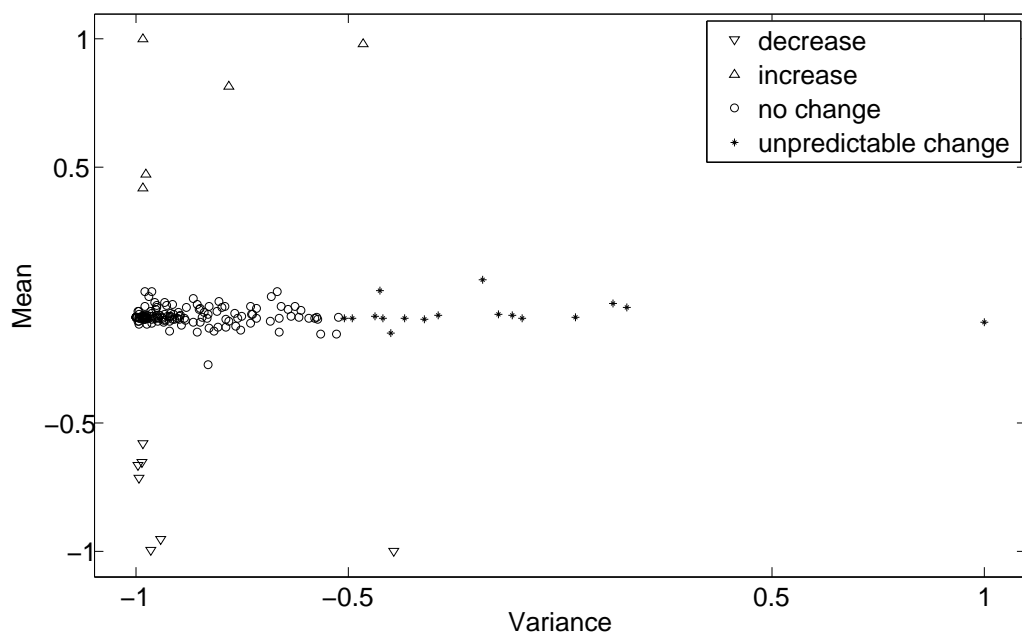


Figure 3.9: Grouping of feature elements into four clusters in the mean-variance space. This grouping is used for assigning '+', '-', '0' and '*', labels to feature elements.

CHAPTER 4

ACQUIRED CONCEPTS

In this chapter, we present the *verb* and *object concepts* acquired using methods introduced in chapter 3. We show that verb concepts can be used as a basis for understanding and imitating others' behaviors or for goal specification tasks on the iCub humanoid robot platform. Next, object concepts acquired through categorization based on objects' affordances are presented. As a final remark, we discuss how our results can lead the robot to discover stable and variable affordances of objects.

4.1 Verb Concepts

As we discussed in section 2.3.1, verbs should correspond to a generalization of behaviors (called *verb concepts*) that achieve the same effect. For deriving *verb concepts*, we find the effect categories in the effect space (figure 3.6) and represent them in a compact way, which we call the *effect prototypes*. We claim that, for a robot, an *effect prototype* can be the label for a *verb concept* which can be linked to a verb.

Using the method described in section 3.5, we extract effect prototypes for our data. Figure 3.9 shows the results of clustering of feature elements in the mean-variance space, that is performed to assign '+', '-', '0' and '*', labels to feature elements. As a result of this clustering, we form the effect prototype strings, that corresponds to *verb concepts* together with mean and variance vectors. The extracted prototype strings are shown in table 4.1.

Now that the robot is able to extract *verb concepts*, it has an understanding of its own and others' behaviors in terms of feature changes their effect creates on the en-

Table 4.1: Effect prototype strings that are extracted using the method introduced in section 3.5. Note that each feature element has an associated mean and variance of the change. θ stands for orientation.

	Position (X-Y-Z)	Shape	Orientation	Size
Pushed Right	+00	0000000000	0	000
Rolled Right	+00	000000**00	0	000
Pushed Left	-00	0000000000	*	000
Rolled Left	-00	0000000*00	0	000
Pushed Forward	0--	0000000000	0	000
Rolled Forward	0--	0000000*00	0	000
No effect	000	0000000000	0	000
Rotated 45°	000	***0000000	+	**0
Rotated 90°	000	***00000*0	+	**0
Lifted	0+-	00*0000000	0	000

vironment. It can exploit this information as a basis for communication with humans or other agents, for example to understand and imitate others' behaviors or for goal specification tasks (see figure 4.1). The robot's ability to understand others' behaviors can be utilized to create a similar effect on objects with similar affordances. Furthermore, by supplying some task in terms of feature changes of the object, the robot can pair the specified goal with one of his behaviors to achieve the goal on an object in the environment. Next, we demonstrate these capabilities of the robot in simple imitation games on the iCub humanoid robot platform.

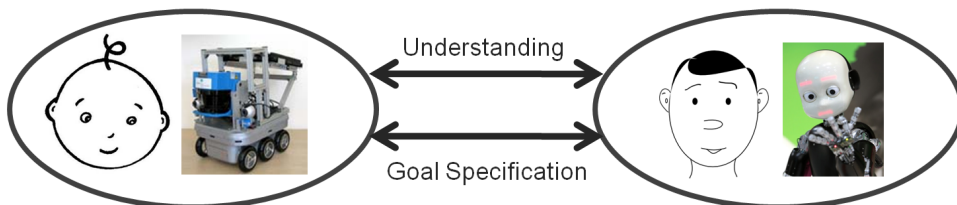


Figure 4.1: A robot can utilize *verb concepts* in order to communicate with humans. We show that the robot can understand and imitate others' behavior or accomplish a specified goal using *verb concepts* he has developed by simple interaction games².

²Images are taken with permission from <http://www.clker.com>. Last access date: 24.09.2010.

4.1.1 Interaction Games

The aim of the first game is to demonstrate that the robot can understand others' behaviors in terms of the *verb concepts* that it has developed. In this scenario, iCub is demonstrated a behavior executed on an object and is required to replicate the effect on a new object. Subsequently iCub selects a behavior in its repertoire to emulate the executed effect corresponding to a *verb concept*. Figure 4.3 shows scenes from two sample runs of game. The steps of the game can be summarized as follows:

- In the demonstration phase, iCub records data before and after the interaction corresponding initial features and final features. Then the effect is calculated by subtracting final features from initial features.
- Next, iCub introduces a new object. Using the affordance learning model described in section 3.4, he goes through all behaviors in his repertoire to infer the effects he is able to create on this object.
- Among the qualified effects, iCub finds the effect that is closest to the demonstrated one. This selection requires that a distance metric exists between the effect categories. We use Mahalanobis distance³ in order to measure the distance between the required effect and prototype of qualified effects. In this comparison, dimensions denoted by a '*' in the prototype string are disregarded, as those correspond to an unpredictable/inconsistent change in the feature elements. As a result, Mahalanobis distance between the required effect f_r and the prototype of the cluster $c^{+,-,0}$ is given by the equation 4.1 where S is the covariance matrix the effect cluster:

$$D(f_r, c^{+,-,0}) = \sqrt{(f_r - c^{+,-,0})^T S^{-1} (f_r - c^{+,-,0})}. \quad (4.1)$$

- As a result, the behavior which results in the selected effect is executed on the new object.

³ In order to measure the distance between a test point and a cluster, Euclidean distance assumes that sample points in the cluster are distributed about the center of mass in a spherical manner which may not be the case for clusters in our effect space (see figure 3.6). Whereas, Mahalanobis distance returns a value proportional to the width of the given cluster in the direction of the test point. Therefore, it gives a better estimation of the cluster's distribution.

Table 4.2 shows the derived confusion matrix between different effect clusters for our data. We calculate distances by taking mean of one cluster and comparing it with the distribution of the other using equation 4.1. As expected, for example, *push right* effect is most similar to itself and then to *rolled right* effect which causes a similar change in the same direction but with a different mean and variance. Likewise, *rotated 45°* is closer to *rotated 90°* effect than other effect, as both cause a variation in the same, namely orientation, direction. The result that *no effect* is very distant to other effects, can be caused by this effect cluster’s small variance in all directions.

Moreover, in a second scenario, we can make iCub to achieve a goal given in the form of a *verb concept* representation. For instance, by providing a string +*****
*****, we may ask him to produce an effect which results in an increase in the first feature, namely x position and disregard the rest. Note that in this context, ‘*’ has an interpretation “disregard the changes in this feature element” rather than the interpretation of “an unpredictable effect”. Therefore, we slightly modify the equation 4.1 to allow for exclusion of the dimensions denoted by a ‘*’ in the given goal string. Figures 4.4(a) and show iCub given two goal strings * — ***** and +***** respectively, which he successfully matches with the right behavior.

4.2 Object Concepts

As pointed out in section 2.3.2, objects can be categorized based on appearance and function. In this section, we show that function-based categorization forms categories of objects based on their affordances and the formed categories can be linked to ‘object concepts’ that we use in language.

We achieve function-based categorization by predicting affordances of objects for different behaviors in the robot’s repertoire (using classifiers trained by our affordance learning model described in section 3.4) and then performing unsupervised clustering of vectors that describe the predicted affordance class of the objects corresponding to different behaviors of the robot. (see section 3.6).

Fig. 4.5 shows categories that emerge as a result of our current setup. We see that the

majority of the ball shaped objects used in the experiments have been gathered in one cluster, namely cluster 2. This shows balls' ability to roll as a result of push behaviors make them easily distinguishable from boxes and cylinders when compared based on their affordances. The other two clusters (cluster 1 and 3) are mostly composed of box and cylinder shaped objects, whereas within these clusters objects are separated in accordance with their sizes. This is the result of box and cylinder shaped objects' similar affordances in response to push and rotate behaviors in the repertoire of the robot. Instead, lift behavior, that is successfully executed on small objects but fails for large objects, decides on the form of these clusters.

Moreover, as shown in table 4.2, the affordance-based categorization proposed in this paper can perform surprisingly well on novel non-symmetrical objects. For example, objects with complex shape features (e.g., the spray, the robot) can be categorized correctly by the system.


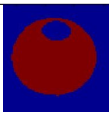

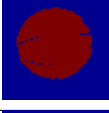

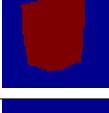

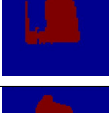

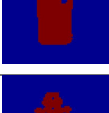
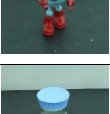
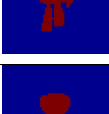

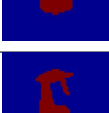
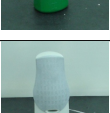
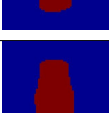


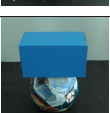
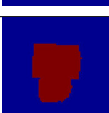




As a result, three categories that correspond to balls, small sized box and cylinders and large sized box and cylinders, emerge from our function-based categorization of objects. This categorization can be further improved as more behaviors are added in repertoire of the robot. More interestingly, a categorization based on both appearance and functional properties of the objects can be performed which yield categories similar to humans.

4.2.1 Stable and Variable Affordances

As noted in section 2.2.1, the aspects of objects that can be changed by manipulation/behaviors of agent can be associated with the variable affordances, whereas others can be associated to stable affordances. Therefore, it is important for an agent to discover the relation between its behaviors and stable/variable properties of objects by himself, considering stable and variable affordances are believed to be sub-served by different neural pathways; namely dorsal and ventral streams. In this section, we show that effect prototypes that represent *verb concepts* in our framework, can also be utilized to infer these properties.

Table 4.1 displays the prototypes of different effects. We see that pushed left, right,

Table 4.2: The detected categories of novel and non-symmetrical objects.

Object name	Object picture	Range image	Category
Lamp shade			O.C. 1
Basketball			O.C. 1
Dried-fruits box.			O.C. 2
Telephone			O.C. 3
Detergent			O.C. 2
Robot			O.C. 2
Tchibo jar			O.C. 1
Spray			O.C. 2
Loudspeaker			O.C. 2
Helmet			O.C. 1
Hybrid Object			O.C. 2
Hybrid Object			O.C. 2

forward and rolled left, right and forward effects cause a consistent change in the position of the objects whereas the rest of the feature vector is irrelevant for these effects. On the other hand, for the rotate effects, the position, shape and the size of the objects are irrelevant but there is considerable change in the estimated orientation of the objects. Looking at these results, one can conclude that position and orientation features of objects used in our experiments, can be changed by manipulation and are therefore variable, whereas size or shape related features remain stable.

We should also note that properties of objects, such as size or orientation, should not be directly referred to as stable and variable affordances, but rather should be associated to an *affordance relation* between an agent and a specific object. For example, although shape is a stable property of objects used in our experiments, it may not be true for an object made of sponge. Likewise, position can be related to a stable affordance in case of a situation where the agent is not able to move an object.

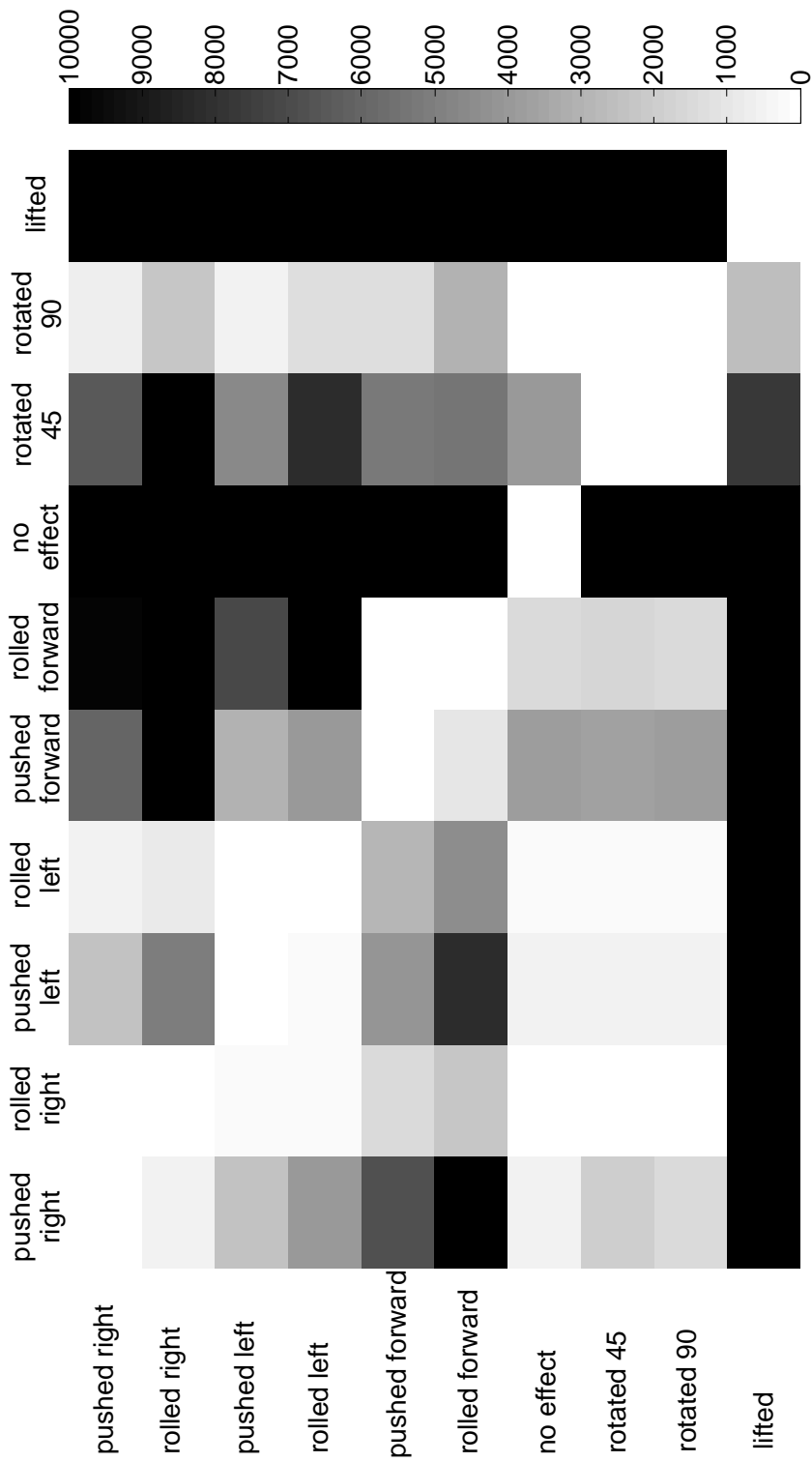


Figure 4.2: The distances between the extracted effect prototypes. We calculate distances by taking mean of one cluster and comparing it with the distribution of the other using equation 4.1.

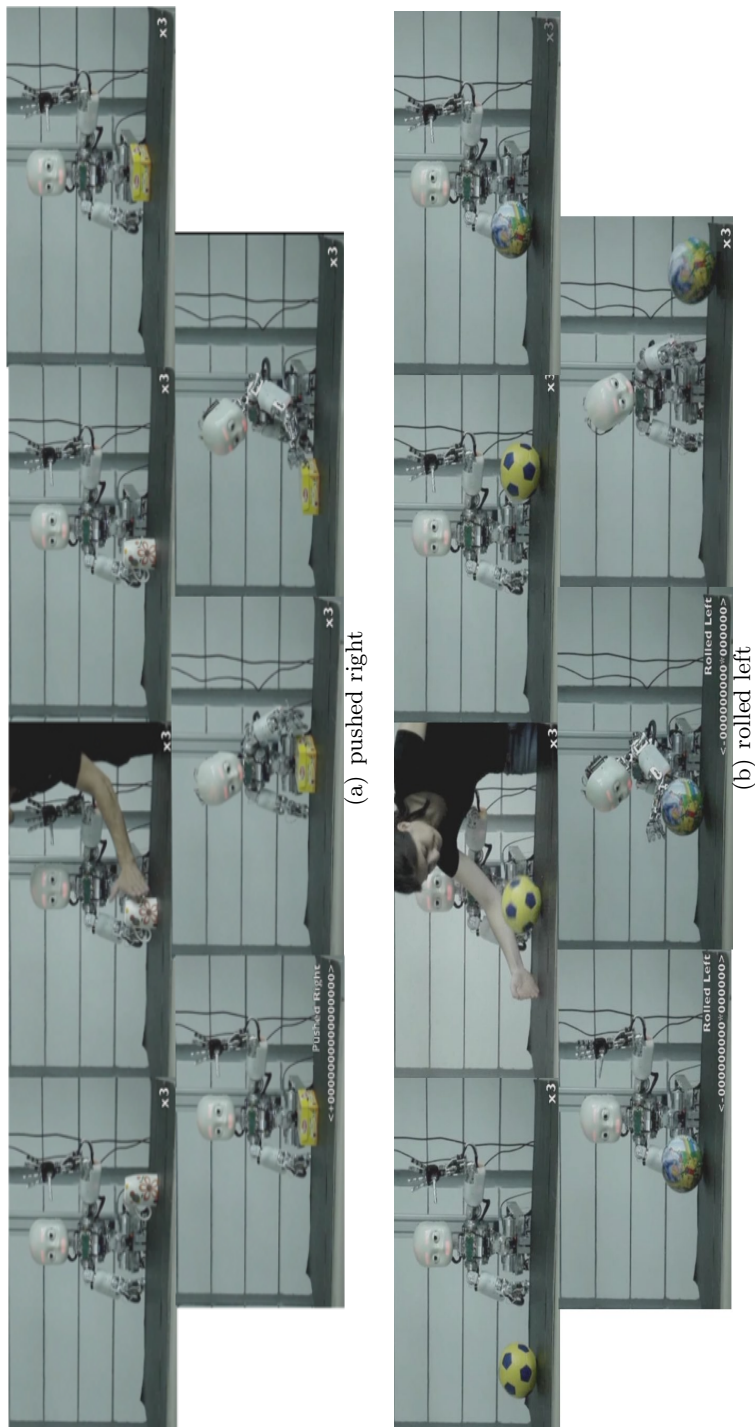


Figure 4.3: (a) In the first line, iCub is demonstrated the *pushed right* effect, which it can successfully match with the corresponding *verb concept*, i.e., +000000000000000000. Then, it is introduced a new object on which it is required to create the same effect. The second line shows iCub executing the *push right* behavior which it successfully chooses among the behaviors in its repertoire. (b) Similarly, iCub is first demonstrated the *rolled left* effect (corresponding to a verb concept, i.e., -000000000*000000) and a new object is put in front of it. It chooses to execute the *push left* behavior to produce the same effect.

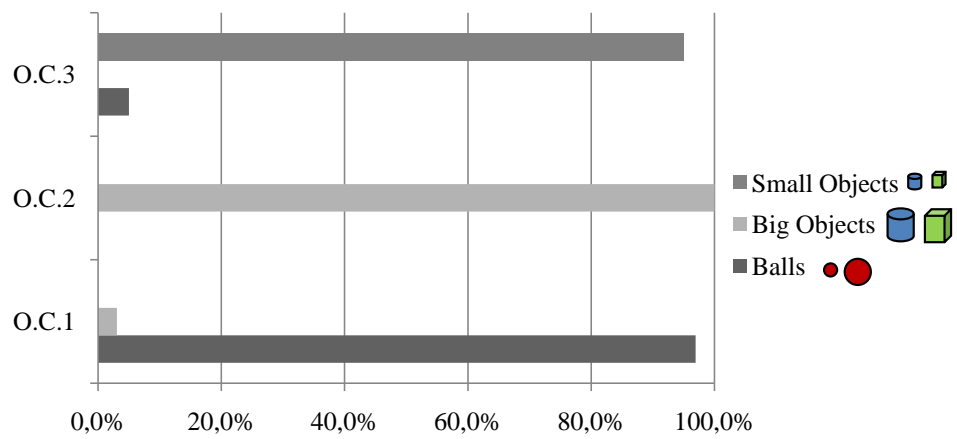


Figure 4.5: Categories of the objects, i.e., object concepts. We see that the interactions of iCub lead to three object concepts O.C.1, O.C.2 and O.C.3, which respectively correspond to balls, big objects (including medium and big cylinders and cubes) and small objects (both cylinders and cubes).

CHAPTER 5

DISCUSSION

In this thesis, we used the notion of affordances to investigate how a humanoid robot can develop *verb* and *object concepts* through interactions with the objects. The acquired concepts are represented in the robot’s own sensorimotor space which means they are grounded in perception and action.

We formalize the interaction of an agent with its environment as an affordance relation between the object in the environment, the behavior of the agent and the generated effect in the environment as a result of executing the behavior on the object. We used an affordance learning model which enables the robot to predict the effects of its behaviors applied on novel objects. As a next step, we argued that the creation of equivalence classes over the components of the affordance relation provides a mechanism for creating abstract categories that can be linked to concepts represented by verbs and nouns.

The literature generally links verbs to specific behaviors of the robot. Instead, we proposed linking verbs with, what we call as, *verb concepts*, which are generalizations over behaviors defined in the effect space of the behaviors. We made experiments in which the robot learns the *effect prototypes* through demonstrations made by a human supervisor. These prototypes represent effects in terms of feature changes in the perception vector of the robot. We demonstrated, that the extracted verb concepts can be utilized by the iCub humanoid robot in order to find the *verb concept* corresponding to an observed behavior, or it can satisfy a goal given in the form of an *effect prototype* corresponding to a *verb concept*.

We also propose to categorize objects using the fact that objects producing similar effects due to the same behaviors are similar and objects producing different effects due to same behavior are dissimilar. We showed that object categories similar to humans can emerge as a result of grouping the predicted affordance classes of objects in an unsupervised manner. We link these categories to *object concepts* corresponding to nouns in language. Additionally, we argued that acquired concepts can lead the robot to discover stable and variable properties of objects which can be associated to stable and variable affordances.

This thesis can be improved in several ways. One way is to incorporate dynamic aspects of behaviors in our model which are known to be important for verb concepts or semantics [36, 37]. Moreover, we can improve our model by integrating appearance and function based categorizations for acquisition of *object concepts* in a developmental manner similar to humans.

REFERENCES

- [1] P. J. Besl and R. C. Jain, “Invariant surface characteristics for 3d object recognition in range images,” *Comput. Vision Graph. Image Process.*, vol. 33, no. 1, pp. 33–80, 1986.
- [2] A. Borghi, “Object concepts and embodiment: Why sensorimotor and cognitive processes cannot be separated.,” *La nuova critica.*, vol. 49-50, pp. 90–107, 2007.
- [3] D. Pecher and R. Zwaan, “Introduction to grounding cognition,” *Grounding cognition: The role of perception and action in memory, language, and thinking*, pp. 1–7, 2005.
- [4] A. Glenberg and M. Kaschak, “Grounding language in action,” *Psychonomic Bulletin & Review*, vol. 9, no. 3, p. 558, 2002.
- [5] P. Fitzpatrick, G. Metta, L. Natale, S. Rao, and G. Sandini, “Learning about objects through action-initial steps towards artificial cognition,” in *IEEE International Conference on Robotics and Automation, 2003. Proceedings. ICRA’03*, vol. 3, 2003.
- [6] H. Kozima, C. Nakagawa, and H. Yano, “Emergence of imitation mediated by objects,” *Lund University Cognitive Studies*, vol. 94, pp. 59–61, 2002.
- [7] A. Stoytchev, “Behavior-grounded representation of tool affordances,” in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation, 2005. ICRA 2005*, pp. 3060–3065, 2005.
- [8] M. Dogar, M. Cakmak, E. Ugur, and E. Sahin, “From primitive behaviors to goal-directed behavior using affordances,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pp. 729–734, 2007.
- [9] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, “Learning object affordances: From sensory–motor coordination to imitation,” *Robotics, IEEE Transactions on [see also Robotics and Automation, IEEE Transactions on]*, vol. 24, no. 1, pp. 15–26, 2008.
- [10] S. C. Want and P. L. Harris, “How do children ape? Applying concepts from the study of non-human primates to the developmental study of imitation in children,” *Developmental Science*, vol. 5, pp. 1–13, 2002.
- [11] B. Elsner, “Infants’ imitation of goal-directed actions: the role of movements and action effects.,” *Acta psychologica*, vol. 124, no. 1, pp. 44–59, 2007.
- [12] A. Turing, “Computing machinery and intelligence,” *Mind*, vol. 59, pp. 433–460, 1950.

- [13] J. R. Searle, “Minds, brains, and programs,” *Behavioral and Brain Sciences*, vol. 3, pp. 417–424, 1980.
- [14] S. Harnad, “The symbol grounding problem,” *Physica*, vol. D, no. 42, pp. 335–346, 1990.
- [15] A. Borghi, “Object concepts and action,” *The grounding of cognition: The role of perception and action in memory, language, thinking*, pp. 8–34, 2005.
- [16] M. Wilson, “Six views of embodied cognition,” *Psychonomic Bulletin & Review*, vol. 9, no. 4, p. 625, 2002.
- [17] G. Rizzolatti, L. Fadiga, V. Gallese, and L. Fogassi, “Premotor cortex and the recognition of motor actions,” *Cognitive brain research*, vol. 3, no. 2, pp. 131–141, 1996.
- [18] A. Glenberg, “Language and action: Creating sensible combinations of ideas,” *The Oxford handbook of psycholinguistics*, pp. 361–370, 2007.
- [19] J. Fodor, *Concepts: Where cognitive science went wrong*. Oxford University Press, USA, 1998.
- [20] L. Barsalou, “Perceptual symbol systems,” *Behavioral and brain sciences*, vol. 22, no. 04, pp. 577–660, 1999.
- [21] D. Roy and A. Pentland, “Learning words from sights and sounds: A computational model,” *Cognitive science*, vol. 26, no. 1, pp. 113–146, 2002.
- [22] Y. Sugita and J. Tani, “Learning semantic combinatoriality from the interaction between linguistic and behavioral processes,” *Adaptive Behavior*, vol. 13, no. 1, p. 33, 2005.
- [23] L. Steels, “Evolving grounded communication for robots,” *Trends in Cognitive Science*, pp. 308–312, 2003.
- [24] A. Cangelosi and T. Riga, “An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots,” *Cognitive science*, vol. 30, no. 4, pp. 673–689, 2006.
- [25] L. Steels and F. Kaplan, “AIBO’s first words: The social learning of language and meaning,” *Evolution of Communication*, vol. 4, no. 1, pp. 3–32, 2002.
- [26] F. Kaplan, P. Oudeyer, and B. Bergen, “Computational models in the debate over language learnability,” *infant and child development*, vol. 17, no. 1, pp. 55–80, 2008.
- [27] J. J. Gibson, *The Ecological Approach to visual perception*. Lawrence Erlbaum Associates, 1986.
- [28] M. Turvey, “Affordances and prospective control: An outline of the ontology,” *Ecological Psychology*, vol. 4, no. 3, pp. 173–187, 1992.
- [29] M. Steedman, “Plans, affordances, and combinatory grammar,” *Linguistics and Philosophy*, vol. 25, no. 5, pp. 723–753, 2002.

- [30] T. Stoffregen, “Affordances as properties of the animal-environment system,” *Ecological Psychology*, vol. 15, no. 2, pp. 115–134, 2003.
- [31] A. Chemero, “An outline of a theory of affordances,” *Ecological Psychology*, vol. 15, no. 2, pp. 181–195, 2003.
- [32] E. Şahin, M. Çakmak, M. R. Doğan, E. Uğur, and G. Üçoluk, “To Afford or Not to Afford: A New Formalization of Affordances Toward Affordance-Based Robot Control,” *Adaptive Behavior*, vol. 15, no. 4, pp. 447–472, 2007.
- [33] A. Milner and M. Goodale, *The visual brain in action*. Oxford University Press, USA, 1996.
- [34] A. Borghi and L. Riggio, “Sentence comprehension and simulation of object temporary, canonical and stable affordances,” *Brain Research*, vol. 1253, pp. 117–128, 2009.
- [35] M. Rudolph, M. Muhlig, M. Gienger, and H.-J. Bohme, “Learning the consequences of actions: Representing effects as feature changes,” *Int. Symposium on Learning and Adaptive Behavior in Robotic Systems*, 2010.
- [36] P. Cohen, C. Morrison, and E. Cannon, “Maps for verbs: The relation between interaction dynamics and verb use,” in *Proceedings of the Nineteenth International Conference on Artificial Intelligence (IJCAI 2005)*, 2005.
- [37] E. Cannon and P. Cohen, “Talk about motion: The semantic representation of verbs by motion dynamics,” in *Language and Space*, Ed: Linda Smith, 2006.
- [38] D. Marocco, A. Cangelosi, K. Fischer, and T. Belpaeme, “Grounding action words in the sensorimotor interaction with the world: experiments with a simulated icub humanoid robot,” *Frontiers in Neurobotics*, vol. 4, no. 7, 2010.
- [39] A. M. Borghi, A. D. Ferdinando, and D. Parisi, “The role of perception and action in object categorization,” in *J.A. Bullinaria & W. Lowe (Eds), Connectionist Models of Cognition and Perception*. Singapore: World Scientific, pp. 40–50, 2002.
- [40] L. Barsalou, “Ad hoc categories,” *Memory and cognition*, vol. 11, no. 3, pp. 211–227, 1983.
- [41] S. Nolfi and D. Marocco, “Active perception: A sensorimotor account of object categorization,” in *From Animals to Animats 7: Proceeding on the Sixth International Conference on Simulation of Adaptive Behavior*, 2002.
- [42] J. Sun, J. Moore, A. Bobick, and J. Rehg, “Learning Visual Object Categories for Robot Affordance Prediction,” *The International Journal of Robotics Research*, vol. 29, no. 2-3, p. 174, 2010.
- [43] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, “The iCub humanoid robot: an open platform for research in embodied cognition,” in *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, pp. 50–56, ACM, 2008.
- [44] A. Saxena, J. Driemeyer, and A. Y. Ng, “Learning 3d object orientation from images,” in *NIPS workshop on Robotic Challenges for Machine Learning*, 2007.

- [45] I. Kononenko, “Estimating attributes: Analysis and extensions of relief,” in *European Conference on Machine Learning* (F. Bergadano and L. D. Raedt, eds.), pp. 171–182, Springer, 1994.
- [46] A. Qin and P. Suganthan, “Robust growing neural gas algorithm with application in cluster analysis,” *Neural Networks*, vol. 17, no. 8-9, pp. 1135–1148, 2004.
- [47] A. Hoover, G. Jean-Baptiste, X. Jiang, P. Flynn, H. Bunke, D. Goldgof, K. Bowyer, D. Eggert, A. Fitzgibbon, and R. Fisher, “An experimental comparison of range image segmentation algorithms,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 18, no. 7, pp. 673–689, 1996.
- [48] M. Jaesik, M. W. Powell, and K. W. Bowyer, “Automated performance evaluation of range image segmentation,” in *Applications of Computer Vision, 2000, Fifth IEEE Workshop on.*, pp. 163–168, 2000.
- [49] E. Ugur, E. Sahin, and E. Oztop, “Affordance learning from range data for multi-step planning,” in *Proc. of the ninth int. conf. on epigenetic robotics*, vol. 146, pp. 177–184, 2009.
- [50] E. Ugur, M. R. Dogar, and E. Sahin, “Learning object affordances for planning,” *ICRA09 Workshop on Sensorimotor Learning*, 2009.
- [51] J. L. Crowley, F. Wallner, and B. Schiele, “Position estimation using principal components of range data,” in *Robotics and Autonomous Systems*, vol. 4, pp. 3121–3128 vol.4, May 1998.
- [52] H. Shum, K. Ikeuchi, and R. Reddy, “Principal component analysis with missing data and its application to polyhedral object modeling,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 9, pp. 854–867, 1995.
- [53] D. G. Lowe, “Object recognition from local scale-invariant features,” *Proceedings of the International Conference on Computer Vision*, 1999.
- [54] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [55] Y. Ke and R. Sukthankar, “PCA-SIFT: A more distinctive representation for local image descriptors,” *IEEE Computer Society*, 2004.
- [56] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1615–1630, 2005.
- [57] H. Bay, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” *Computer Vision–ECCV 2006*, pp. 404–417, 2006.
- [58] T.-W. R. Lo and J. P. Siebert, “Local feature extraction and matching on range images: 2.5d sift,” *Computer Vision and Image Understanding (2009)*, doi: 10.1016/j.cviu.2009.06.005, 2009.
- [59] T.-W. R. Lo, J. Siebert, and A. Ayoub, “Robust feature extraction for range images interpretation using local topology statistics,” in *Proceedings of MICCAI 2006 Workshop on Craniofacial Image Analysis for Biology, Clinical Genetics, Diagnostics and Treatment*, pp. 75–82, 2006.

- [60] J. Flusser and T. Suk, “A moment-based approach to registration of images with affine geometric distortion,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 2, 1994.
- [61] A. Carkacioglu and F. T. Yarman-Vural, “Set of texture similarity measures,” *Electronic Imagi 97, Machine Vision Applications in Industrial Inspection, SPIE Proceedings*, vol. 3029, pp. 118–127, 1997.
- [62] V. N. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, September 1998.
- [63] M. Aizerman, E. Braverman, and L. Rozonoèr, “Theoretical foundations of the potential function method in pattern recognition learning,” *Automation and remote control*, vol. 25, no. 6, pp. 821–837, 1964.
- [64] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [65] K. Kira and L. A. Rendell, “A practical approach to feature selection,” in *Ninth International Workshop on Machine Learning* (D. H. Sleeman and P. Edwards, eds.), pp. 249–256, Morgan Kaufmann, 1992.
- [66] M. Robnik-Šikonja and I. Kononenko, “Theoretical and empirical analysis of relieff and rrelieff,” *Mach. Learn.*, vol. 53, pp. 23–69, October 2003.
- [67] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, “The WEKA data mining software: An update,” *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

APPENDIX A

Feature Extraction Methods

In this chapter, we present a number of methods that are used in literature to extract features from range images.

A.1 Surface normal based methods

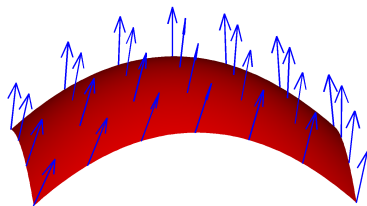


Figure A.1: Surface normals¹.

The surface normal (n_x, n_y, n_z) corresponding to a point represented by coordinates (x, y, z) can be calculated using the first order partial derivatives as in equation A.1 [47, 48]:

$$(n_x, n_y, n_z) = \frac{\frac{\partial z}{\partial x} \times \frac{\partial z}{\partial y}}{\left| \frac{\partial z}{\partial x} \times \frac{\partial z}{\partial y} \right|}. \quad (\text{A.1})$$

[49, 50] make use of the frequency histograms of normal vector angles in latitude and longitude. These histograms are very robust for distinguishing between objects with different shapes. However, this approach tightly couples orientation and shape information and views of the same object from different orientations may result in shifted histograms.

¹Image is taken with permission from http://commons.wikimedia.org/wiki/File:Surface_vectors.png. Last access date: 22.09.2010.

A.2 Principal Component Analysis (PCA) based methods

Principle Component Analysis is a method to find a new basis for the data in which variables co-vary as little as possible. In this new basis, initial principle components accounts for large variances and assumed to correspond to important dynamics, whereas succeeding components accounts for low variances corresponding to noise.

Studies by [51, 52] make use of principle components of the range data sets to accomplish different tasks. [51] constructs an eigenspace from the principal components of a large number of range data sets to be used in the position and orientation estimation of the robot. [52] shows that image modelling from a sequence of range images using PCA is problematic due to missing data and tries to generalize it as a weighted least squares problem.

A.3 Image descriptor based methods

A.3.1 Scale Invariant Feature Transform (SIFT)

Scale Invariant Feature Transform (SIFT) algorithm, originally proposed in [53] and refined in [54], uses local image descriptors sampled at different locations in order to achieve efficient object recognition. The features extracted from SIFT algorithm are invariant to image scaling and rotation. They are also partially invariant to illumination and 3D camera viewpoint changes. The algorithm first identifies potential key-points in the Gaussian scale space of the image and then eliminates some of them in order to decide on actual key-points. Next, a consistent orientation is assigned to each key-point using the peak of the orientation histogram formed considering a region around the point. Finally, a 128 element descriptor is calculated for each key-point which is proved to be highly distinctive for performing object or scene matching.

SIFT has other variations such as PCA-SIFT (Principal Components Analysis applied to SIFT descriptors)[55], GLOH (Gradient Location and Orientation Histogram)[56] and SURF(Speeded-Up Robust Features)[57] which show similar performance on descriptor matching for object recognition.

A.3.2 2.5 SIFT

The general SIFT algorithm is actually intended to be used with 2D intensity images. 2.5 SIFT algorithm, presented in [58, 59] adapts the general SIFT algorithm to be effectively used with range images. They derive feature descriptors based on shape index histograms and range gradient orientations of key-point locations. Additionally, using range image normals, they assign a consistent canonical local slant and local tilt in order to facilitate invariance to 3D rotational changes.

A.4 Image Moments

Affine image moment invariants can be useful for representing shape from 3D point clouds. Given the central moment μ_{pq} of order $(p + q)$ as:

$$\mu_{pq} = \int \int_G (x - x_t)^p (y - y_t)^q dx dy, \quad (\text{A.2})$$

where (x_t, y_t) is the center of gravity for the object G , one can define the affine image moment invariants measuring the area, distribution of the points around the center, symmetry of the shape (and more) (see [60] for details).

A.5 Methods inspired from texture representation

Textures are repetitive patterns in images. Their extraction, representation and classification is a long-studied topic in computer vision [61]. Range data processing approaches might benefit from the study of texture. Below, we list a set of approaches and metrics which might be easily transformed to range data (by changing gray levels with depth information, for example):

- **First-order statistical methods:**

Assume that G is the number of gray levels, and h_i is the number of pixels in an image with gray level i , and the normalized histogram H_i is defined as $H_i = h_i/N$.

- **Mean gray level:**

$$\mu = \sum_{i=0}^{G-1} iH_i. \quad (\text{A.3})$$

– **Gray level standard deviation:**

$$\sigma = \sqrt{\sum_{i=0}^{G-1} (i - \mu)^2 H_i}. \quad (\text{A.4})$$

– **Coefficient of variation:**

Coefficient of variation is a measure of relative dispersion:

$$cv = \frac{\sigma}{\mu}. \quad (\text{A.5})$$

– **Skewness:**

Skewness γ_1 is a measure of the symmetry of the (gray level) histogram:

$$\gamma_1 = \frac{1}{\sigma^3} \sum_{i=0}^{G-1} (i - \mu)^3 H_i. \quad (\text{A.6})$$

– **Kurtosis:**

Kurtosis γ_2 is a measure of whether the histogram is peaked or flat relative to a normal distribution:

$$\gamma_2 = \frac{1}{\sigma^4} \sum_{i=0}^{G-1} (i - \mu)^4 H_i - 3. \quad (\text{A.7})$$

– **Energy:**

Energy measures the non-uniformity of the histogram:

$$e = \sum_{i=0}^{G-1} H_i^2. \quad (\text{A.8})$$

– **Entropy:**

Entropy measures the uniformity of the histogram:

$$s = - \sum_{i=0}^{G-1} H_i \log H_i. \quad (\text{A.9})$$

- **Second-order statistical methods:** Second-order methods rely on gray-level relations between pixels. The methods depend on c_{ij} which is the number of pixels having gray level j displaced h (a constant given by the user) relative to a point having gray level i . C_{ij} is the normalized version of c_{ij} , i.e., $C_{ij} = c_{ij}/N_h$. Given C_{ij} , we can define the following second-order features:

– **Energy or Angular Second Moment:**

Homogeneity (in gray level) in the image can be measured with:

$$\epsilon = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} C_{ij}^2. \quad (\text{A.10})$$

– **Entropy:**

Randomness of the image can be measured with:

$$S = - \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} C_{ij} \log C_{ij}. \quad (\text{A.11})$$

– **Contrast:**

Local variations of gray levels, i.e., coarseness of texture, can be measured with:

$$C = - \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} (i - j)^2 C_{ij}. \quad (\text{A.12})$$

– **Homogeneity:**

Monotonicity in gray levels can be defined as:

$$H = - \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \frac{C_{ij}}{1 + |i - j|}. \quad (\text{A.13})$$

– **Auto-correlation:**

Gray-level linear dependencies:

$$\rho = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \frac{(i - \mu_x)(j - \mu_y)C_{ij}}{\sigma_x \sigma_y}. \quad (\text{A.14})$$

– **Diagonal Moment:**

The difference in correlation for both high and low gray levels:

$$D = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} |i - j|(i + j - \mu_x - \mu_y)C_{ij}. \quad (\text{A.15})$$

See [61] for more details and for more complicated features.

APPENDIX B

Support Vector Machine (SVM) Classifier

Support Vector Machines, introduced by Vladimir Vapnik in 1998 [62], are methods used for statistical classification and regression analysis, i.e., given a set of labelled training instances, they build a model in order to predict which class a new sample will fall into.

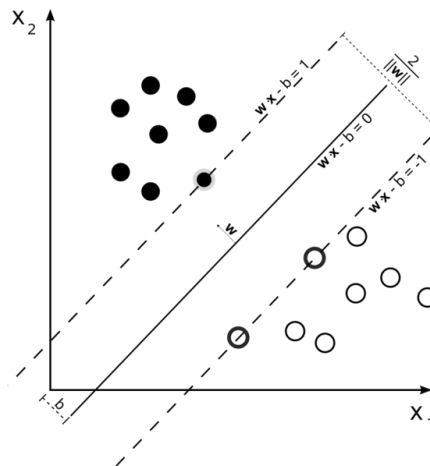


Figure B.1: Support vector machines use nearest instances on each side, called support vectors, to find a maximum margin hyper-plane in the feature space².

SVM classifiers first perform a nonlinear mapping of the data into a high-dimensional feature space, in which data is assumed to be linearly separable, using kernel functions. Since linear algorithm in the range space of kernel function is equivalent to the non-linear algorithm in the input space, calculations are still performed in the input space and therefore are still efficient (see *kernel trick* [63]). In the next step, the optimal hyper-plane (or hyper-planes for multi-class classification) in the high-dimensional

²Image is taken with permission from http://commons.wikimedia.org/wiki/File:Svm_max_sep_hyperplane_with_margin.png. Last access date: 22.09.2010.

feature space that separates classes is found. SVM's try to find the *maximum-margin hyperplane*, that maximizes the distance from it to the nearest data point on each side. These nearest data points in the training set are called the *support vectors*. (B.1). Then, a class label of a new instance is predicted based on location of the instance with respect to the hyperplane in the feature space.

In our implementation we use LIBSVM[64], an integrated software for support vector classification.

APPENDIX C

Relief and ReliefF Algorithms for Feature Selection

In real-world problems, data can potentially be represented by many features, although only a few of them are relevant to the task at hand. For example, in an experiment, objects may be represented by different features related to color, size, shape or texture of the object, whereas only size related features may be relevant for a task such as grasping. Feature selection methods can help to speed up learning algorithms by reducing number of dimensions and can improve quality by eliminating irrelevant and noisy features.

The Relief algorithm, first proposed by Kira and Rendell[65], is a statistical method which returns a weight vector which estimates quality of features based on their discriminative power between instances that are near each other. Unlike other methods which assume that features are independent, this method is also applicable to domains with strong dependencies between features.

Algorithm 1 Pseudo code for Relief[66]

Input: For each training instance a vector of attribute values and the class value.

Output: The vector W of estimations of the qualities of attributes.

```
set all weights  $W[A] := 0.0$ ;  
for  $i = 1$  to  $m$  do  
    randomly select an instance  $R_i$ ;  
    find nearest hit  $H$  and nearest miss  $M$ ;  
    for  $A = 1$  to  $a$  do  
         $W[A] := W[A] - \text{diff}(A, R_i, H)/m + \text{diff}(A, R_i, M)/m$ ;  
    end for  
end for
```

Relief algorithm is given in algorithm 1. For an instance R_i , the algorithm finds *item*

(1) nearest neighbor from the same class (nearest hit H), *item* (2) nearest neighbor from the different class (nearest miss M). Then, for each feature A , its weight $W[A]$ is decreased by the difference between the values of the attribute A for instance R_i and its nearest hit H (i.e. $\text{diff}(A; R_i; H)$). Also each weight is increased by the difference between the values of the attribute A for instance R_i and its nearest miss M . (i.e. $\text{diff}(A; R_i; M)$). This procedure is repeated m times, a parameter defined by the user.

Relief can be used for evaluating weights of features with nominal and numerical values. However, it cannot deal with noisy, incomplete and multi-class data sets. ReliefF algorithm[45] is an extension of Relief to address these problems.

ReliefF algorithm is similar to Relief, however it searches not only one but k of its nearest neighbors from the same class (nearest hits) and k nearest neighbors from each of the different classes (nearest misses). The weight of each feature is updated considering the average of all k hits and misses. This algorithm can also deal with missing data by probabilistically estimating missing values.

In our experiments, we use ReliefF implementation of WEKA[67], an open source software with a large collection of machine learning algorithms.