

A GROUNDED AND CONTEXTUALIZED WEB OF CONCEPTS ON A  
HUMANOID ROBOT

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

BY

HANDE ÇELİKKANAT

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR  
THE DEGREE OF DOCTOR OF PHILOSOPHY  
IN  
COMPUTER ENGINEERING

SEPTEMBER 2015



Approval of the thesis:

**A GROUNDED AND CONTEXTUALIZED WEB OF CONCEPTS ON A  
HUMANOID ROBOT**

submitted by **HANDE ÇELİKKANAT** in partial fulfillment of the requirements for  
the degree of **Doctor of Philosophy in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver  
Dean, Graduate School of **Natural and Applied Sciences**

\_\_\_\_\_

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

\_\_\_\_\_

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

\_\_\_\_\_

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

\_\_\_\_\_

**Examining Committee Members:**

Prof. Dr. Göktürk Üçoluk  
Computer Engineering Department, METU

\_\_\_\_\_

Assist. Prof. Dr. Sinan Kalkan  
Computer Engineering Department, METU

\_\_\_\_\_

Prof. Dr. Aydan Erkmen  
Electrical and Electronics Engineering Department, METU

\_\_\_\_\_

Prof. Dr. Billur Barshan  
Electrical and Electronics Engineering Department, Bilkent Uni.

\_\_\_\_\_

Assist. Prof. Dr. Didem Gökçay  
Health Informatics, METU

\_\_\_\_\_

**Date:**

**08.09.2015**

**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: HANDE ÇELİKKANAT

Signature :



# ABSTRACT

## A GROUNDED AND CONTEXTUALIZED WEB OF CONCEPTS ON A HUMANOID ROBOT

Çelikkanat, Hande

Ph.D., Department of Computer Engineering

Supervisor : Assist. Prof. Dr. Sinan Kalkan

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

September 2015, 140 pages

In this thesis, we propose a formalization for a densely connected representation of concepts and their contexts on a humanoid robot platform. Although concepts have been studied implicitly and explicitly in numerous studies before, our study is unique in placing the relatedness of concepts to the center: We hypothesize that a concept is fully meaningful only when considered in relation to the other concepts. Thus, we propose a novel densely connected web of concepts, and show how utilizing the relatedness of concepts can take cognition one step forward from the conventional approach that treats them individually. Then we use this densely connected framework for determining the context of encountered scenes. Although unanimously accepted as one of the pillars of cognition, our study is the first attempt to provide a dedicated and general formalization of context in a robotics setting. We follow a developmental approach in which the robot determines the existing contexts in its environment in an unsupervised manner, associates seen objects and whole scenes with these contexts as appropriate, and further utilizes this extracted contextual information in reasoning and planning. As required by the developmental paradigm, the programmer's input to the robot in terms of informational bias is kept at a minimum, and the robot is expected to deduce the important characteristics of the environment itself, such as the number of contexts hidden in its environment, if and when to introduce another context to its world model, and how these contexts probabilistically give rise to the

related concepts in this world.

**Keywords:** Concepts, Concept Web, Context, Conceptualization, Symbol-Grounding, Developmental Robotics, Cognitive Robotics

## ÖZ

### İNSANSI ROBOTLAR İÇİN TEMELLENDİRİLMİŞ VE BAĞLAMSAL BİR KAVRAM AĞI

Çelikkanat, Hande

Doktora, Bilgisayar Mühendisliği Bölümü

Tez Yöneticisi : Yrd. Doç. Dr. Sinan Kalkan

Ortak Tez Yöneticisi : Doç. Dr. Erol Şahin

Eylül 2015 , 140 sayfa

Bu tezde, bir insansı robot platformunda kavramların (*concepts*) bir ağ formasyonu kullanılarak temsil edilmesini ve ilgili bağlamlarının (*contexts*) çıkarılmasını destekleyen bir formalizasyon önermekteyiz. Literatürde kavramlama yetisinin doğrudan ya da dolaylı olarak incelendiği çok sayıda çalışma bulunsa da, bu, kavramların birbirleriyle olan bağlantılarına odaklanan ilk çalışmadır: Temel hipotezimiz, bir kavramın ancak diğer kavramlarla ilişki içinde ele alındığında tam olarak anlam kazanacağı yönündedir. Bu hipotezden yola çıkarak, tezin ilk kısmında, kavramları temsil etmek için yoğun bağlantılı (*densely connected*) bir kavram ağı önermekte, sonrasında ise bu yaklaşımın, kavramları birbirinden bağımsız olarak işlemeye kıyasla sunduğu avantajları göstermekteyiz. Tezin ikinci kısmında ise, bu yoğun bağlantılı kavram ağını, robotun karşılaştığı ortamlarda bağlamı tespit etmek için bir temel olarak kullanmaktayız. Bağlamın çeşitli bilişsel yetilerdeki etkin rolü neredeyse evrensel olarak kabul edilse de, bu çalışma, robotlarda bağlamı genel anlamıyla, önceden belirlenmiş uygulama senaryolarına kısıtlı kalmadan formalize etmeye yönelik ilk çalışmadır. Gelişimsel robotik yaklaşımına uygun olarak, robotun ortamındaki bağlamları eğitici (*unsupervised*) bir şekilde, kendi kendine keşfetmesi, karşılaştığı nesnelere ve sahneleri bu bağlamlarla ilişkilendirmesi ve keşfettiği bağlamsal bilgiyi kendi akıl yürütmesini ve davranışını yönlendirecek şekilde kullanabilmesi sağlanmıştır. Gelişimsel yaklaşımın gerekleri doğrultusunda, programcının robota beslediği hazır bilgi girdisi

en düşük seviyede tutulmakta ve robotun, ortamın önemli karakteristiklerini, örneğin ortamda gizli bağlam sayısını, dünya modeline yeni bir bağlamın ne zaman eklenmesi gerektiğini ve bağlamların kavramlarla olasılıksal olarak nasıl ilişkili olduğunu kendi kendine keşfetmesi beklenmektedir.

**Anahtar Kelimeler:** Kavramlar, Kavram Ağı, Bağlam, Kavramlama, Sembol Temelendirme, Gelişimsel Robotik, Bilişsel Robotik

*To my family, who loved me more than anything,  
and to the ones who came to my life, and loved me, and volunteered to be my family.  
Not a day goes by without being thankful for you.*

## ACKNOWLEDGMENTS

All in all, The Thesis is not totally unlike Middle Earth: The journey is so unbelievably long and tricky that you “meet many foes, some open, and some disguised; and find friends along your way when you least look for it.” The Thesis, for me, is much more than the Work, it is what gave me my life-long friends, by showing there are people willing to make a stand with me in the face of unknown and nothingness. To those friends, for their love and patience, I dedicate this humble work of long years.

To Sinan Kalkan, known in Europe as the Turkish avalanche. Supervisor, wise person, best friend, helper in dark times, an ‘advise’-er in the most general sense of the word. It is difficult to know where one ends and the other begins. For never giving up belief in me. For guiding me calmly through all the downs of this sometimes very downy process. For keeping me sane with an incessant supply of humor through trademark one-liners and practical jokes, made with a completely straight face. For showing me perfection is a dedication foremost. For setting the example of what a researcher and a human should be. Most of all, for the unyielding support that he will always be.

To Erol Şahin, my incredible, tireless, perfectionist, super-hero supervisor. What an amazing vision! Every time I think small, focusing only on the completion of some little task, I think of you, try to imitate you, in your vision of building a lab from nothing, in your dream of developing something tangible, in believing “Why can’t we do even better?” What our country needs is more from you, who are not satisfied with what is, but constantly dreams of what can be. You have raised us all, now almost fifty, made us into scientists. We are what we are, only because you have believed.

To Göktürk Üçoluk, for sharing many a laughs! For the great teacher who showed me how good you can teach a person something really complicated (ie, how a damn computer actually works), that she remembers every word perfectly more than 10 years later. For gifting our struggling country quite a number of capable computer scientists who can actually think. All of us learned from you, and we will have you to try to resemble, in character, in integrity, and in the neverending enthusiasm to teach.

To Emre Uğur, for sometimes caring for my future more than myself. For being the tireless, perfectionist researcher that I always envy. From inviting this shy newly-graduate to her first seminar in the her very first day, to pestering non-stop to get me looking for my next step, you never stopped provoking thought, and pushing me to question what I am too lazy to think about.

To Erhan Öztop, one of the rare people that I cannot decide whether his heart or his

mind is brighter. My role model, for not a single day had he drawn a line, imagined a hierarchy between himself and his students. For making me think that all geniuses must be kind hearted. For glimpses of insights you unknowingly bestow. For brilliant days in Japan. Most of all, for your hearty support that I can always count on.

To Angelo Cangelosi, for his immense support when I was away from home. For it is difficult to find such a genuinely kind, thoughtful person. For never forgetting that his students are actual human beings, each with personal problems of their own, putting in every support he can summon and leaving no stone unturned to help them.

To Didem Gökçay, for brilliant comments and future suggestions for this thesis, but also for her determination, bravery, and unyielding stance against all wrong. For never separating herself from her students in her mind. For working for her students more than themselves. Finally, for the admirable wit with which she laughs at all irony.

To Aydan Erkmén for her incredible support, endless energy, for most creative comments, for always seeing a different side in which I have become too familiar to question. With special thanks for her encouragement about the “postpartum” syndrome.

To Billur Barshan for her incessant kindness and thoughtfulness. Lucky are your students to have such an amazingly nice person to guide them.

To Fatih Gökçe, for being my friend all through this sometimes-long journey. Every moment of trouble was lighter through laughing over it. For his unwavering support that I always found in times of trouble, when I felt completely alone. For teaching me what real dedication to good, solid work and attention to detail is. Your place for me as the eternal pillar of support will forever stand. May you be always happy and together with your family in your next adventure.

To Ali Emre Turgut, for teaching me that everything passes, but friends remain; therefore no matter what, we are a friend first. And for teaching me what academic honesty is by setting the best example, that we are humans first, researchers second. Here is to every bad result of mine, acknowledged with pride, in honor of you!

To Osman Tursun, for literally being the joy of the lab. For teaching me that “to want” and “to wish” are totally different things. My little brother, your wisdom extending much further than your years, your passion for life unconquerable, I have no doubt you will be taking yourself and your kindred closer to the horizon. To Güner Orhan, we have gone through all kinds of trouble together, and the good overweighs the bad. To Akif Akkuş, for his immense positiveness. The person so pure at his own heart, that he can never see a fault at anyone. May you always remain so genuine, so heartily happy in the neverending interestingness of life. Selda Eren, Mehmet Durna, Erkin Bahçeci, Maya Çakmak, Onur Soysal, İlkey Atıl, Barış Akgün, Doruk Tunaoğlu, Mustafa Parlaktuna, Yiğit Çalışkan, Kadir Fırat Uyanık, Asil Kaan Bozcuoğlu, Erinç İnci, Sertaç Olgunsoylu, Onur Yürüten, Levent Bayındır, Yaman Çakmakçı, Fariba

Yousefi, Mehmet Çelik, Gaye Topuz, Nihal Tarkan, Alper Karamanlıoğlu, İlker Bozcan, Çağrı Erciyes, Metin Balaban, Barış Özkuşlar, and now Irmak Doğan and Cemal Aker. Troubles diminished until all there were was laughs and joy with you.

To Selma Süloğlu, for incessantly asking what she can do to help. No, she could not make my flu go away, thank her very much, but she did make my life worth having, by being with me whenever I felt sad or angry or disappointed, by sharing most of my worst thoughts and bearing with me, for literally knowing me and loving me all the same. If someone knows the dark side of me truly, this is you. Thank you for being half of me nonetheless. Thank you for making me trust that no matter what happens, there is one constant, one anchor, one safe haven in my life. To Ömer Nebil Yaveroğlu, my eternal friend, source of wisdom, calmness and soothing. There is no limit to the amount of troubles we have brought on the table, discussed, dissected, and beaten. If any self-help guru instructs me to “imagine myself at my safe place”, I should think directly of your room at the department, with the merry sun of Ankara shining in from the windows, your head bouncing happily left and right. They were some of the best days, but here is also to walking together, towards even better ones! To Burçin Sapaz, my crying-wall of ten years, who has never missed stopping everything he was doing to listen to his friends in need. For the support you have given me, countless times you have listened to me, thought of my well-being more than me, but perhaps more importantly for teaching me how to use words, not tears or deadly stares, in solving arguments and settling disputes. I am forever grateful to you for gifting me this new tool, that I understand is called “communication”. To Ayşegül Sapaz, the physical embodiment of the notion of strength, the problem-solver, the unceasing determination. For being practical and level-headed every time I lost my head in the clouds and was in need of solid advice. For forgetting her own problems in trying to calm me down. For seeing the ironies in the world through the same laughing eyes. And to my favorite niece Tanem, with the hopes of many happy years together! To the most loyal person I know, Nilgün Çelik, for she is the only person I know who loves unconditionally, never wavering in her love in the haze of arguments and anger. For showing me what true loyalty is. And also for demonstrating how a person can advance herself every single day, always improving, step by step, and enjoying the process in all sincerity. To Anıl Çelik, for expanding my horizons with many wise discussions and comments. For rising my awareness, making me a better person. To Kübra Kültür, for many a yorgunluk kahvesi. How surprising to find someone so utterly like yourself, after so many years! For sharing two identical minds, for taking the same breaks, for *tsk-tsk*'ing with the same aunty-demeanor, for laughing at the same absurdities. For her creative and completely unnecessary solutions (washing your hand in the dark?) For her determination, strong will, and kind heart. For endless discussions, endless efforts on being better, more mature people. For reminding me that we are never done growing up. To Gökçe Yıldırım Kalkan, the kind soul, who is always on the side of the under-dog. Always supportive, thinking of all possible angles before condemning anyone. From you, I have learned what it



is to be just: Because justness requires hearing everyone's story, giving all a chance. From you I have learned to take your time before you make your decision, because there is always more than one side. To Emre Başeski, for his incredible magic with the cards. Or incredible genius. I'm not sure which. But anyway it was so much fun.

To Elif Sarımay Cenk, for being the reason that I am what I am today, by setting the wise, compassionate, full-of-life example for me to imitate. For holding my hand for who knows how many years. Across the ocean, my heart still beats with yours. To Ekin Eroğul, an enigma, an idol, the very first 'charisma: 4' entity most people have seen, but first and foremost and forever, my friend. For showing me how there is a world bigger than your intimidations. How interesting the wide world really is. For teaching me how one is stronger than their fears, interdependences, doubts. For being one of the strongest, cleverest, most determined, most loyal people I've ever known. To Mari Fukami, my friend, my soul-mate from the other end of the world. Who took my loneliness away and gave me her limitless love instead. Who made Japan my second home, and myself a little Japanese at the end. A modern day *samurai*, who can take on all the struggles of the world with courage, grace, and grit. To dearest Marina Wimmer and Nicolas Pugeault, for making their home my home. Making me a part of their lives. Long I have traveled, and yet I fail to understand how these two warmest, most giving, most loyal hearts could happen to find each other. The loving mom and dad, the no-nonsense scientists, they defy the stereotypes that one is either a "heart-person" or a "mind-person". And to Aurélie, welcome to us! To Frank Guerin, whose heart remains as pure and untarnished in this selfish era as science should ever be. With thanks for many eye-opening conversations, wise guidance, incessant idealism, many laughs, and more importantly reminding me why we are all in this. To Çağla Okutan, my thoughtful friend, for unstoppable support and genuine goodness in her heart, for being with me through thick and thin, where everything else has failed me. To Charu Raghavan, Salomon Ramirez Contla, Paola Massyel Garcia Meneses. Words fail me perhaps the first time. What immense love, what boundless friendship, what an "unbelievable" surprise to have found you when most unexpected. To my co-conspirators, partners in crime, with gratitude. To Anna-Lisa Vollmer and Nikolas Hemion, for their immense kindness, for making me at home in a country so unlike my own. To Lucy Davies, the "crazy Welsh lady" who I will give anything to laugh together with again. My idol in her awareness and care about all the world. For teaching me that there are more things in the world to be concerned about than merely ourselves. To Shaun Lewin, for expanding my mind with his every remark.

To my "family" in the department, and such a family that one can find only once in life. Who made me the person I am, by giving me all their support, by sharing laughs and tears alike. In our cooperations and in our statements, our dreams and our stances alike, we have been together. To Gökdeniz Karadağ, who taught me to laugh loudly in the face of distress (not totally unlike a berserker). Who taught me that life may be hard, but we can still be damn happy. To this eternal geek, loyal friend, dedicated

teacher, lucky-as-a-four-leaf-clover nerd. To Erdal Sivri, the self-professed nihilist who is more widely known for his giant heart. To the ever-stable pillar of support who would literally do anything and everything (including listening to crying girls) for his friends. To Onur Deniz, forever the opposition, forever the philosopher. My too-clever-for-his-own-goodness friend, here is to having our hearts at peace while toiling away on our neverending quests for the right and the just. To Anıl Sınacı, perfectionist, idealist, warrior, in an imperfect, non-ideal, selfish world. As the water flows to find its way, let us take a rest once in a while, and enjoy our little time with loved ones. To Can Eroğul, as kind, as thoughtful as they make it. With his immense wisdom, calming advice, grand heart, the secret big brother of us all. To Umut Eroğul, a huge “Günaydın!”. One of the most pure-hearted people I have ever known, not a single doubt, not a single prejudice, but pure, utter, complete acceptance of who and what you are. In your happy acceptance of all I see the way how the world can be happy, not by perfecting, but simply by accepting. To Çelebi Kocair, the last Mohikan, the last knight, the last samurai. The last memory of an honorable kind that is now dying from the world. In you, I see how it has been in a time less selfish, and still a faint hope for mankind in the future. To Utku Erdoğan, for incredible times. For unforgettable “knowledge” on everyone, for comments so true to deserve being quoted. For such wisdom that I’m not sure how you have collected. To Özgür Kaya, for being the “father” of us all, with the knowledge that anyone bothers us, he will “take care” of them for us. For being the strong rock that we all depend on. For and stopping everything to help you in times of need. To Merve Aydınlılar for her incessant support, kindest smile, great wisdom, and to Kerem Hadımlı and Can Hoşgör, for brightening many dark days by making me laugh all the time. To Serdar Çifçi for his forever thoughtfulness and kindness. To Müge Sevinç, Orkan Bayer, Alev Mutlu, Levent Eksert, Okan Tarhan Tursun, Sinem Demirci, Dilek Önal, Fatih Titrek, Hüsnü Yıldız, Hilal Kılıç, Ayşegül Yaman, Mine Yoldaş, Özlem Erdaş, Aslı Gençtav for their companionship in the most troubled times. To Erkut and Aykut Erdem, for their unyielding loyalty and boundless friendship. To the loving, compassionate Deniz Karadağ, for endless laughs over the weirdest videos, for her immense kindness and patience with kids who had so many troubles, and in particular, for sharing my lunacy and giving me the relief that I’m not a lunatic alone, at least.

To my family, with so much gratitude that is it meaningless to try to describe. To Mom and Dad, to my Grannies, for loving me so, for being the wisest people I have ever known, for giving me all my principles, wits, and love I possess today. To my Uncles, for loving me more than I thought possible for us humans. To my Cousins or daughters, for I forget which one is true sometimes, for loving me unspeakably and also giving me the endless worries for them that will never leave me. Whatever I do, it is yours. Whatever I love, it is because they are like you.

I also thank The Scientific and Technological Research Council of Turkey (TÜBİTAK) for funding this work by the project no 111E287, and the graduate fellowship no 2211.

# TABLE OF CONTENTS

ABSTRACT . . . . .	v
ÖZ . . . . .	vii
ACKNOWLEDGMENTS . . . . .	x
TABLE OF CONTENTS . . . . .	xv
LIST OF TABLES . . . . .	xx
LIST OF FIGURES . . . . .	xxiii
LIST OF ALGORITHMS . . . . .	xxx
LIST OF ABBREVIATIONS . . . . .	xxxi

## CHAPTERS

1	INTRODUCTION . . . . .	1
1.1	Contributions of this Thesis . . . . .	3
1.2	Outline of the Thesis . . . . .	6
2	ON CONCEPTUALIZATION AND CONTEXT . . . . .	7
2.1	Concepts . . . . .	7
2.1.1	Theories of Concepts . . . . .	8
2.1.2	Grounding Concepts . . . . .	10

2.1.3	Spatial Concepts . . . . .	13
2.2	Structural Representation of Concepts . . . . .	19
2.2.1	Structural Representation in Humans . . . . .	19
2.2.2	Structural Representation in Robotics . . . . .	26
2.3	Context . . . . .	29
2.3.1	Context in Robotics . . . . .	30
2.3.2	Context in Related Fields . . . . .	33
2.3.3	Structural Representation of Context . . . . .	34
3	THE EXPERIMENTAL SETUP . . . . .	37
3.1	Object Set . . . . .	38
3.2	Behaviors . . . . .	39
3.3	Features and Data Collection . . . . .	40
4	THE CONCEPT WEB . . . . .	45
4.1	Individual Concepts . . . . .	45
4.1.1	Conceptualization of a Category . . . . .	45
4.1.2	Category Prediction from Features Only . . . . .	48
4.2	A Probabilistic Web of Concepts . . . . .	51
4.2.1	Building a Web from Individual Concepts . . . . .	51
4.2.2	Concept Web as a Markov Random Field . . . . .	51
4.2.3	Belief Propagation in MRF . . . . .	53
4.2.4	Inferences in Concept Web Using Loopy Belief Propagation . . . . .	55

4.3	Experiments and Results . . . . .	57
4.3.1	Scenario 1: Perception-driven activation of concepts in the web . . . . .	57
4.3.2	Scenario 2: Interaction-driven activation of concepts in the web . . . . .	59
4.3.3	Scenario 3: Action-driven activation of concepts in the web . . . . .	62
4.4	Summary . . . . .	63
5	SPATIAL CONCEPTS AND THE REPRESENTATION OF WHOLE SCENES . . . . .	67
5.1	Data Collection and Feature Extraction . . . . .	68
5.2	Representing Spatial Concepts with Prototypes . . . . .	69
5.3	Hybrid Markov Random Field . . . . .	69
5.4	Scene Representation . . . . .	71
5.5	Experiments and Results . . . . .	72
5.5.1	Scene Interpretation . . . . .	72
5.5.2	Correcting Wrong Interpretations . . . . .	73
5.5.3	Human-Robot Interaction . . . . .	74
5.6	Summary . . . . .	76
6	A FORMALISM AND COMPUTATIONAL MODEL FOR CONTEXT . . . . .	77
6.1	A Formalism of Context . . . . .	78
6.2	Latent Dirichlet Allocation . . . . .	81
6.3	Modeling Contextual Information with LDA . . . . .	83

6.4	An Incremental and Online Version: Incremental-LDA . . . .	83
6.5	LDA versus the Requirements of Contextual Modeling . . . .	86
6.6	Making Use of Context: Feeding the Contextual Information back to the Concept Web . . . . .	87
6.7	Entropy-Based Evaluation of the System . . . . .	88
6.8	Experiments and Results . . . . .	88
6.8.1	Performance of Incremental-LDA and K-Incremental Gibbs Sampling . . . . .	90
6.8.2	Context from the Concept Web against Context from Raw Features . . . . .	95
6.8.3	Using Context, Part 1: Making Sense of Pure- and Mixed-Context Environments . . . . .	96
6.8.4	Using Context, Part 2: Object Recognition in Con- text . . . . .	96
6.8.5	Using Context, Part 3: Planning in Context . . . .	100
6.8.6	Running Time Performance of the System . . . . .	104
6.9	Summary . . . . .	104
7	CONCLUSION . . . . .	107
7.1	Discussion . . . . .	108
7.1.1	A Web of Concepts . . . . .	109
7.1.2	Spatial Concepts as Nodes . . . . .	109
7.1.3	Basing Context on the Concept Web . . . . .	110
7.1.4	Lifelong and Developmental Learning in Robots .	111
7.1.5	Planning in the Real World . . . . .	111

7.2	Limitations and Future Work . . . . .	112
	REFERENCES . . . . .	117
	CURRICULUM VITAE . . . . .	137

## LIST OF TABLES

### TABLES

Table 3.1 The frequencies of instances in the dataset in which specified noun and adjective pairs co-occur together (out of 60 objects in the dataset). . . .	39
Table 3.2 Possible applicable set of behaviors with respect to object categories. $X \in \{Left, Right, Forward, Backward\}$ ; A: <i>Applicable</i> ; N/A: <i>Not-Applicable</i> . . . . .	40
Table 3.3 The audio, haptic and visual features extracted from the interactions of the robot. . . . .	40
Table 4.1 Extracted prototypes for noun, adjective and verb concepts. [Adapted from [6] ©2015 IEEE.] . . . . .	49
Table 4.2 A comparison of noun and adjective predictions using the concept web. One object from each noun category is used for demonstration. Images depict RGB-D images from the Kinect sensor. Columns 2 and 3: SVM predictions, Columns 4 and 5: ReliefF feature selection + SVM, Columns 6 and 7: Prototype-only predictions, Columns 8 and 9: Concept web estimations. Parentheses: Prediction confidences. Bold text: Correct decisions. Stroked text: Wrong decisions. [Adapted from [6] ©2015 IEEE.] . . . . .	60
Table 4.3 The predictions as corrected by the activation on the concept web, when there is no direct perceptual access to certain features of the object. The iCub is not allowed to grasp the ball object, and therefore makes initial predictions using only the available visual features (columns 2 and 3). The visual parts of the concept prototypes ( <i>i.e.</i> , features [1-67]) are used for this comparison. This initial activations are then allowed to spread on the concept web, which converges to the significantly more accurate a posteriori predictions displayed in columns 4 and 5. The haptic and audio predictions are corrected through the spreading of activation. Columns 8 and 9: Concept web estimations. Parentheses: Prediction confidences. Bold text: Correct decisions. Stroked text: Wrong decisions. [Adapted from [6] ©2015 IEEE.] . . . . .	62



Table 4.4	The selection of objects to which sample commands are applicable. The selection is performed by the spreading activation on the web, which disperses to the related verb concepts as well. <i>throw</i> verb concept activates selectively non-cup and non-plate objects. Selection confidences are indicated in parentheses. Images depict RGB-D images from the Kinect sensor. [Adapted from [6] ©2015 IEEE. Best viewed in color.] . . . . .	64
Table 5.1	The features used for extracting spatial concepts, including (1) binary projective features extracted from two objects in relation to each other, and (2) the individual visual features of the two objects. [Adapted from [67] ©2015 IEEE.] . . . . .	69
Table 5.2	Prototypes extracted for spatial concepts. [Adapted from [67] ©2015 IEEE.] . . . . .	70
Table 5.3	A sample scenario of scene interpretation. Some of the extracted relations for the presented 3D view are indicated. [Adapted from [67] ©2015 IEEE. Best viewed in color.] . . . . .	73
Table 5.4	A sample scenario of correcting wrong interpretations of spatial relations. Prediction confidences with and without concept web are indicated. The spatial estimation of ball A is corrected through the concept web. Bold text: Correct decisions. Stroked text: Wrong decisions. [Adapted from [67] ©2015 IEEE. Best viewed in color.] . . . . .	74
Table 5.5	Accuracies of concept web estimations of the noun concepts and spatial relations in the scene, and the accuracies of spatial-direction based human-robot communication. [Adapted from [67] ©2015 IEEE. Best viewed in color.] . . . . .	74
Table 5.6	A sample scenario of human-robot interaction based on spatial-directions. Objects found by iCub in response to sample queries on the given 3D scene are indicated. [Adapted from [67] ©2015 IEEE. Best viewed in color.] . . . . .	75
Table 6.1	The correspondence between the LDA terms and the notation used in this work. [Adapted from [195] ©2015 IEEE.] . . . . .	83
Table 6.2	Prediction confidences are indicated in parentheses. Bold text indicates correct decisions whereas stroked text indicates wrong decisions. [Adapted from [195] ©2015 IEEE. Best viewed in color.] . . . . .	97

Table 6.3 Object recognition in context. Prediction confidences are indicated in parentheses. Bold text indicates correct decisions whereas stroked text indicates wrong decisions. [Adapted from [195] ©2015 IEEE. Best viewed in color.] . . . . . 98

# LIST OF FIGURES

## FIGURES

Figure 1.1 (a) Existing cognitive systems have concepts which have links to perceptual features and motor actions which were programmed by a designer or trained in context-free environments. (b) The concept web model without context: A densely connected concept web connecting perception, action and language; however, there is no notion of context in this model. (c) We propose a system that learns in context the links between concepts and sensorimotor primitives, based on the statistics of its interactions in real-life environments. For clarity, only a few links and concepts are shown. . . . . 3

Figure 2.1 The “connected” concepts representation we propose in this thesis. Semantically related concepts are linked together with connections. They are also grounded in perception, action, and language. White color indicates “active” concepts, blue color indicates inactive ones, see Chapter 4 for details. Adapted from [6] ©2015 IEEE. . . . . 12

Figure 2.2 The interference of perceived functionality on the interpretation of the spatial prepositions, experiments by Coventry *et al.* [45]. When the object with a canonical functional usage, *i.e.*, the umbrella that is supposed to protect the man from the rain, is in a position that is able to carry out this functionality, people are more likely to say that the man is *under* the umbrella (middle row). On the contrary, when the object is perceived as ineffective to perform its functionality (bottom row), people do not judge the man as under the umbrella anymore, but rather, as *below* it. Figure taken from [45]. . . . . 15

Figure 2.3 The interference of everyday knowledge with the interpretation of spatial relations, experiments by Coventry *et al.* [46, 47]. (a) Although the object is at the same location with respect to the bowl in both figures, people are likely to judge it *in* the bowl only in the left figure, where the object indirectly “touches” the bowl via the other objects in between. (b) Although both the bowl (left) and the jug (right) are full to the same amount, people are likely to judge the oranges on the left as *in* the bowl, while the apples on the right as *out* of the jug. Coventry *et al.* hypothesize that this difference is due to the canonical usage of the jugs to hold liquids, which can then physically be filled only until the rim. The bowl, on the other hand, is commonly used to hold solid objects. Therefore, people have in their minds a larger canonical containment range for the bowl, as compared to the jug (c). Figures taken from [47]. . . . . 16

Figure 2.4 The representation of spatial concepts in our system, as bona fide concepts themselves, that connect other noun, adjective, and verb concepts but are also flexible and subject to refining and reconsideration through the integration of knowledge coming from neighboring concepts. White color indicates “active” concepts, see Chapter 4. Adapted from [67] ©2015 IEEE. . . . . 18

Figure 2.5 The somatotopic activation of the premotor cortex to the words “pick”, “lick”, and “kick”, compared to the motor areas that actually activate while performing picking (hand-related), kicking (foot-related), and licking (mouth-related) actions. Figure taken from [75]. . . . . 21

Figure 2.6 The single neuron that fires selectively to Luke Skywalker, whether presented visually, as a text string, or in audio modality, while also firing for a picture of Yoda [101]. This neuron is therefore showing selective activation to the encapsulating Star Wars concept, regardless of the specific modality or detail of the presented cue, providing support for the single concept cells hypothesis. Figure taken from [101]. . . . . 22

Figure 2.7 The KNOWROB system proposed by Tenorth and Beetz [118]. Everything in the world, be it an object, action, or event, is a part of a (tree-structured) ontology, and derives from the most general class of “Thing”. Figure taken from [118]. . . . . 27

Figure 2.8 The associative web of conceptual memory proposed by Baxter *et al.* [123]. Different modalities of the same concept are connected together with associative links. The information about which modalities are included in a concept are learned from the zoo database of UCI Machine Learning Repository . Figure taken from Baxter *et al.* [123] . . . . . 28

Figure 2.9	The effect on visual context on object recognition. Note how two visually identical items can be judged as both a hairdryer (on the left) and a drill (on the right). Figure taken from [5]. Best viewed in color. . . . .	34
Figure 3.1	Experimental setup including the iCub robot platform and the Kinect RGB-D depth camera. [Adapted from [195] ©2015 IEEE.] . . . . .	37
Figure 3.2	The noun concepts used in this study, and the associated objects. [Adapted from [6] ©2015 IEEE.] . . . . .	38
Figure 3.3	The adjective concepts used in this study, and the associated objects. [Adapted from [6] ©2015 IEEE.] . . . . .	39
Figure 3.4	Extraction of entity features and effect visual features. $\mathbf{e}_v$ and $\mathbf{e}'_v$ are the visual features of an object before and after a behavior is applied. $\mathbf{f} = \mathbf{e}'_v - \mathbf{e}_v$ is the effect visual feature. $\mathbf{e}$ is the multi-modal feature incorporating visual, haptic, proprioceptive, and audio information. [Adapted from [6] ©2015 IEEE.] . . . . .	42
Figure 4.1	The schematic presentation of the concept web, which connected related concepts to each other and to their counterparts in the language, action, and perception spaces. Information can flow in from the perception space, through a feature extraction mid-level, or from the language and action spaces as well. A number of nodes are randomly illustrated with white color to exemplify active concepts. [Adapted from [6] ©2015 IEEE.]	46
Figure 4.2	Schematic visualization of the extraction of a concept prototype. If a feature has a consistently high contribution, marked with a high mean and low variance distribution, it is indicated with a '+' sign. Those with a consistently low contribution, marked with a low mean and low variance distribution, are assigned a '-' sign, whereas those with a high variance are marked with a '*' to indicate inconsistent contribution. Sample features are illustrated for the <i>hard</i> concept. [Adapted from [6] ©2015 IEEE.] . . .	50
Figure 4.3	A snapshot of the concept web iCub has constructed. Connections between related concepts are denoted with gray links. Noun concepts are indicated with red, adjective concepts with blue, verb concepts with green, and superordinate concepts with cyan. The graph is created using Ubigraph graph visualization library [199]. [Adapted from [6] ©2015 IEEE. Best viewed in color.] . . . . .	52

Figure 4.4 (a) A sample 2D Markov Random Field. The Markovian property holds in Markov Random Fields, by which a random variable (*i.e.*, the black node), given its immediate neighbors (the gray nodes), is independent of all other random variables. (b) A maximal clique (with 3 nodes) is indicated in an MRF with 5 nodes. . . . . 53

Figure 4.5 A schematic representation of MRF modeling of the concept web. Initial predictions about the concepts are used to initialize concept node probability values. Conformance to initially predicted values are maintained by minimizing the sum of unary potential functions  $\psi_C$ . Meanwhile, clique potentials are initialized from the cooccurrence information from the training data, and conformance to the cooccurrence information is maintained through minimizing the sum of clique potentials  $\psi_K$ . [Adapted from [6] ©2015 IEEE.] . . . . . 54

Figure 4.6 Sample Markov Random Field chain of variable nodes . . . . . 54

Figure 4.7 The conversion of a MRF graph as a factor graph, as input to the Loopy Belief Propagation. [Adapted from [6] ©2015 IEEE.] . . . . . 55

Figure 4.8 Divided subtrees of the graph in Figure 4.7. [Adapted from [6] ©2015 IEEE.] . . . . . 55

Figure 4.9 Schematic presentation of Scenario 1. iCub is presented with a cup and expected to predict the type and properties of the object, as well as what kind of behaviors can be applied on this object. ML: *Move Left*, MR: *Move Right*, MB: *Move Backward*, MF: *Move Forward*, PL: *Push Left*, PR: *Push Right*, PB: *Push Backward*, PF: *Push Forward*. [Adapted from [6] ©2015 IEEE. Best viewed in color.] . . . . . 57

Figure 4.10 Schematic presentation of Scenario 2. iCub is presented with a ball, and is expected to guess its properties, as well as applying the *push* behavior on it. It predicts the category and the properties of the object correctly, even though it is not allowed to touch the object beforehand, and is therefore unaware of its proprioceptive, haptic (*soft*), and auditory (*silent*) properties beforehand. PL: *Push Left*, PR: *Push Right*, PB: *Push Backward*, PF: *Push Forward*. [Adapted from [6] ©2015 IEEE. Best viewed in color.] . . . . . 61

Figure 4.11 Schematic representation of Scenario 3. The sample *box*, *cup*, and *plate* objects are given to the system and *knock down* behavior is commanded to iCub. iCub selects any one of these objects if the commanded behavior is applicable. In this scenario, the *box* object is selected and its activated concepts are shown. The action space and verb concepts are contoured with green, whereas blue and orange colors represent the noun and adjective categories for the object, respectively. The gray and smaller fonts show inactive concepts in the web, while bigger fonts and colored nodes represent activated concepts. There are other concepts that are not shown for clarity. [Adapted from [6] ©2015 IEEE. Best viewed in color.] 63

Figure 5.1 The concept web combining concept web instantiations of perceived objects, and their spatial relations. Shaded areas correspond to concept webs of the individual objects. These are fed by the extracted features of the objects, as well as by the language and action planes. The spatial relations between the objects combine these individual object representations, and are fed by the relative and the individual features of the two objects, and again by the language and action planes. [Adapted from [67] ©2015 IEEE.] . . . . . 68

Figure 5.2 Extraction of directed cliques in a hybrid Markov Random Field, and which is converted into a factor graph with two clique nodes. [Adapted from [67] ©2015 IEEE.] . . . . . 71

Figure 6.1 Average log likelihood  $\hat{l}$  for varying  $\sigma$  (Equation 6.11,  $\sigma = 0$ : Pure contextual information,  $\sigma = 1$ : Pure concept web decision). The interval  $[0.4, 0.5]$  is depicted as maximizing  $\hat{l}$ . [Adapted from [195] ©2015 IEEE.] 89

Figure 6.2 A comparison of the entropy ( $\tilde{H}$ ) evolution (Equation 6.13) of K-Incremental Gibbs solver, versus the standard batch Gibbs solver. The K-Incremental Gibbs solver is fed a partial solution for 2 contexts and then run for  $K = 3$  contexts. The batch Gibbs sampler is directly run for  $K = 3$  contexts. [Adapted from [195] ©2015 IEEE.] . . . . . 90

Figure 6.3 The effect of encountered scene counts and varying context counts  $K$ . Note that Incremental-LDA would itself stop at  $K = 3$ , however we force increasing  $K$  for the sake of comparison. (a) Effect of increasing  $K$  on the number of uncertain concepts,  $|\mathbb{C}_{low}|$ , for varying number of scenes. By  $K = 3$  contexts  $|\mathbb{C}_{low}|$  diminishes to 0, therefore Incremental-LDA would stop adding new contexts at this point. (b) Effect of encountered scenes on the number of uncertain concepts,  $|\mathbb{C}_{low}|$ , for different context counts. (c) Effect of increasing  $K$  on the entropy of the system,  $\tilde{H}$ , for varying number of scenes. (d) Effect of encountered scenes on the entropy of the system,  $\tilde{H}$ , for different context counts. In all the experiments, 10 test sets of  $|\mathbb{D}|$  scenes each are used. The mean values for the 10 test sets are plotted, while the standard deviations are indicated with error bars. In (b) and (d), the  $x$ -axis is in log-scale. [Adapted from [195] ©2015 IEEE. Best viewed in color.] . . . . . 92

Figure 6.4 The performances of LDA over raw features only, versus of LDA over MRF-based concept web, presented as prediction accuracies scaled to [0,1]. Presented values are the predicted likelihoods of “correct contexts” in each corresponding case. For evaluation, the ground truth data of the expected contexts were extracted via supervision.  $\alpha$  and  $\xi$  are the trade-off parameters from Equation 6.10. (a) Using only the raw features as input to LDA, for varying discretization bin counts and increasing numbers of encountered scenes ( $\alpha = 0.1, \xi = 0.1$ ). (b) Using only the raw features as input to LDA, for varying settings of  $\alpha$  and  $\xi$  (50 scenes, 10 bins). (c) Using the concept web as input to LDA, for increasing numbers of encountered scenes. ( $\alpha = 0.1, \xi = 0.1$ . Discretization is not necessary, therefore the result vector is 1-dimensional.) (d) Using the concept web as input to LDA, for varying settings of  $\alpha$  and  $\xi$  (50 scenes). [Adapted from [195] ©2015 IEEE. Best viewed in color.] . . . . . 94

Figure 6.5 The combined results of object recognition in context, over all 15 objects in the test set. The prediction accuracies over all determined noun and adjective concepts, using (i) only perceptual features, (ii) the concept web, and (iii) contextual information are compared. In the plot, the red lines denote the median values, the boxes denote the data that fall between the 25<sup>th</sup> and 75<sup>th</sup> percentiles, the whiskers cover the extreme data that are not outliers, and stars indicate the outliers. [Adapted from [195] ©2015 IEEE.] . . . . . 99



Figure 6.6 The performance of the individual prototype-based predictions, versus context enhanced concept web predictions, under artificially added noise, presented as prediction accuracies scaled to [0,1]. The noise probability denotes the probability of artificial noise being added to each single concept, via reversing its prototype-predicted probability from  $p\%$  to *reversed* to  $(100 - p)\%$ .  $\sigma$  refers to the trade-off parameter in Equation 6.11. [Adapted from [195] ©2015 IEEE. Best viewed in color.] . . . . . 99

Figure 6.7 Pruning of forward planning trees by integrating contextual information. (a) iCub’s workspace. (b) First planning scenario. iCub is expected to move a cup from position 8 to position 5. Since pushing and knocking actions are dangerous in the kitchen context, these nodes are pruned without further expansion. Pruned branches are indicated with crosses. (c) Second scenario. iCub must bring a ball from position 7 to 1. Pushes are pruned, since pushing a ball causes it to roll down from the table. PX: Push left/right/forward/backward, MX: Move left/right/forward/backward, KD: Knock down, SH: Shake, TH: Throw, DP: Drop, G: Grasp. [Adapted from [195] ©2015 IEEE.] . . . . . 101

Figure 6.8 The node counts of *unpruned* vs. *pruned* planning trees of 10000 random scenarios, grouped by their contexts. The Kitchen context is subject to more pruning, as expected, due to a large number of *NA* behaviors. The Workshop context, on the other hand, is not subject to any pruning, since all behaviors are potentially applicable. In the plot, the boxes denote the data that fall between the 25<sup>th</sup> and 75<sup>th</sup> percentiles, and stars indicate the outliers. [Adapted from [195] ©2015 IEEE.] . . . . . 103

## LIST OF ALGORITHMS

### ALGORITHMS

Algorithm 1	Derivation of a prototype from the exemplars of a category [Adapted from [6,38].]	47
Algorithm 2	Batch Gibbs sampling algorithm. [Adapted from [3,210].]	82
Algorithm 3	The proposed Incremental-LDA algorithm. [Adapted from [195] ©2015 IEEE.]	85
Algorithm 4	The K-Incremental Gibbs sampling approach we propose as a companion to Incremental-LDA. [Adapted from [195] ©2015 IEEE.]	85
Algorithm 5	Breadth-first forward planning with context-dependent prun- ing. [Adapted from [195] ©2015 IEEE.]	102

## **LIST OF ABBREVIATIONS**

LBP	Loopy Belief Propagation
LDA	Latent Dirichlet Allocation
MFCC	Mel-Frequency Cepstral Coefficients
MRF	Markov Random Field
PCL	Point Cloud Library
RGNG	Robust Growing Neural Gas
SVM	Support Vector Machine



# CHAPTER 1

## INTRODUCTION

An “exchange of ideas” with a 5-year-old is always interesting. This little kid, who was not in this world only a couple of years ago, not only understands you perfectly, but he is also sure to surprise you with one or two very well aimed comments on life, truth, or yourself in particular. What is it that these little creatures are capable of, and yet our most “artificially intelligent” robots are clueless at? How do they progress from “Hey what is this? It is a plane!” to logical talk, commentary, irony, and intentional lying? Somehow they not only soak up a plethora of new concepts every day, but also rapidly form countless connections between them - until eventually they have a perfect world model in their minds that identically replicates the rules and relations of the outer world. No later than they meet with a new concept, they connect it with what they already know, which will, in the long term, give rise to context in their minds, and prevent them, for instance, from swearing in the presence of their mother, unless of course in case of dire emergency.

We learn concepts, as well as the contexts that they frequently occur in, very early in life, and mostly do not consciously realize our dependence on our ability to conceptualize, or contextualize, any more. Yet, we depend on these internalized contexts for numerous things, for understanding which categories a novel object falls into, for planning how we can perform a certain goal with it, for reasoning how it fits within our world. Concepts even shape our language by guiding how we think: Nouns occur because they represent groups that are important enough in life for us to attend to, adjectives lead us to care about some properties of objects, such as being yellow or shiny, more than being made-up-of-a-single-type-of-atoms or not. By

deciding what is important and what is not, concepts also drive our categorization, which Lakoff emphasized as, very elegantly, in his paper titled “*Women, Fire, and Dangerous Things*” [1];

“Categorization is not a matter to be taken lightly. There is nothing more basic than categorization to our thought, perception, action and speech [...] Whenever we intentionally perform any kind of action, say something as mundane as writing with a pencil, hammering with a hammer, or ironing clothes, we are using categories [...] Without the ability to categorize, we could not function at all.”

There are no doubt many facets of conceptualization as a cognitive facility, which possibly include the forming of new concepts in our minds, the statistically and probably-also-socially guided decision of what comprise a distinct concept and what do not, how these concepts are structurally represented in our minds, how we place a first-time seen object among the countless concepts we have acquired for years, how concepts are held in relation to each other, how we learn about and utilize these relations, and how these relations lead the emergence of different contexts in our life with time. The questions are endless, but so are also the possibilities: We stand to enormous gain, theoretically and practically, of the study of such a comprehensive subject.

In this thesis, we make an initial attempt at a formalization of concepts and context in robots. Concepts, for sure, have been studied implicitly or explicitly numerous times before, in robotics or elsewhere, but we share the viewpoint of Deacon [2], and place the *relatedness* of concepts to the very center of our formalization. Indeed, he claims, primates are able to learn certain concepts as well, but humans are unique in learning them in relation to each other, connecting every symbol in a great map of symbols, being therefore able to manipulate them at will, and truly utilize their power. Context, on the other hand, to the best of our knowledge, in spite of being intuitively one of the cornerstones of cognition, has never been tackled formally and with adequate generality before in a robotics work. It has been worked on, for instance, in Natural Language Processing domain for making sense of text documents (*e.g.*, [3]), in psychology for explaining automated behavior (*e.g.*, [4]), or in computer vision for

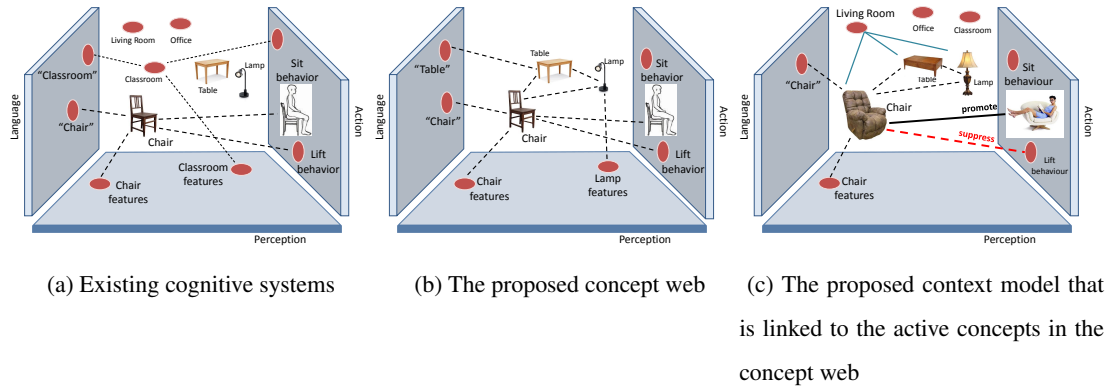


Figure 1.1: (a) Existing cognitive systems have concepts which have links to perceptual features and motor actions which were programmed by a designer or trained in context-free environments. (b) The concept web model without context: A densely connected concept web connecting perception, action and language; however, there is no notion of context in this model. (c) We propose a system that learns in context the links between concepts and sensorimotor primitives, based on the statistics of its interactions in real-life environments. For clarity, only a few links and concepts are shown.

facilitating object recognition [5], but it has not been shown if and how a robot can grow developmentally while at the same time acquiring and utilizing different contexts, much like a human child would do. In this thesis, we make the first systematic attempt at such a robot with a formal understanding of grounded and contextualized concepts represented as a densely connected web.

## 1.1 Contributions of this Thesis

Our contributions in this thesis can be collected under two general headers:

- **Development of a densely connected web of concepts**

- We propose a novel densely connected web representation of concepts, which can exploit relatedness of different concepts to facilitate reasoning. The idea of grounding concepts is well-researched in the literature (Figure 1.1a), however, the proposal of grounding them *together and in relation*

*to each other* is novel to our work (Figure 1.1b). We use a Markov Random Field as the basis of our representation. This model is able to use statistically gathered a priori information about co-occurrences of concepts in order to correct its initially naive interpretations of objects. **This part of the contribution has been done in collaboration with Güner Orhan. [6]**

- We postulate that spatial relations, corresponding to prepositions in the language, need to be represented as *bona fide* concepts themselves in our web, and be considered similarly in relation to the other concepts. For representing the unidirectional, ordered nature of the spatial relations, we propose a hybrid variant of the standard Markov Random Field model.
- Taking advantage of the addition of spatial concepts into the framework, we show how whole scenes can be represented in the system, through activating separate instantiations of the concept web for each object, which are then connected to each other via the spatial concepts. Since our concept web is associative in nature, we need such a distinction of *instantiations* for effective representation of possibly-semantically-conflicting objects present in the scene at the same time.

- **Formalization of context over this web of concepts**

- On this connected web of concepts, we develop a novel framework for contextual interpretation of the scenes (Figure 1.1c). The robot is able to deduce the contexts of scenes, and of individual objects existing in these scenes, using statistical properties of the scenes it has encountered in its lifetime. We propose a novel, online extension to the Latent Dirichlet Allocation methodology, so that the robot is able to learn detecting the contexts developmentally.
- We provide an explicit formalization for context. To the best of our knowledge, this is the first time that context is tackled with systematically, or modeled *per se*, as a separate entity but also in direct relation with other conceptual entities, in a robotics scenario.
- The robot is then able to feedback this contextual information to the more primitive concept web layer, thereby guiding the reasoning in the concept



web according to context. It is also able shape and prune its planning according to contextual information, achieving significant computational efficiency.

These contributions have been disseminated as the following papers:

**Journal papers:**

- **Hande Çelikkanat**, Güner Orhan, Nicolas Pugeault, Frank Guerin, Erol Şahin, and Sinan Kalkan, *Learning Context on a Humanoid Robot using Incremental Latent Dirichlet Allocation*, accepted for publication by IEEE Transactions on Autonomous Mental Development, 2015.
- **Hande Çelikkanat**, Güner Orhan, and Sinan Kalkan, *A Probabilistic Web of Concepts on a Humanoid Robot*, IEEE Transactions on Autonomous Mental Development, vol: 7, no: 2, pp.92-106, 2015.

**Conference papers:**

- **Hande Çelikkanat**, Erol Şahin, and Sinan Kalkan, *Integrating Spatial Concepts into a Probabilistic Concept Web*, IEEE International Conference on Advanced Robotics (ICAR), 2015.
- **Hande Çelikkanat**, Güner Orhan, Erol Şahin, and Sinan Kalkan, *İnsansı Robotlar için Olasılıksal Bir Kavram Ağı*, Türkiye Robotbilim Konferansı (TORK), 2015 (submitted).
- **Hande Çelikkanat**, Güner Orhan, Nicolas Pugeault, Frank Guerin, Erol Şahin, and Sinan Kalkan, *Learning and Using Context on a Humanoid Robot Using Latent Dirichlet Allocation*, IEEE Joint Conference on Development and Learning and on Epigenetic Robotics (ICDL-Epirob), pp.201-207, 2014.
- **Hande Çelikkanat**, Güner Orhan, Nicolas Pugeault, Frank Guerin, Erol Şahin, and Sinan Kalkan, *İnsansı Robotlarda Bağlamın Öğrenilmesi*, Türkiye Robotbilim Konferansı (TORK), 2014.

## 1.2 Outline of the Thesis

The outline of the thesis is as follows:

Chapter 2 presents the literature and state-of-the-art knowledge on concepts and context, in terms of their structural representation in humans as well as in other robotics and computational studies.

Chapter 3 introduces the iCub humanoid robot platform that we work on. This section also details our data set, the specific concepts that we use in this work, the process of data collection, and the representation of the feature vectors of the objects.

Chapter 4 introduces our concept web formulation, explains how it is learned from the training set, and inference is later conducted on it, and presents results that demonstrate the superiority of such a connected approach as compared to an individual-concepts view.

Chapter 5 advances this concept web by proposing a holistic scene representation, via instantiations of the concept web for individual objects, and spatial concepts connecting them.

Chapter 6 builds the idea of context on top of the concept web, describes an unsupervised method to extract contextual information, and shows how this information can later be fed back to the concept web to guide reasoning. This chapter also presents performance increase results in object recognition, scene interpretation, and planning scenarios when under contextual guidance. Moreover, the benefits of using the concept web formulation as the basis are investigated.

Chapter 7 discusses various important features and design choices within the proposed model, and concludes with ideas for future work.

## CHAPTER 2

### ON CONCEPTUALIZATION AND CONTEXT

We begin this chapter by surveying the literature on *conceptualization*. This is a very broad subject, indeed one of the pillars of rational thought, and therefore has been studied from various points of views. From a vast literature, we try to summarize the works that are most relevant to our approach, combining the rich background of psychological, neurological, computational, and robotics works together. We follow this discussion with the idea of *context*, talking about its implications in different areas of life and science. Finally, we summarize our contributions in this thesis, in relation to the existing findings and the literature.

#### 2.1 Concepts

A concept is “a representation that allow us to make sense of the world by enabling us to categorize the continuous high-dimensional sensorimotor space” [6]. How does this representation first form in our minds? How can the formation of one concept facilitate the formation of *other* concepts, once it is securely internalized? More technically, how is it, structurally, *represented* in our minds? If we were to understand the exact structural mechanisms of its representations in humans, then perhaps we could try to mimic a similar representation for our robots, try to be inspired from it for the sake of obtaining more robust, more flexible reasoning in artificial minds. This section presents the current literature in philosophy, psychology, neuroscience, and robotics, that tries to provide some answers, albeit in progress rather than definite and finalized, to these questions.

### 2.1.1 Theories of Concepts

Given the fundamental and central place of concepts in our thought, maybe it is not surprising to trace back the questions on the nature of conceptualization to the Ancient Greeks. Indeed, in the dawn of philosophy, we see the explicit efforts to place the roots, development, and mechanisms of conceptual thought in the human mind.

In Plato's dialogues, we see Socrates arguing that knowledge comes from inside, rather than from outside. In this sense, he claims that we *remember*, rather than *learn* knowledge, that it is *divine*, not humane. Plato builds on this view in his thesis of *ideas*: That the world we live in is not the real one, but an image of a perfect world, a world of "ideas". Human souls, which originally belonged to this perfect world, still remember it and yearn for it. When they see a bird, or a tree, they recognize it because these things remind them of the perfect bird, the perfect tree. Therefore there is no learning, just remembering, with a sense of nostalgia, of archetypes or concepts, too absolute and ideal that everyday instances of them can never duplicate, but only weakly imitate.

Aristotle, on the other hand, firmly protests against this idea. An experimentalist to the core, he argues that it is not that we are born with some ideas of concepts in our minds. It is the opposite, we are born with empty minds, and form our concepts as we encounter their exemplars one by one: When we see a couple of trees, we form a rough idea of a tree; when we encounter a thousand, we then have a very solid tree theory. He also forms a Hylomorphism theory, in which he divides the forms of *substances* into substantial and accidental forms. Substantial forms are the essential properties, *i.e.*, essences, of a substance, they define what makes the substance. The accidental forms, on the other hand, are non-vital, those can change and show variance, without causing the substance to become another substance.

The discussion was naturally not limited to ancient Greece. Many scientists have likewise pondered on the elusive origin and the exact representation of concepts in our minds. The following main theories have thus emerged as a result of centuries of philosophical and empirical work:

- **The Classical (Rule-based) View:** According to the classical view, each con-

cept is describable with a strict rule. Memberships of instances are decided according to their being compatible or not with this rule, resulting in a crisp *yes* or *no* answer (see, *e.g.*, [7]). The following, for instance, may be used as a rule for defining the concept BIRD:

$$\text{has-wings}(obj) \wedge \text{flies}(obj) \wedge \text{lays-eggs}(obj) \wedge \text{has-beak}(obj) \wedge \dots \quad (2.1)$$

Sparrow obviously satisfies this rule, and is therefore an instance of the BIRD concept. However, this view is now generally accepted to have certain shortcomings, the most prominent of which being its inability to distinguish between *typical* and *non-typical* instances of a concept [8]. Take for instance, a very non-typical bird, the penguin, which does not fly, and therefore violates one of the core criteria of the rule. Yet the penguin is still a bird, albeit a marginal one. It is therefore plausible, even common, that instances can be *partially* compatible with the rules, a situation that is completely disregarded by this view.

- **The Prototype View:** In this view, the concepts are defined by their prototypes (*e.g.*, [9]). The *prototype* of a concept describes how a perfect member of the category should look like, by combining hypothetical “ideal” features. The membership of an instance is then decided by comparing its features with that of the prototype. Memberships are not crisp, so an exemplar with a high similarity to the prototype might be deemed a very typical member of the concept, while an instance with only marginal likeliness can still be allowed as a non-typical member. (See also [10] for an interesting geometrical interpretation of this view.)
- **The Exemplar View:** The exemplar view holds that concepts are composed of a collection of previously encountered exemplars, retained in memory for future reference. Incoming instances are compared against this exemplar set of the concept, to check for membership. Note that more commonly occurring exemplars will naturally with a higher frequency in this representation, and therefore will serve to distinguish typical members from non-typical members by a number of vote. Meanwhile, non-typical members can still be represented and accepted by a less number of exemplars. The BIRD concept, for instance, is expected to be represented by many flying bird instances, and a small number

of non-flying ones.

All of these views have their positive and negative points, see for instance [11] for a review. Evidence supporting each of these claims have indeed been put forward separately, so it is not possible, as of yet, to converge on a single optimal conceptual representation theory, discarding all others. Furthermore, given the high centrality of conceptualization to our thinking, it is quite possible that different strategies are being employed at different tasks [12], which may as well imply an underlying hybrid representation [13]. An important computational model combining these different strategies is Learning Vector Quantization [14], in which concepts are held by prototypes which are updated online via incoming exemplars.

### 2.1.2 Grounding Concepts

A major question related to conceptualization is how we manage to make sense of concepts, and how we can ever share them. Harnad points out in his seminal work [15] that any agent that is merely capable of symbol manipulation might be communicating with us to a certain degree, however it will have no internal *understanding* of these symbols. Called the “symbol grounding problem”, this coins one of the main problems of implementing concepts in an artificial agent: How can we build an intelligence that can not only talk, but also understand what it is talking about?

Barsalou’s pioneering Perceptual Symbol Systems hypothesis [16] suggests that perceptual input may be triggering bottom-up activation in sensorimotor cortices, which are then partially reactivated through top-down mechanisms over the association cortices to simulate these *perceptual* symbols. These distributed and yet connected cortices can then work together to implement a basic conceptual system, out of which emerges conceptualization and categorization.

The famous robotics counterpart of the answer is the embodiment of the agent (*e.g.*, [17–28]). A robot, which is not only an intangible mind, but which instead has a *body* in the world, can understand the world in terms of its own interactions: A *soft* object is one you can squeeze, and *squeezing* is an action that changes the form of a soft object, but not of a hard one. Therefore, sensorimotor experiences can form the basis

of the the internal meanings we attribute to concepts.

Steels' Recruitment Theory of Language [29] approaches the problem from a functional view point. The proposed hypothesis replaces the idea of a dedicated, evolved language "module", with a dynamic, agent-based development of linguistic abilities, through "piggybacking" over previously evolved, more primitive cognitive functions. The specific cognitive modules that prove efficient for developing communication abilities, by providing the linguistic power, may be selected by the agent and kept through developmental stages. In this sense, any region that contributes to the formation of concepts, thereby facilitating interpersonal communication, could be utilized. Perhaps the most obvious candidates, in line with the above theory, would be the primary sensory and motor cortices.

How then a community converges on a common language, with common names for specific concepts? Cangelosi [30] draws attention to the "double function" of language, making the distinction between the communicative function of the language between social agents, versus the cognitive facilitation of language within an individual's own reasoning. With his collaborators, they demonstrate how a shared, grounded lexicon can emerge among a group of agents that struggle to forage in a challenging world [31, 32]. In the course of developing this mode of communication, they necessarily ground the labels on their own sensorimotor experiences. Moreover, these conceptualizations can be transferred from a teacher to a student. Similarly, Steels [33] demonstrates how playing a collaborative naming game between the agents can simulate the cultural transfer of language between agents, when they aim to maximize communicative success in order to "win" in the game. Belpeame and Morse [34], in an attempt to explain how young children learn concepts, compare cross-situational learning of concepts with socially guided learning, to show both are feasible, but social learning is even more efficient. Finally, Cangelosi and Riga [18] discusses how symbol grounding for basic actions can later be transferred to more complicated, higher-order actions. All these works support that symbol grounding need not be strictly individual or on word-basis only: Grounded symbols can be efficiently transferred between agents and concepts, and in fact, it is more efficient to do so, rather than grounding every single concept individually.

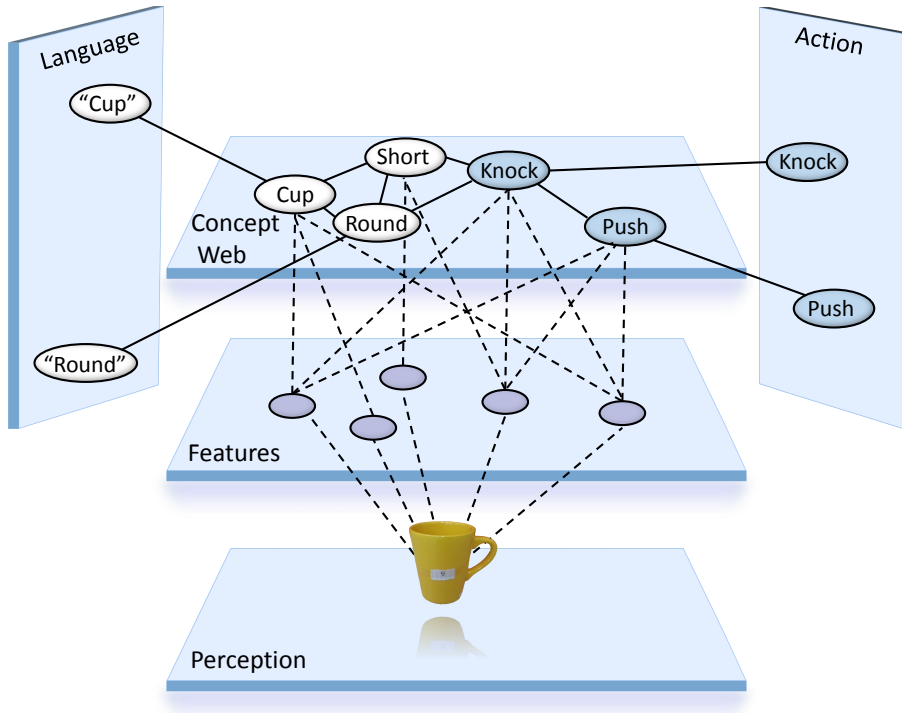


Figure 2.1: The “connected” concepts representation we propose in this thesis. Semantically related concepts are linked together with connections. They are also grounded in perception, action, and language. White color indicates “active” concepts, blue color indicates inactive ones, see Chapter 4 for details. Adapted from [6] ©2015 IEEE.

From a different point of view, Hashimoto and Masumi [35] show how concepts can be modeled as *attractors* in a dynamical system. The transitions between the attractors then naturally corresponds to the manipulation of these symbols during language use. (An interesting finding is that their model did not emerge any regularities in the transitions, therefore this system has not developed an explicit “syntax”.)

Grounded conceptualization of “nouns” and “adjectives”, which may be considered as “object classes” and “object properties”, has received its rightful attention in literature, for instance see [36, 37]; however an organized attempt towards the conceptualization, and especially *generalization*, of actions into verbs is rather recent. Kalkan *et al.* [38] and Rudolph *et al.* [39], propose that behaviors can be defined and subsequently generalized in terms of their effects. Via the effects, we may arrive at a generalization over a rather continuous space of actions: Take, for instance, the be-



havior of “passing the salt”, which can be performed using either the left or the right hand, employing either a power or a precision grasp. Nevertheless, the same goal is achieved in all of these scenarios, irrespective of the actual physical movement, therefore resulting in the realization of the same verb concept.

As we have detailed above, physical grounding is a well-researched solution for developing concepts in artificial agents, however the general approach is to ground each concept *separately*. In this thesis, we would like to go one step further, by proposing to ground concepts *together* via connecting them to not only their perception, action, or language counterparts, but also to other “semantically similar” concepts. In this sense, a “cup” and a “plate” are semantically similar, since both tend to exist in kitchens, and co-occur together often in eating situations. Similarly, a cup is usually short, round, and hard, so these properties are also intrinsically related to the cup concept (Figure 2.1). We continue this discussion further in Section 2.2.

### 2.1.3 Spatial Concepts

A perhaps more advanced question is, how do we conceptualize spatial relations between objects? An understanding of spatial relations of objects in the world is crucial to everyday actions, not only we communicate with each other using them (“Give me the cup on the table.”), but we also plan subconsciously using these relations all the time (*e.g.*, pouring the milk *into* the pot in order to be able to warm it *on* the oven). There is unmistakable evidence collected proving that our parietal cortices are active constantly in order to detect these abundant relations non-stop (see, for instance, [40–42]). Arguably, the only way we could have survived as animals is by carrying an accurate spatial model of the world in our minds: We can close our eyes at any moment and recount the *relative* positions of the objects around us to an astounding accuracy - “I am sitting at a chair standing in front of a table, on which there are two cups, one red and one yellow, a number of books stacked on top of each other standing next to a calendar, which is to the left of the monitor”. But in addition to this instantaneous and automatic spatial modeling, we also have a virtually perfect intuitive *understanding* of the spatial concepts as dictated by the laws of physics: We know we cannot (easily) stand on a basketball because it is round, we understand we

can place a book beneath the monitor in order to raise it, but not an orange, and so on. Therefore there is more to spatial relations than basic world-modeling: A spatial concept is just like any other concept in that it is most meaningful only when considered in relation to the other concepts in our mind.

We argue that prepositional spatial relations should also be regarded as concepts in a web of concepts, in relation to other concepts, just like a noun or adjective concepts. However, since they are binary *and* directed in nature (since, e.g., a ball can stand on a box, but not vice versa, since the ball is round, therefore the spatial relations have *order*), we present a Hybrid Markov Random Field model, which is a variant of the Markov Random Field-based concept web model [6], enhanced to include *directed* connections between spatially-related graph nodes.

How to represent spatial concepts have necessarily attracted much attention in robotics, either from a cognitive point of view, in order to understand and implement their human-like development, or from a pragmatic, utilitarian point of view, in order to endow robots with cleverer acting abilities in the real world. Kuipers [43] had both of these goals in mind, proposing a multi-layered cognitive map called the Spatial Semantic Hierarchy, which can represent the real world in hierarchical levels as information continues to flow, which integrate eventually into flexible semantic representations. A seminal work on the psychology of spatial conceptualization was conducted by Landau and Jackendorf [44], who argued that people do not take into account every detail when extracting spatial relations, instead relying on approximations. For instance, languages commonly provide only the crude descriptions of “in” and “not in” for the important conceptual relation of containment, but there is no detailed propositions describing, for instance, “being in a round object”, or “being inside and also in contact with the inner surface of an object”, etc. Instead, they point out, languages tend to elaborate on nouns, while abstracting over useless details of spatial concepts.

The closest point of view to the one followed in this thesis is raised by Coventry, Garrod, and colleagues, who propose that spatial concepts are as much about *functionality* and *knowledge about conventional usage*, as they are about geometrical relations. For instance, asking subjects to describe the location of a protective object (*e.g.*, an



Figure 2.2: The interference of perceived functionality on the interpretation of the spatial prepositions, experiments by Coventry *et al.* [45]. When the object with a canonical functional usage, *i.e.*, the umbrella that is supposed to protect the man from the rain, is in a position that is able to carry out this functionality, people are more likely to say that the man is *under* the umbrella (middle row). On the contrary, when the object is perceived as ineffective to perform its functionality (bottom row), people do not judge the man as under the umbrella anymore, but rather, as *below* it. Figure taken from [45].

umbrella) with respect to a reference object (*e.g.*, a man) [45] showed that people were more likely to describe the man as *under* the umbrella if it was perceived as effective in protecting the man from the rain (Figure 2.2). Conversely, the description of the man as being *below* the umbrella was more affected by relative geometrical orientation of the man and the umbrella. Therefore, functional and geometric states can differentially effect the interpretations of *above/under* vs. *over/below*. Similarly, Coventry *et al.* [46–49], as well as Garrod *et al.* [50, 51], conducted detailed experiments depicting that, when interpreting the *containment* relation, people make heavy use of the understanding of functionality, and related a-priori knowledge: A ball in

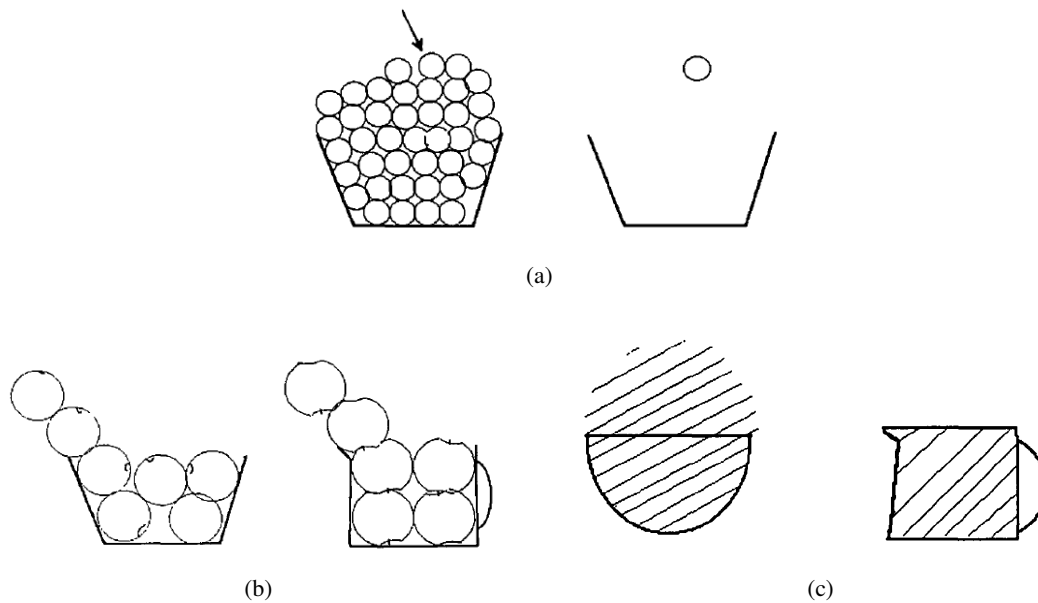


Figure 2.3: The interference of everyday knowledge with the interpretation of spatial relations, experiments by Coventry *et al.* [46, 47]. (a) Although the object is at the same location with respect to the bowl in both figures, people are likely to judge it *in* the bowl only in the left figure, where the object indirectly “touches” the bowl via the other objects in between. (b) Although both the bowl (left) and the jug (right) are full to the same amount, people are likely to judge the oranges on the left as *in* the bowl, while the apples on the right as *out* of the jug. Coventry *et al.* hypothesize that this difference is due to the canonical usage of the jugs to hold liquids, which can then physically be filled only until the rim. The bowl, on the other hand, is commonly used to hold solid objects. Therefore, people have in their minds a larger canonical containment range for the bowl, as compared to the jug (c). Figures taken from [47].

contact with the inner surface of a bowl is *in* the bowl, even if this is “indirect” contact, whereas without contact, the same ball at the same location is not considered in the bowl (Figure 2.3a). Similarly, when too many oranges are stacked in a bowl so that some are overflowing from above the bowl, people still judge these oranges as *inside* the bowl (Figure 2.3b). However, when the same scenario is performed with a jug instead of a bowl, they do not judge the oranges as inside the jug. Coventry *et al.* [46, 48] explain this behavior by mentioning the regular usage of bowls with solid items, which can overflow the bowl but still remain in place, versus the regular usage of jugs with liquids, that spills out in a similar scenario. Therefore, they conclude that

people have a larger *space of containment* idea for bowls, which effectively extends their range. (Figure 2.3c) Once again, the a priori intuition of the rules of physics, meddles with our conceptualization.

Trying to apply the principles of spatial labeling to Human-Robot Interaction, Fischer [52] investigated the variables affecting people’s choice of spatial instructions when interacting with a robot. Stopp *et al.* [53] studied how a robot can anchor verbal spatial descriptions to its physical environment, thus grounding them, proposing a compositional variant of spatial potential fields. Gold *et al.* [54] showed how prepositions, together with pronouns, can be extracted and represented as word trees, depending on entropy and information gain metrics applied on the physical environment. Moratz and Tenbrink [55] developed a system for iterative interpreting of projective relations in human-robot interaction scenarios, in order to enable mutual identification of objects in the environment between the robot and the human partner. Roy *et al.* [56] proposed that a physical simulation-based mental world model can be employed for allowing a robot to *shift* between his own perspective (“my left”) and that of the human partner’s (“your left”).

Van de Weghe and colleagues [57, 58] pointed out the qualitative (as compared to quantitative) nature of spatial representations in humans, for example describing something crudely as “on the right” rather than as “at  $\theta^\circ$  to the right”, and proposed Qualitative Trajectory Calculus (QTC) as a qualitative formalization of the relative motions of two agents. Such a generalized and qualitative representation naturally allows focusing on the core aspects of the spatial interactions, instead of getting lost in quantitative descriptive details of motions. Hanheide, Bellotto and Van de Weghe then showed how such a representation could be used in a naturalistic manner in robots, for instance for comparing and modeling behaviors of humans to guide the robots, and providing the common understanding for human-robot spatial interactions [59–61]. Dealing with the reverse problem, Iliopoulos *et al.* [62] showed that it is also possible to inverse-map qualitative QTC descriptions back into trajectories that satisfy them. Yi *et al.* [63] propose a topological-semantic distance map, which considers the “spatial contexts” of objects and robots (*e.g.*, robot on the left of X, X is in front of Y, etc.) using a probabilistic Bayesian model, which similarly allows qualitative descriptions of environments, in order to allow building more flexible spatial representations.

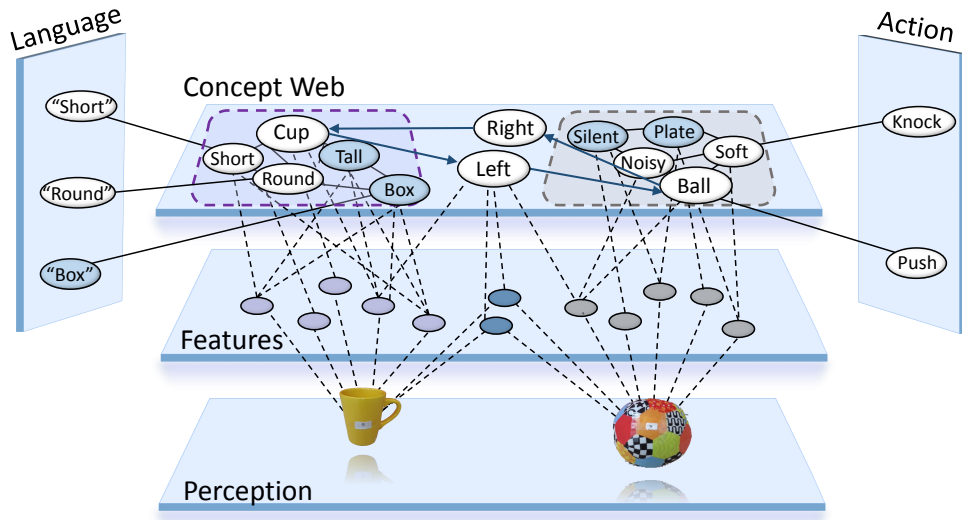


Figure 2.4: The representation of spatial concepts in our system, as bona fide concepts themselves, that connect other noun, adjective, and verb concepts but are also flexible and subject to refining and reconsideration through the integration of knowledge coming from neighboring concepts. White color indicates “active” concepts, see Chapter 4. Adapted from [67] ©2015 IEEE.

Investigating the formation of spatial concepts, Golland *et al.* [64] showed that, when trying to minimize the risk of miscommunication between two collaborative agents, *discovering* descriptive spatial labels is more effective than sticking with pre-determined labels. Inspired by this result, Guadarrama *et al.* [65] proposed a system to learn spatial prepositions and object representations simultaneously, combining strategies of template matching, syntactic parsing, and probabilistic analysis. Tellex *et al.* [66] showcased a robotic forklift scenario, to be controlled by natural language commands, as a testbed for language grounding, in which they try to learn the parameters for a probabilistic graphical model from a corpus of commands.

Out of the existing modeling studies, the most relevant to our approach is of Anand *et al.* [68], who used spatial relations between noun concepts to guide a visual search via contextual information, using a Markov Random Field. In their work, each object part corresponds to a node in MRF, and detected spatial relations between the parts are used to connect the nodes to each other. Hard-coded, rule-based spatial relations such as “on top of”, “in front of” are then integrated into the model as edge potentials to improve the accuracy. A similar approach is also taken by Misra *et al.* [69], who

assume simple geometric relations between objects to define contexts. Our approach is different in that the spatial prepositional relations are themselves concepts that link other noun, adjective, and verb concepts that are related to the objects. Therefore, the spatial concepts are subject to reasoning and reconsidering: Instead of assuming perfect geometrical perception of the spatial configuration, our system is able to incorporate a priori knowledge about the world into its assessment of the spatial concepts. This hypothesis is in line with the psychological experiments of Coventry *et al.* [46–49] and Garrod *et al.* [50,51], who point out to the significant level of interplay between everyday world knowledge and the interpretation of spatial relationships between objects. Figure 2.4 provides an overview of the modeling of spatial concepts in our system.

## 2.2 Structural Representation of Concepts

What are the exact cortical mechanisms that hold concepts in the human brain? And how are these mechanisms replicated to date in artificial agents? In this section, we try to present the literature related to the findings and models about the structural representation of conceptualization.

### 2.2.1 Structural Representation in Humans

What are the insights we currently have regarding the structural representation of concepts in the human brain? Where and how are the concepts held, where and how are they connected to each other? There are two main hypotheses on the exact mechanism of conceptual representation in the brain, on both of which, interestingly, strong arguments have been proposed. The two hypotheses are the *distributed representation* and the *localist representation* hypotheses.

The initial proposal of the **distributed representation** hypothesis is owing to Wernicke and Meynert (see [70] and [71] for detailed discussions). Their proposal is that concepts are made of modality-specific engrams, which reside in their corresponding primary sensory or motor cortices. Since they are fully connected, any hint of the concept, by its name, sound, or taste, would activate the whole web, rapidly calling

into attention holistic knowledge about the concept.

The popularity of this hypothesis is not arbitrary: There exists significant supportive neuroimaging evidence in favor of it. Goldberg *et al.* [72] and Kellenbach *et al.* [73] demonstrated modality-specific cortical activations during semantic retrieval and decision-making tasks. Goldberg *et al.* [72] showed how (1) accessing tactile information activates somatosensory, motor, and premotor cortical areas, (2) flavor information retrieval activates orbitofrontal region, (3) visual information retrieval activates ventral temporal cortex, and finally, (4) auditory information retrieval activates superior temporal sulcus. Kellenbach *et al.* [73] similarly demonstrated increased posterior inferior temporal activation for color-related decisions, posterior superior temporal gyrus activation for sound-related decisions, and right medial parietal cortex activation for size-related decisions. Similar results have also been obtained related to conceptualization of actions. In his famous work, Pulvermuller demonstrated that motor and premotor cortices activate somatotopically during purely linguistic usage of action words, specifically tongue-related, peri-sylvian area activating for the word “lick”, finger-related, lateral area for “pick”, and foot-related, dorsal area for “kick” [74–76], Figure 2.5. The detection of such somatotopy is especially important, since it hints a systematic co-activation of these motor areas respectively for each category of verbs. Chao and Martin [77] similarly conducted a tool viewing-and-naming task, which causes selective activation in left ventral premotor, as well as left posterior parietal cortices. The importance of these findings are better appreciated when we acknowledge that grasping a tool for using it is an integral part of the tool concept: Therefore, spatial and motor areas must be highly relevant to the semantics of the tool.

Recently, support for the concepts-as-combinations-of-concepts hypothesis has also come from a pioneering study of Mitchell *et al.* [78], which proposes a paradigm shift in neuroimaging: That it is possible to be *predictive* about neuroimaging research. In this study, Mitchell *et al.* show that the fMRI activations for complex words, such as *celery*, can be *predicted* by superposing previously known fMRI activations for a set of 25 basic nouns, including for instance *eat, taste, see, hear, smell, manipulate, touch, say, and move*. Apparently, these activations can simply be *added* together to achieve the expected activation of the complex word. What is especially striking is



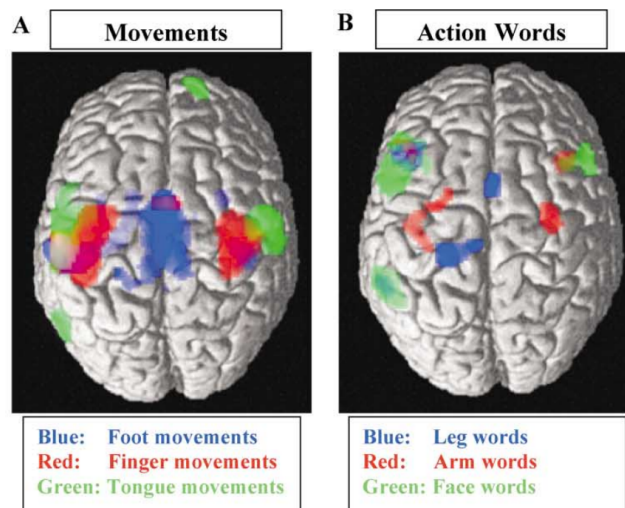


Figure 2.5: The somatotopic activation of the premotor cortex to the words “pick”, “lick”, and “kick”, compared to the motor areas that actually activate while performing picking (hand-related), kicking (foot-related), and licking (mouth-related) actions. Figure taken from [75].

that the multiplicative weight of each simple word’s activation pattern in the whole sum can be taken simply as this word’s *co-occurrence* with the target word in a large text corpus. Literally, every simple concept is related to the complex concept as much as they *co-occur*. O’Toole *et al.* [79] also find that object categories, such as chairs, faces, houses, etc., which have common visual attributes (*e.g.*, different types of chairs), are represented by “shared neural structures” in the ventral temporal cortex, supporting a “feature-based representation of objects”.

Given the large amount of supportive findings, it is not surprising to see in the literature a plethora of theories that regard the representational basis of concepts as a connected-web-of-cortical-areas (See for instance Pulvermuller [74], Damasio [80], Bryson [26], and Deacon [2]). Note that this theory can elegantly integrate initial physical experience, subsequent memory retrieval, and even high-level reasoning. The reader should also note the strong relevance with the theories of Barsalou [16] and Steels [29].

The evidence supporting the distributed representation hypothesis is very strong, and in fact could be deemed decisive, if not for the rivaling and equally, if not more, strong evidence supporting the exact opposite, the **localist representation** hypothesis. Note

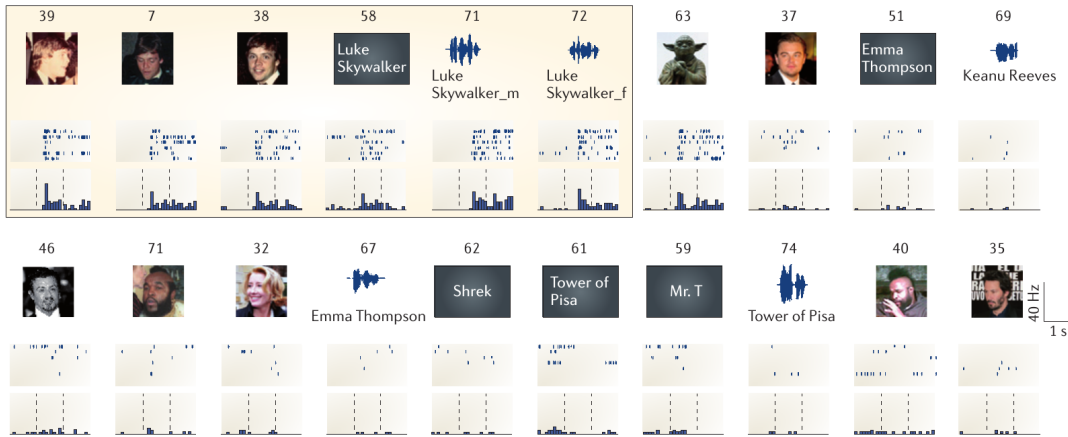


Figure 2.6: The single neuron that fires selectively to Luke Skywalker, whether presented visually, as a text string, or in audio modality, while also firing for a picture of Yoda [101]. This neuron is therefore showing selective activation to the encapsulating Star Wars concept, regardless of the specific modality or detail of the presented cue, providing support for the single concept cells hypothesis. Figure taken from [101].

that it had already been well-researched that specific concepts are being recognized through dedicated cortical areas in the brain [81], the most clearly identified ones including the human face [82–88], human body parts [89–91], outdoor scenes [92, 93], and body movements [94, 95]. However, much more conclusive evidence came from single-cell recordings, performed both in animals, and via intracranial recordings of human epilepsy patients (for a comprehensive survey, see [96]). It has been shown that there are single cells in Medial Temporal Lobe (MTL), the so-called *concept cells*, which are highly concept specific, and fire when the subject is presented with a “concept”, regardless of the specific visual pose or visual context: Specifically, neurons have been identified which fire selectively when the subject is shown a picture of a certain category (*i.e.*, a face, animal, house, scene, famous person, car) as opposed to visuals of a different category [97, 98]. Interestingly, of the “face-selective” cells in the work of Fried *et al.* [97], most also fired selectively according to the gender, emotional expression, and novelty of the face. Kawasaki *et al.* [99] similarly identified neurons in the human cingulate cortex that are selective for one *emotion class*. Gothard *et al.* [100] found cells in the monkey amygdala, which responded selectively to monkey faces, human faces, and objects.

Even more strikingly, Quian Quiroga and colleagues [102] could identify cells, again

in MTL, that fire to the photos of Jennifer Aniston, and to that of Jennifer Aniston only, instead of any random face. These cells showed equally strong activation irrespective of the exact pose or the context in the picture, and very interestingly also activated slightly to the photos of Lisa Kudrow, who co-starred with Jennifer Aniston in the popular TV show, Friends. A similar finding was also recorded for a neuron that fires selectively to “Luke Skywalker”, but also to Yoda (Figure 2.6). In subsequent studies, Quian Quiroga and colleagues [101, 103] demonstrated quite convergent results to this end, locating cells that fire for (1) photos of “Saddam Husein”, as well as to his name as pronounced from a computer, and (2) to photos of Halle Berry, even when she was masked as the “cat woman”, to the visualized string “Halle Berry”, and again to the pronunciation of the name. Suthana and Fried [104] could localize a neuron that fires selectively to the Sydney Opera House. (Interestingly, this neuron also fired for the Bahai Temple in India, which the patient in the study verbally reported confusing with the Sydney Opera House.)

These “concept cells” naturally drew huge interest, and were investigated under various conditions: Lin *et al.* [105] showed the existence of “nest” cells in mouse hippocampus, which fire when the mouse encounters a potential nest, irrelevant of its exact shape, structure, size, or material. The only identified factor in this study was the perceived functionality of the nest: The cells selectively fired to nests that are potentially usable, for instance, with an unblocked entrance. Yoshida and Mori [106] located odor-category specific cells in the olfactory cortex, that are able distinguish between food types, say, between a watermelon and a grape. Sugase *et al.* [107] demonstrated how the time evolution of such concept cells matters: They showed that the firing pattern of a concept cell can start by discriminating between gross classes, say between human and monkey faces, and become more specific in time to provide more detailed information, in the work discriminating between identity of the face, or its emotional expression. Hung *et al.* [108] were able to *read* the seen objects identity, as well as category, from a small population of 100 neurons in the macaque Inferior Temporal Cortex. These neurons were also highly correlated with the *consciousness* of having seen a concept: When Kreiman *et al.* [109] and Quian Quiroga *et al.* [110] *flashed* photos of concepts in a difficult-to-perceive paradigm, they found out that these cells fired *only when* the subjects were consciously *aware* of what they

saw. Recently, Roy [96] presented a very in-depth literature survey on the subject, as well as discussing in detail the potential computational and biological advantages on having indeed local, concept-specific centers in the brain, instead of merely relying on a widespread distributed representation.

With ample support for the both opposing views, there is certainly great need for further investigation and theories. Recently though, a number of studies emerged, which may possibly provide a synthesis of these two opposite antitheses [71, 111–113]. What they pointedly try to determine is whether this “cortical web” of primary cortices is enough to represent a concept, or if there is a dedicated region that orchestrates the combination of these low-level cortical activations into the coherent concept meaning. (Damasio [80] and Damasio *et al.* [114] also hint a similar idea when they mention high-level, amodal convergence zones, from which the time-locked activation in primary cortices is orchestrated.) Lambon Ralph [71] and Patterson *et al.* [111] conduct lesion and neuroimaging studies of Semantic Dementia (SD), in which the semantic knowledge is selectively and progressively lost, while the other cognitive abilities remain intact. Patterson *et al.* in [111] recounts a striking anecdote: “When we asked one of our patients to name a picture of a zebra, she replied: ‘It’s a horse, ain’t it?’ Then, pointing to the stripes, she added, ‘But what are these funny things for?’” In semantic dementia, the primary cortices and their association areas are intact, therefore the patient can classify the picture as a horse-like animal. What’s more, she can successfully detect the stripes visually. However, the *concept of a zebra* is lost, therefore she converges on the next close concept (*horse*) that is still available. Another example is a patient (who is competent in all other cognitive facilities) asking, “What are those things?” to a herd of sheep. This unusual form of dementia is associated with degeneration in the Anterior Temporal Lobe (ATL) (for instance, [115–117]. Kellenbach *et al.* [73] also interestingly recount unexpected activation in ATL during a semantic task, which was not a specific region-of-interest in their study, but whose activation is actually meaningfully accountable by this hypothesis).

Lambon Ralph [71] and Patterson *et al.* [111] then propose the ATL-as-a-“semantic hub” hypothesis, in which they propose that ATL may be the region that connects the widespread cortical webs into meaningful entities, corresponding to concepts. They

suggest that the often complex and nonlinear boundaries of concepts might require such a “hub”: After all, although concepts are indeed collections of features, these features usually bind together in nonlinear and complex fashions. Lambon Ralph *et al.* [113] suggests that the extra ATL layer might indeed have a functionality that is similar to that of the hidden layer’s in a multi-layer neural network. A single-layer neural network can only bind together linear features, and is therefore unable to represent certain functions. Incorporating even a single additional layer allows generating potentially any function. The progressive nature of semantic dementia also allows testing this hypothesis explicitly: Working with mild and severe semantic dementia patients, Lambon Ralph [71] and Patterson *et al.* [111] show that, patients with mild dementia have a tendency to make under- and over-generalizations when categorizing rather non-canonical entities (camels with humps, pumpkins as vegetables, etc.) These mistakes are become more prominent in severe semantic dementia, where patients cannot remember object features that are not prototypical of their category: They may, for instance, draw ducks with four legs, as typical of the animal category [111], failing to make an exception for the duck sub-concept.

In a recent and paradigm-wise pioneering study, Huth *et al.* even proposed that the representation of thousands of known entities could be spread on the cortex as a *continuous* semantic space. The main hypothesis of this work is that since virtually (tens of) thousands of items are recognizable by humans, it is more likely and space-efficient to find these representations as a continuous semantic map, rather than individually in a non-topographic manner. Conducting a pioneering principal component analysis of the semantic space of 1705 action and object categories used as stimulus, they are able to show that identified principle components correlate with the voxel activation map - meaning that the activation of these concepts in the cortical map follows a semantic spatial distribution. Still, these results are similarly debatable, see for instance [101] for arguments in favor of a non-topographic conceptual organization, for instance reporting how Halle Berry and Mother Teresa are found to be represented by neighboring neurons. And yet there is the common point that both of these figures are humans, and *famous* humans as well. Quian Quiroga *et al.* [101] also theorizes that concepts are held in sparse cell assemblies rather than in single cells, and semantically connected concepts may share some cells in their assemblies, which

can also be a factor to consider when evaluating these results. To sum it all, there are numerous interesting findings, whose degree of compatibility with each other varies, suggesting that conceptualization is a very complex cognitive function with a non-trivial structural representation mechanism, and which will certainly promote more in-depth investigations in the future.

### **2.2.2 Structural Representation in Robotics**

Concepts and their representation have inspired numerous computational and robotics studies as well, some of which trying to unveil the mystery by presenting testable models, while others mainly aiming to solve the perennial learning and adaptation challenge in robots. One of the most organized attempts of formalizing concepts for robotics use came in the form of a knowledge processing framework, KNOWROB, proposed by Tenorth and Beetz [118, 119], Figure 2.7. Their main point was to develop a system which can process information efficiently as humans do in real life: By filling in the gaps in conversation with background knowledge. The system can connect to external information sources, such as the Internet or dedicated databases, and possesses manipulators with which it can utilize accessed unformatted information in various tasks freely. Information is kept unformatted (“virtual”) until it is needed, and then can be searched freely for associations. Concepts in KNOWROB can be objects, actions, events, or places, and are organized in a hierarchical manner with more specific concepts inheriting from more general ones. Multiple inheritance is allowed, which enriches membership definitions. Actions are defined as recipes, events as change of states, and all of these are inherited from a general “thing” entity, which is the common ancestor of all nodes (object, action, event, or place) in the ontology. Later on, this system has been extended by Palmia [120] with the aim of mutual understanding and cooperation between multiple robots. Another notable example is Tamosiunaite *et al.*'s [121] utilization of syntactic bootstrapping [122] for the robot to learn, from conversations with humans and online images depicting events, what actions can be used on what kind of objects, effectively generalizing objects with respect to actions in the process. This approach allows the robots to conduct flexible reasoning on huge amounts of data acquired from the Internet, as well as to perform error handling and/or guide the supervisor by asking questions if necessary.

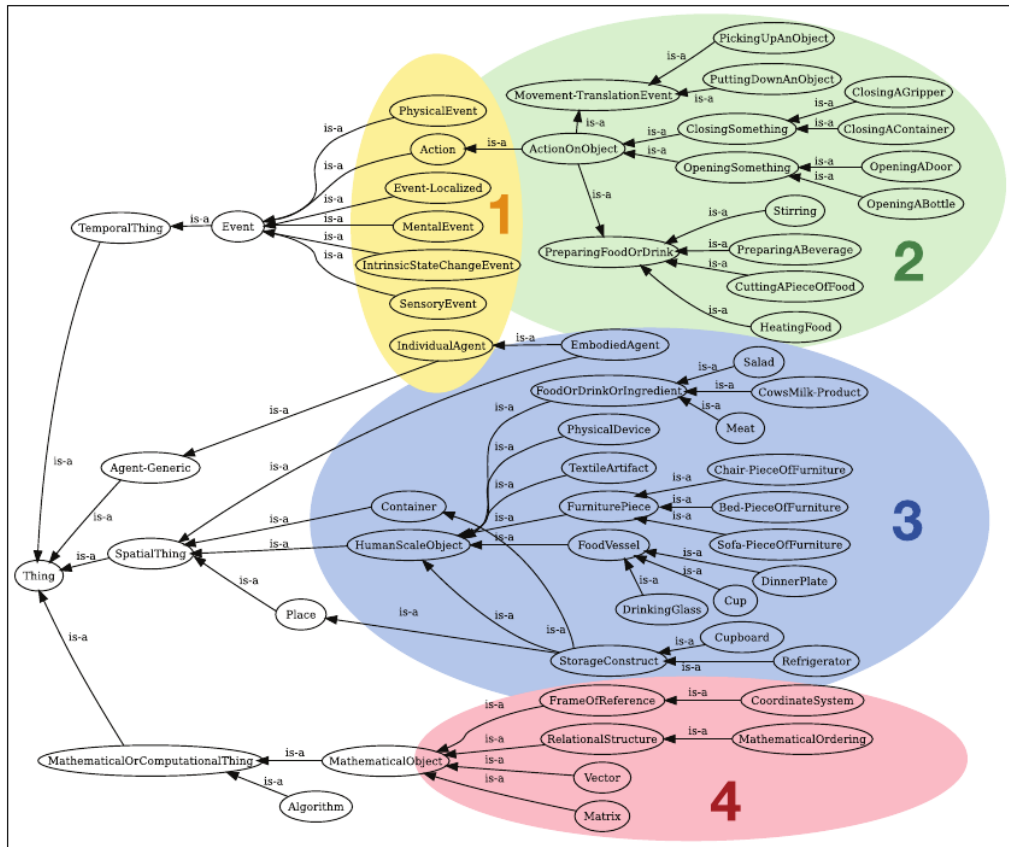


Figure 2.7: The KNOWROB system proposed by Tenorth and Beetz [118]. Everything in the world, be it an object, action, or event, is a part of a (tree-structured) ontology, and derives from the most general class of “Thing”. Figure taken from [118].

On the other side of the fence, there are studies which aim to close the gap with what we know of human cognition. Baxter *et al.* [123] propose a connected developmental architecture of conceptual memory (Figure 2.8). The membership of instances to concepts are defined in terms of Euclidean distance of all features to concept prototypes. They also learn associative links between different feature spaces in a developmental manner, reminiscent of Hebbian learning. (However, these associative links connect only different modalities of the same concept, and not different but semantically related concepts.) In yet another attempt to bring together different modalities, Morse *et al.* [124] use that the “body” of an agent as a “hub” to connect the visual, auditory, and spatial information, enabling the grounding of concepts such as red and cup. [37, 125–127] use the formalization of affordances to ground actions.

Another prevalent approach for conceptual representation in robotics is assuming that

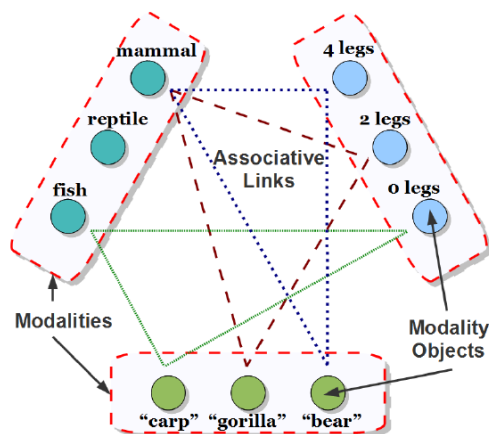


Figure 2.8: The associative web of conceptual memory proposed by Baxter *et al.* [123]. Different modalities of the same concept are connected together with associative links. The information about which modalities are included in a concept are learned from the zoo database of UCI Machine Learning Repository . Figure taken from Baxter *et al.* [123]

concept formation occurs in an incremental manner in the form of a hierarchical structure; *i.e.*, a hierarchical representation is assumed of concepts [128–131]. In this hierarchy, upper concepts represent the general concepts, whereas lower or terminal concepts refer to the specific properties or instances. The connections imply *is-a* type relations between concepts. Instances can be placed into lower or terminal nodes. The tree structure of the hierarchy also provides an option for branching. Top-down classification of instances depends on selecting the best branch or a set of branches to go deeper in a tree, similar to Quinlan’s decision tree approach [132]. One of the earliest attempts is the Elementary Perceiver And Memorizer (EPAM) model [128, 129], which holds nodes with attribute-value pairs in a tree structure. Each edge coming out of a node represents a certain value for a comparison criteria. Leaf nodes correspond to specific *images* of instances. EPAM makes a distinction between classification and prediction tasks as two different processes. This model has later been extended by UNiversal MEMory (UNIMEM) [130] to include confidence and feature frequency statistics, nominal values and images in non-terminal nodes. COBWEB [131] has been inspired by these models, as well as CYRUS [133], and introduced an evaluation function which rewards intra-class similarity and inter-class dissimilarity. Finally, CLASSIT [129] enhanced COBWEB by including mean and



standard deviation values for attributes. The common point in all these hierarchical methods is that they use the hill-climbing search method and each concept node has their own attribute-value pairs.

The general missing point in all these works is the lack of a global structure of associativity. Connected together are only modalities of concepts, or concepts stemming from the same ancestor, such as “cup” as a container, and “glass” as a container. On the contrary, none of these models present a feature of long-distance associativity between seemingly different, but semantically related concepts, such as “water” and “glass” should be related by means of “drinking” action. Moreover, the concepts these studies are generally not grounded: They rely on either ontologies or Internet-based information, or a hand-designed set of features. Therefore, they lack a major extendibility from a developmental point of view [25]. Figures 1.1a and 1.1b contrasts our proposal with the existing systems visually.

### **2.3 Context**

Why is context important? We as humans should know perfectly well, given the fact that we use it shamelessly. Context determines how we walk [4], how we talk [134], how we think [135, 136]. Basically, even if we would like to, it is virtually impossible to get rid of the context in our minds when assessing things, that is why humans are so prone to famous cognitive biases such as stereotyping [137], the halo effect [138], the primacy effect [139], and the framing effect [140]. Barsalou [141–143] notes:

“...[C]oncepts are not typically processed in isolation but are typically situated in background settings, events and introspections. When representing bicycle, for example, people do not represent a bicycle in isolation but represent it in relevant situations [...] [P]eople situate concepts for the following reason: if the brain attempts to simulate a perceptual experience when representing a concept, it should typically simulate a situation, because situations are intrinsic in perception. At any given moment in perception, people perceive the immediate space around them, including agents, objects and events present. Even when people focus attention

on a particular entity or event in perception, they continue to perceive the background situation - the situation does not disappear.” (Quotation belongs to [141].)

### 2.3.1 Context in Robotics

Robotics has achieved significant success in terms of both theory and applications in the past five decades [144]; however, research involving context has focused on the environmental aspect only, *e.g.*, in scene interpretation [5], urban search for rescue tasks [145], home security [146] and elderly people’s living environments [147], object recognition in daily activities [68, 148], and trying to fulfill possibly incomplete natural language instructions of humans [69].

Kim *et al.* [149] propose an architecture for facilitating contextual reasoning among network-based service agents, however they resort to a tuple-based, naive context representation, in which they assume perfect knowledge of ungrounded, high-level information. For instance, the age information is assumed to be part of a context regarding a human. Types of information related to various contexts are connected assumed to be a priori designed by the programmer, and is therefore not flexible or adaptable. Due to these reasons, this study rather represents an attempt towards a context representation at a more structured environment with high reliability.

Wibisono *et al.* [150] aim at a highly specific application of contextual information, namely vehicle-to-vehicle environments and autonomous driving. Contexts of Low-risk, Conflict, and High-risk situations are pre-defined and are tried to be identified, which would rather guide the behavior of the autonomous vehicle. There are assumed to be again pre-defined context-attribute values, including the existence of lane-change behavior, following-distance with the neighboring car, existence of emergency-braking, and so on, all of which readings are also augmented with confidence values. These readings are combined according to their confidence values and the resulting risk-context is deduced. As mentioned above, this is again an interpretation of context for a highly specific environment, which could not be trivially transferred to another domain.

Wang *et al.* [151] proposes using OWL web ontology language, for devising contextual information on a knowledge base of 300 objects. They assume that every object has been fully analyzed through the ontology, and every relevant information about the object has been extracted, *i.e.*, that the function, shape, location, tasks, etc. of the object has already been identified perfectly. They use these attributes of the objects to detect the context. Their results are presented on simulated knowledge only, and has not been validated on a real robotics scenario. The most important restriction is the assumption of the ontology, and the dependence on the perfect identification of object attributes with respect to the ontology, which again restricts the proposed system to a very structured and well-defined environment.

As mentioned above, Yi *et al.* [63] use the idea of a “spatial context”, such as *nearby(object, robot)*, *left-front(object, robot)*, in order to qualitatively describe localizations and to build a topological-semantic distance map. However, this work is again restricted to a single domain (spatial localization), in which the spatial context is described by hand-coded relations (“left”, “right”, “front”, etc).

The works of Mastrogiovanni *et al.* [152] and Padowitz *et al.* [153] are more relevant to our work in terms of aiming for a more general formalism of context. The main aim in both of these works is to provide a formalization of context on which multiple agents can contribute at the same time, improving each other’s guesses. However, both of these works again resort to a rule-based, crisp-logical formalization of context (defined as conjunctions of predicates), which suffers from the same disadvantages that we listed for the rule-based concept theories above in Section 2.1.1, namely an inflexible representation in which all-or-nothing membership is required, forcing either 100% existence of a context or none, as required by the *rule* defining the context. Moreover, the contexts that can possibly exist in the world are assumed to have been perfectly identified by the programmer a priori, who should also perfectly define the conjunctive *rules* describing possible contextual states. These rules are assumed to be static. Therefore, the agent does not have a real understanding of the possible context, but is merely responsible with comparing certain attribute-value readings with the predefined rules, while also considering other agents’ interpretations. The main contribution of Mastrogiovanni *et al.* [152] is the incorporation of a temporal domain, they extend the logic-based representation to include time, and therefore can

define, for instance, precedence relations between events (*i.e.*, Event A has happened before event B), and can reason on when a contextual state starts and when it ends. Padowitz *et al.* [153], on the other hand, introduces a geometrical interpretation of context, in which the *contextual state* is a point in a multi-dimensional space, where the dimensions correspond to the attributes that affect the context. They also define the context as a weighted combination of these attributes, thereby allowing some attributes to be more central, or possibly defining the context themselves. On such a multi-dimensional space, they then identify hyper-volumes, which correspond to the states of the attributes that are consistent with given contexts. The main point of both works is to provide a collaborative reasoning environment, in which they can merge highly uncertain information coming from different agents, arriving to a system that is capable of performing well in an unstructured environment. However, they depend hugely on the assumption of the possible “contexts” being known by the programmer, and being encoded by stable rules, thereby resulting in non-adaptive frameworks. For a detailed review of the above robotics works, the reader is referred to [154].

Anand *et al.*'s work [68] stands out from the robotics point of view, for, although not proposing a formalized account of context, still making use of contextual cues in a semantic search and labeling application. They likewise use a graphical model to represent visual features and shape cues, which they augment with the geometric context information, such as monitors being usually found on-top-of a table, chairs next-to a table, etc. They show that learning and using where object is most likely to be found is beneficial for later searches of the object, as well as interpreting given scenes. Therefore, in their work, the notion of “context” is strictly limited to geometrical context, with spatial relations assumed to be perfectly sensed. Misra *et al.* [69] is also similar to our approach for utilizing a graphical model. Their main aim is to detect prerequisites in possibly ambiguous commands issued by a human partner. They treat context as multiple-choice values of the states of known objects in the environment (*i.e.*, microwave door is *closed* or *open*), and check the desired states of these objects before attempting to perform commands, for instance, the refrigerator door might need to be in the *open* state, before the robot may be able to take the milk out. These “contexts”, on the other hand, are restricted to previously defined states of previously defined objects, rendering this model rather more suitable to highly struc-

tured and static environments, and requiring very specific domain knowledge from the programmer.

### 2.3.2 Context in Related Fields

Although context has not been studied either formally or with adequate generality in a robotics domain, it has stimulated significant thought in a variety of other disciplines, a fact not surprising given the abundant implications of the topic. The field that is most obviously interested in context is perhaps the study of language, in which there exists virtually universal agreement on its importance. Natural languages are inherently ambiguous, and contextual information is vital in disambiguating the meaning [155–158]. Coventry *et al.* [45–48, 158] and Garrod *et al.* [50, 51], that spatial prepositions are as related to *functional context* as they are simple geometrical relations. In psychology, similarly, there seems to be universal recognition of the vitality of context. Schank and Abelson [159] proposed the importance of “scripts” for reasoning about common situations in daily life: When in a restaurant, for example, we know that we can find nearby a menu, dishes, a waiter, a chef, and so on. This idea has motivated the creation of formal ontologies later on. Moreover, context has a clear effect on affordance perception, where the perceptions of affordances are affected when objects embedded in certain scenes (*e.g.*, [160]) or surrounded by specific objects (*e.g.*, [161–163]). As mentioned above, Barsalou drew attention to the explicit need for understanding and situating of concepts in the context of other concepts [141, 143]. Planning in humans, as well, being a complicated and time-consuming ability [164–166], is shown to be tremendously facilitated with contextual awareness [167–170].

Although potentially most obvious, language and psychology are not the only fields recognizing the importance of context, though. In their seminal paper [171] Biederman *et al.* showed that visual recognition of objects were highly facilitated when perceived in environments that are *conceptually relevant*. In fact, context seemed to have a dominant effect in visual object recognition, for instance as showcased by Bar [5], in which people can judge a *visually identical* object as either a hairdryer or a drill, depending only on the visual context (Figure 2.9). Driven with the same hypoth-



Figure 2.9: The effect on visual context on object recognition. Note how two visually identical items can be judged as both a hairdryer (on the left) and a drill (on the right). Figure taken from [5]. Best viewed in color.

esis, Torralba [172] proposed a Bayesian model of object detection based on context, exploiting the strong regularities on the statistical correlations of low-level features of both the object and the scene, and demonstrating emergent effects of object priming, context driven attention, and automatic scale-selection effects on this model. Rabinovich et al. [173] propose using context of visual scenes as a post-processing cue to enhance categorization performance. This is done by incorporating contextual information into a Conditional Random Field, which then enforces semantic constraints. Marszalek et al. [174] demonstrated how contextual information between actions and their typical settings can be learned, and this exploiting this correlation improves the detection performance of both actions and also scenes.

Even in AI, which has been ambivalent with the notion of context traditionally, McCarthy [175] proposed the “rectification” of context in classical AI, and argued that AI needs to revise its point of view by putting the notion of a context to the center stage. he argues that intelligent machines “must construct or choose a limited context containing a suitable theory whose predicates and functions connect to the machine’s inputs and outputs in an appropriate way” [176].

### 2.3.3 Structural Representation of Context

Acknowledging the importance of contextual information in object and scene recognition, Bar [5] tries to identify the relevant structural mechanisms in the brain that

enable the extraction and utilization of this context. Apparently, context can be extracted with surprising rapidness in the visual stream [177]: People can “understand” visual scenes as early as 100 ms [178, 179], and even gather semantic conceptual information from very short exposures of 80 ms [180]. Bar [5] proposes a low vs. high spatial frequencies hypothesis for this rapid processing: According to his proposal, the global cues, encoded within the low frequency band in the view might be processed first and very rapidly, providing essential quick assessment of the context. Afterwards, attention might be turning towards the high frequency features, which can give detailed information about the scene in exchange of greater computational cost.

Another attribute of contextual processing is that it commonly does not need a deliberate attention or awareness. Apparently, it can be processed and learned implicitly [181–183], and in spite of the absence of attention [184].

Bar [5] argues that context is related to associative processing (a hypothesis that is central to this thesis, as well), and is therefore related to the cortical areas in the brain that are known to deal with associative information: namely medial temporal lobe (MTL), which includes the hippocampus, parahippocampal cortex (PHC), the perirhinal cortex, and the entorhinal cortex. In addition, there is abundant evidence on the participation of a certain area in the PHC, called the parahippocampal place area (PPA for short), in the processing of topographic and spatial information - which is significantly related to the contextual information. There is also evident relation between contextual processing and the famous N400 signal [185], which is a strong negative response in the event-related potential of the brain at the 400 ms mark from the onset of the input, *in case* the input is contextually incongruent. Therefore, the N400 response signals the detection of inconsistency with context, or in other words, “senseless sentences”. This response is given not only to verbal input, but also to visual input [186]. There has been recordings of prefrontal cortex (PFC) activation in studies of the N400 response [187–189], which was also synchronized in time with and is therefore possibly related to the activation in MTL [190, 191]. Finally, retrosplenial cortex (RSC) was found to be included in the processing of highly contextual spatial information, such as environmental landmarks [192–194]. In conclusion, given the evidence above, Bar [5] argues for the possible contribution of the parahippocam-

pal (PHC), prefrontal (PFC), and retrosplenial (RSC) cortices in the processing of contextual information.

In spite of the piecemeal attempts for a contextual representation in robotics and elsewhere as mentioned above, to the best of our knowledge, there is not yet a systematic study of formalizing context from a developmental robotics point of view, where the robot discovers and then makes use of general contextual cues as it tries to discover its world. Moreover, in general, context studies do not comply with the developmental paradigms, instead depending heavily on the programmer's background knowledge on what constitutes a context, in general how many contexts exists in the world, and so on. In this thesis, we try to tackle these problems by proposing a fully developmental formalization of context for a robotic setting. In line with the hypotheses of Yeh and Barsalou [143] as well as Bar [5] and many others mentioned above, we assume that "visual objects are contextually related if they tend to co-occur in our environment" [5], and through quantitative experiments prove the feasibility of this hypothesis, and how it can quantitatively improve object recognition. Figure 1.1c depicts the proposed model of context visually.



## CHAPTER 3

### THE EXPERIMENTAL SETUP

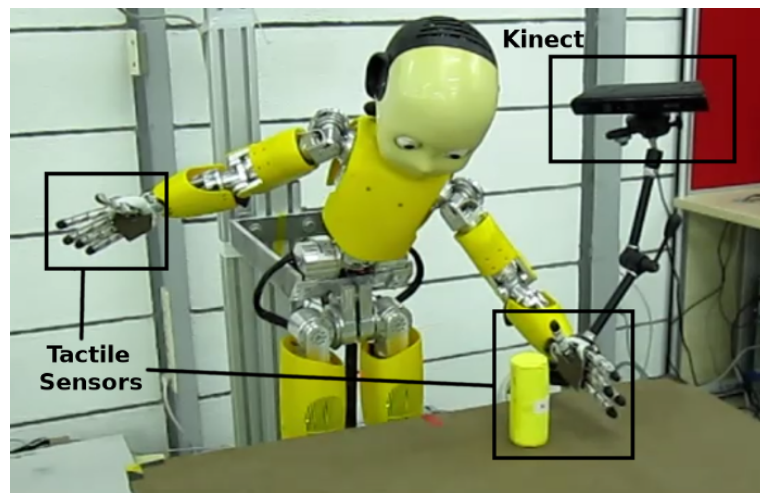


Figure 3.1: Experimental setup including the iCub robot platform and the Kinect RGB-D depth camera. [Adapted from [195] ©2015 IEEE.]

We conduct our experiments using iCub humanoid robot platform (Figure 3.1). It has 53 joints (DoF), six for head, 16 for each arm, three for torso, six for each leg. It also has tactile sensors in each fingertip to detect the degree of grasping an object and gather the relevant information about the hardness of it. Although iCub has two cameras, we utilize Kinect camera to get 3D information due to the calibration related problems. We have also external microphone to record the sound of objects.

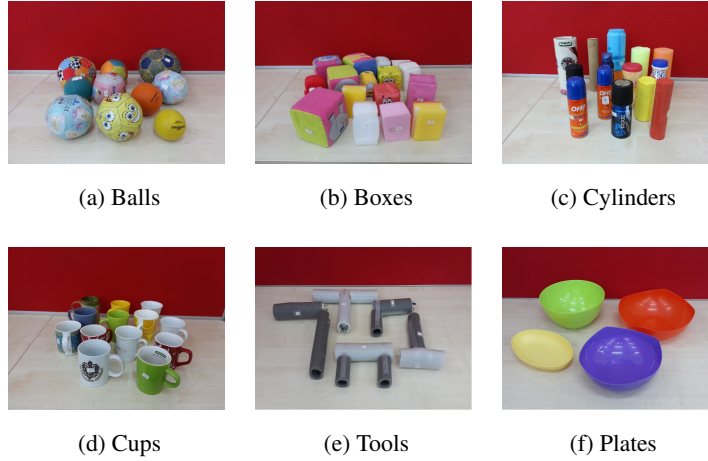


Figure 3.2: The noun concepts used in this study, and the associated objects. [Adapted from [6] ©2015 IEEE.]

### 3.1 Object Set

We have in total 60 objects, arbitrarily divided into a training (45 objects) and a testing set (15 objects). The training objects are classified with respect to their adjective categories ( $\mathbb{A} = \{hard \times soft, noisy \times silent, tall \times short, thin \times thick, round \times edgy\}$ ) and noun categories ( $\mathbb{N} = \{box, ball, cylinder, cup, plate, tool\}$ ). The grouped objects with respect to noun and adjective categories can be seen in Figures 3.2 and 3.3, respectively. In addition, we have a set of 13 verb (behavior) categories ( $\mathbb{V} = \{push\ left, push\ right, push\ forward, push\ backward, move\ left, move\ right, move\ forward, move\ backward, grasp, knock\ down, throw, drop, and\ shake\}$ ). iCub performs all or the partial set of behaviors on objects. We also define 6 spatial relationships that can exist between any two objects ( $\mathbb{S} = \{on, below, left\ of, right\ of, in\ front\ of, behind\}$ )

Each object in a noun category has its own adjectives, which shows the general characteristic of that noun category. For instance, all the objects which are classified as “box” are “edgy”, that is, there is strong correlation between them. The strength of the correlation, i.e. the co-occurrence information, between adjective and noun categories can be seen in Table 3.1.



Figure 3.3: The adjective concepts used in this study, and the associated objects. [Adapted from [6] ©2015 IEEE.]

Table 3.1: The frequencies of instances in the dataset in which specified noun and adjective pairs co-occur together (out of 60 objects in the dataset).

	Hard	Soft	Noisy	Silent	Tall	Short	Thin	Thick	Round	Edgy
Box	2	14	2	14	0	16	0	16	0	16
Ball	3	7	7	3	0	10	1	9	10	0
Cylinder	14	0	5	9	10	4	9	5	14	0
Cup	11	0	1	10	0	11	0	11	11	0
Tool	5	0	5	0	5	0	0	5	5	0
Plate	4	0	0	4	4	0	0	4	4	0

### 3.2 Behaviors

In our experiments, we have 13 behaviors ( $\mathcal{V}$ ) in total. To extract the relevant information over objects, iCub interacts with them by applying these behaviors. Although iCub can theoretically perform any one of these behaviors on each object, there is a limitation. Since plates and cups are fragile, we prevent iCub from performing “drop”, “shake”, “throw” and “knock down” behaviors on these type of objects. The applicable behaviors for objects with respect to their noun categories are shown in Table 3.2. Moreover, iCub can grasp all kinds of objects. it extracts all information for different modalities as the “grasp” behavior also includes the “shake” behavior, also. We have also two types of grasp, *top* and *side* grasps. The selection depends on the height and the depth of an object. If height is relatively more than depth, then iCub grasps an object from side, otherwise, it directly grasps from top. This selection is hard-coded.

Table 3.2: Possible applicable set of behaviors with respect to object categories.  $X \in \{Left, Right, Forward, Backward\}$ ; A: *Applicable*; N/A: *Not-Applicable*

	Push-X	Move-X	Drop	Grasp	Shake	Knock Down	Throw
Box	A	A	A	A	A	A	A
Ball	A	A	A	A	A	A	A
Cylinder	A	A	A	A	A	A	A
Cup	A	A	N/A	A	N/A	N/A	N/A
Tool	A	A	A	A	A	A	A
Plate	A	A	N/A	A	N/A	N/A	N/A

Table 3.3: The audio, haptic and visual features extracted from the interactions of the robot.

Feature Type	Feature	Position
Visual ( $\mathbf{e}_v$ )	Position: $(x, y, z)$	1-3
	Object dimensions: $(width, height, depth)$	4-6
	Normal zenith histogram bins	7-26
	Normal azimuth histogram bins	27-46
	Shape index histogram bins	47-66
Audio ( $\mathbf{e}_a$ )	13 bins of MFCC (max - min)	67-79
Haptic ( $\mathbf{e}_h$ )	Change for index finger	80
	Min values for index finger	81
	Max values for index finger	82
	Mean for index finger	83
	Variance for index finger	84
	Standard deviation for index finger	85
Proprioceptive ( $\mathbf{e}_p$ )	Change for index finger	86
	Min values for index finger	87
	Max values for index finger	88
	Mean for index finger	89
	Variance for index finger	90
	Standard deviation for index finger	91

### 3.3 Features and Data Collection

As we have mentioned in Section 3.2, we have multi-modal learning method, consisting of visual, auditory, haptic and proprioceptive information. Each of this information are kept as a set of features of our feature vector. The features that are extracted from each object is shown in Table 3.3.

The first 66 features are related the visual features ( $\mathbf{e}_v$ ) of an object including different properties. The visual features are the most crucial set of features, affecting the quality of the learning the concepts and prediction of the categories. Therefore, the extracted features have to represent the characteristic properties of an object. For

instance, a “cup” has to be placed with appropriate orientation such that the handle of it must be clearly discerned, since the one of the most prominent property that discriminate a “cup” from a “cylinder” is its handle. The first six features carry the information about the place and the dimensional properties of an object in world. The following 40 features are the histogram bins for azimuth (20-bin) and zenith (20-bin) angles of normal vectors. The normal vectors are extracted for all points, forming an object. The last 20 features are the histogram bins of the shape index information [?]. Shape index value is calculated as:

$$\mathcal{S} = \frac{\mathcal{K}_1 + \mathcal{K}_2}{\mathcal{K}_1 - \mathcal{K}_2}, \quad (3.1)$$

where  $\mathcal{K}_1$  and  $\mathcal{K}_2$  is the maximum and minimum principal curvatures, respectively. This set of features tells the surface property of an object, such as plane, saddle. etc. All of the visual features are extracted using Point Cloud Library (PCL) [196]. It is widely used in many studies requiring three dimensional information.

The following 13 features consist of auditory features ( $\mathbf{e}_a$ ) to determine whether an object has internal sound. We use MFCC (Mel-Frequency Cepstrum Coefficients) on the raw audio file. After executing the MFCC algorithm, we obtain a set of 13-features vector, varying with respect to the duration of the recording. To combine the relevant audio information into one vector, the maximum and minimum values are subtracted for each column.

Haptic and proprioceptive features ( $\mathbf{e}_h$  and  $\mathbf{e}_p$ ) are obtained using only the index finger of iCub. The other finger are used for the sake of convenient grasping operation. Haptic and proprioceptive values are collected throughout the action. As was the case in the auditory feature extraction, we also apply some statistical calculations on this set of values to obtain relevant information. The final and the initial values of instantaneous sensor reading for haptic and proprioceptive data are subtracted in column-wise and stored as features, Moreover, the minimum and maximum values are stored. The other features are directly extracted by finding the mean, variance, and the standard deviation of sensor value for index finger.

We have two types of feature vectors, namely *entity* and *effect*. The former one includes the clues about the adjective and noun categories of an object, whereas the latter one tells the information about applied behavior to an object. The entity feature

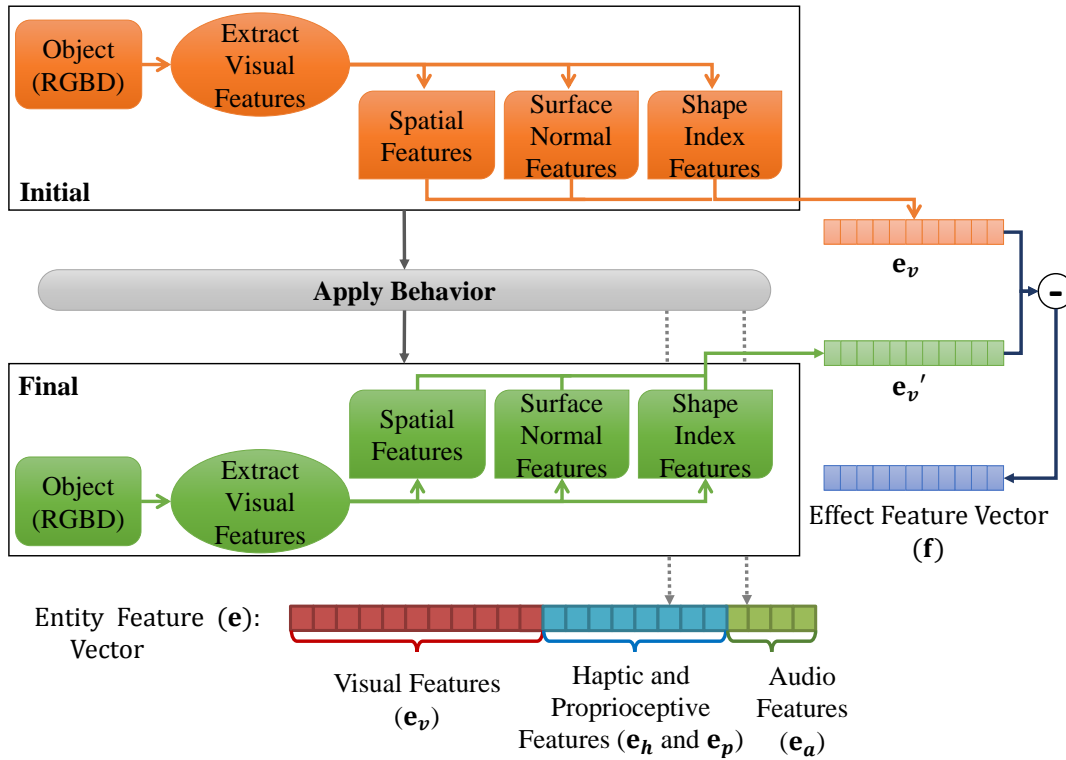


Figure 3.4: Extraction of entity features and effect visual features.  $e_v$  and  $e'_v$  are the visual features of an object before and after a behavior is applied.  $f = e'_v - e_v$  is the effect visual feature.  $e$  is the multi-modal feature incorporating visual, haptic, proprioceptive, and audio information. [Adapted from [6] ©2015 IEEE.]

vector consists of all features belonging to all modalities ( $e_v + e_a + e_h + e_p$ ), and extracted by following the below steps:

1. We put an object on the table, and iCub looks at it to extract visual features ( $e_v$ ).
2. iCub starts to grasp the object while storing the haptic and proprioceptive features ( $e_h$  and  $e_p$ ).
3. After grasping the object, iCub shakes it to collect auditory features ( $e_a$ )
4. iCub stops shaking and grasping; and finalizes collecting haptic, proprioceptive, and auditory features.
5. The final visual features ( $e'_v$ ) are collected.

The effect features only includes the visual information. As previously mentioned, they tell the verb of an action. Therefore, we collect visual features before and after

the behavior, and subtract the final features ( $\mathbf{e}'_v$ ) from the initial features ( $\mathbf{e}_v$ ), which is depicted in Figure 3.4.





## CHAPTER 4

### THE CONCEPT WEB

In this thesis, we propose a densely connected web representation for concepts (Figure 4.1). The proposed architecture is a Markov Random Field, whose structure of connectedness is learned from the training data. A newly encountered object is then examined in this web for association with known concepts. This part of the thesis has been conducted in collaboration with Güner Orhan, and disseminated in IEEE Transactions on Autonomous Mental Development [6].

#### 4.1 Individual Concepts

First we describe how we conceptualize noun, adjective and verb categories, and how they can be used individually to predict the concepts of new instances.

##### 4.1.1 Conceptualization of a Category

We define concepts by their prototypes (see Section 2.1 for theories of concepts) following our previous work on prototype-based conceptualization of verbs, nouns and adjectives [38, 197]. In this approach, given a category of exemplars where each exemplar is represented by a fixed-length feature vector, we look at the distribution of features in each single dimensionality. When we look at the mean and variance of each dimension of the features extracted for the exemplars in a category, there are four possible cases in general:

1. Its mean value over all instances is high compared to the means of other fea-

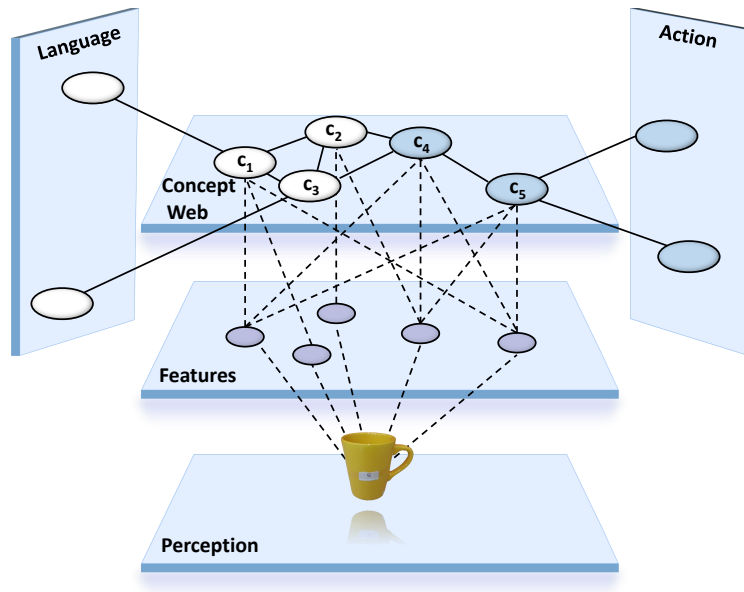


Figure 4.1: The schematic presentation of the concept web, which connected related concepts to each other and to their counterparts in the language, action, and perception spaces. Information can flow in from the perception space, through a feature extraction mid-level, or from the language and action spaces as well. A number of nodes are randomly illustrated with white color to exemplify active concepts. [Adapted from [6] ©2015 IEEE.]

tures, then we can deduce that this feature dimension correlates positively for this category, and therefore labeled with a ‘+’ sign in the prototype.

2. If the mean value is small, and variance is again small, then this feature category is negatively correlated, and labeled with a ‘-’.
3. If the variance of the feature is too high to be meaningful, then this feature category cannot represent our category. Therefore, it is labeled as ‘\*’ meaning highly variant and inconsistent.
4. For behaviors only, the fourth case includes features that show no change before and after the application of the behavior. This case is marked with a ‘0’, and used only for verb prototypes.

In effect, for the nouns and adjectives, dimensions marked with ‘+’ and ‘-’ symbols refer to the characteristic properties of these concepts, which are independent of the

---

**Algorithm 1:** Derivation of a prototype from the exemplars of a category  
 [Adapted from [6, 38].]

---

**for all** concepts  $c \in \mathbb{C}$  **do**

**for all** feature dimensions  $d$  **do**

    Compute the mean  $\mu_{cd}$ :

$$\mu_{cd} = \frac{1}{N} \sum_{i \in \mathbb{I}(c)} i_d, \quad (4.1)$$

    where  $\mathbb{I}(c)$  is the set of training instances of concept  $c$ , with cardinality  $N = |\mathbb{I}(c)|$ , and  $i_d$  is the  $d^{\text{th}}$  feature of instance  $i$ .

    Compute the variance  $\sigma_{cd}^2$ :

$$\sigma_{cd}^2 = \frac{1}{N} \sum_{i \in \mathbb{I}(c)} (i_d - \mu_{cd})^2. \quad (4.2)$$

**end for**

  Concatenate  $\mu_{cd}$ 's and  $\sigma_{cd}^2$ 's to obtain the vectors  $\mu_c$  and  $\sigma_c^2$ .

**end for**

**for all** concepts  $c \in \mathbb{C}$  **do**

  Apply Robust Neural Growing Gas algorithm in  $\mu_c \times \sigma_c^2$  space:

**if**  $c \in \mathbb{N} \cup \mathbb{A}$  **then**

    - Manually assign one of the labels '+', '-', or '\*' to the dimension  $d$ , considering the cluster that  $d$  falls into:

**if** cluster is high on  $\mu$  axis and low on  $\sigma^2$  axis **then**

      assign '+' to  $d$

**else if** cluster is low on both  $\mu$  and  $\sigma^2$  axes **then**

      assign '-' to  $d$

**else if** cluster is high on  $\sigma^2$  axis **then**

      assign '\*' to  $d$

**end if**

**else**

    - Manually assign one of the labels '+', '-', '\*', or '0' to the dimension  $d$ , considering the cluster that  $d$  falls into:

**if** cluster is high on  $\mu$  axis and low on  $\sigma^2$  axis **then**

      assign '+' to  $d$

**else if** cluster is low on both  $\mu$  and  $\sigma^2$  axes **then**

      assign '-' to  $d$

**else if** cluster is close to 0 on  $\mu$  axis and low on  $\sigma^2$  axis **then**

      assign '0' to  $d$

**else if** cluster is high on  $\sigma^2$  axis **then**

      assign '\*' to  $d$

**end if**

**end if**

**end for**

---

behavior applied. On the other hand, ‘+’ and ‘-’ for the verb concepts which are extracted from the visual effect features ( $\mathbf{f}$ ) refer to the characteristically changed properties of objects through the associated behaviors, such as increased x-position after a *push (forward)* behavior.

To obtain the prototypes for each category, first the objects and the effects of the interactions are labeled with noun, adjective and verb labels using cross-situational learning. Then, on each category, we use the Robust Growing Neural Gas (RGNG) algorithm [198] for clustering every feature dimension into one of ‘+’, ‘-’, ‘\*’, ‘0’ classes. The exact procedure is depicted in Algorithm 1, which is applied on  $\mathbf{e}$  in the case of nouns and adjectives, and  $\mathbf{f}$  in the case of verbs. Eventually, we obtain 29 prototypes, one prototype for each category or concept: 6 for nouns, 10 for adjectives, and 13 for verbs (for *push* and *move* behaviors, a separate prototype is obtained for each possible argument, due to different features being relevant in each case). The resulting prototypes are shown in Table 4.1. The process is visually explained in Figure 4.2.

#### 4.1.2 Category Prediction from Features Only

The prediction procedure takes as input the above prototypes of concepts and the feature vector ( $\mathbf{e}$  or  $\mathbf{f}$ ) of a new object or an interaction (for the sake of clarity, in the remainder of this subsection, the methodology will be described for an object although the same procedures are applied for both). For evaluating membership for a concept, only meaningful features (labeled with ‘+’, ‘-’ or ‘0’ in the corresponding prototype) are considered. On these meaningful dimensions, an Euclidean distance to the mean values of the concept’s prototype is calculated as follows:

$$D(c, o) = \frac{1}{|\mathcal{R}_c \setminus \mathcal{R}_c^*|} \sum_{i \in \mathcal{R}_c \setminus \mathcal{R}_c^*} i \mathbf{e}_o - i \mu_c, \quad (4.3)$$

where  $o$  is the new object;  $\mathcal{R}_c^*$  is the set of indices that are ‘\*’-signed (*i.e.*, inconsistent) for concept  $c$ ;  $\mathbf{e}_o^i$  is the  $i^{th}$  feature of test object  $o$ , and  $\mu_c$  is the mean feature vector of objects in concept  $c$ .

$D(c, o)$  is the closeness of the new object to the selected concept. We can convert it



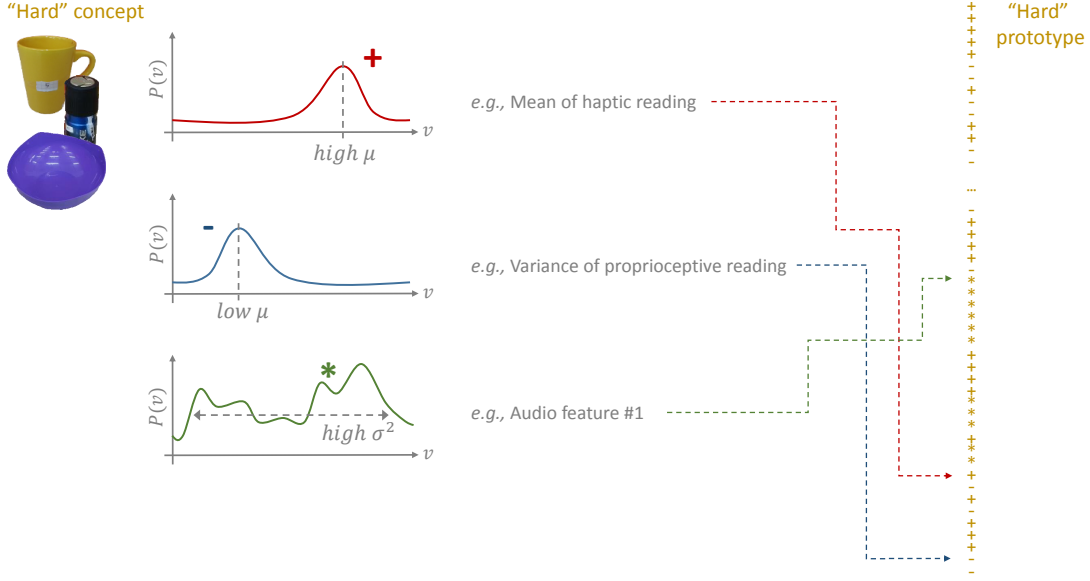


Figure 4.2: Schematic visualization of the extraction of a concept prototype. If a feature has a consistently high contribution, marked with a high mean and low variance distribution, it is indicated with a ‘+’ sign. Those with a consistently low contribution, marked with a low mean and low variance distribution, are assigned a ‘-’ sign, whereas those with a high variance are marked with a ‘\*’ to indicate inconsistent contribution. Sample features are illustrated for the *hard* concept. [Adapted from [6] ©2015 IEEE.]

into a probability estimate as follows:

$$s_{perc}(c, o) = \frac{\prod_{r \in \mathbb{C} \setminus \{c\}} D(r, o)}{\sum_{r \in \mathbb{C}} \left( \prod_{r_t \in \mathbb{C} \setminus \{r\}} D(r_t, o) \right)}, \quad (4.4)$$

where  $\mathbb{C}$  is the set of possible group of concepts. These groups can be nouns ( $\mathbb{N} = \{box, ball, cylinder, cup, tool, plate\}$ ), adjective pairs (e.g.,  $\mathbb{A}_{pair} \in \mathbb{A} = \{hard, soft\}$ ), and verbs ( $\mathbb{V} = \{push\ left, push\ right, push\ forward, push\ backward, move\ left, move\ right, move\ forward, move\ backward, grasp, knock\ down, throw, drop, shake\}$ ), separately. Equation 4.4 defines the probability that a new instance  $o$  belongs to a concept  $c$ . Note that  $s_{perc}(\cdot, \cdot)$  is based only on the features extracted from the instance and it does not use the co-occurrences between concepts.

## 4.2 A Probabilistic Web of Concepts

In the previous section, we described the conceptualization of individual categories. In this section, we present how we represent the concept web in a probabilistic model, namely, Markov Random Field. Each node of the constructed Markov Random Field corresponds to a concept (noun, adjective, verb, or superordinate) in our web.

### 4.2.1 Building a Web from Individual Concepts

Let us call  $\mathbb{C}$  be the set of all concepts; *i.e.*,  $\mathbb{C} = \text{N} \cup \text{A} \cup \text{V}$ , the concatenation of noun, adjective and verb concepts. Let us denote  $W$  to be the concept web constructed from the interactions of the robot. The web  $W$  can be represented as a graph  $G(\mathbb{C}, \mathbb{E})$ , where each  $c \in \mathbb{C}$  is treated as a node in  $W$ .

The edges in the web  $\mathbb{E}$  are established based on the co-occurrences of the concepts. Namely, an edge between concept  $c_i$  and  $c_j$  (*i.e.*,  $\mathbb{E}(c_i, c_j)$ ) is placed in the web if concepts  $c_i$  and  $c_j$  have co-occurred in an interaction. The web constructed from the interactions is provided in Figure 4.3.

$$s_{\text{cooc}}(c_1, c_2) = \frac{|S_{c_1}^{c_2}|}{|S_{c_1}|}. \quad (4.5)$$

### 4.2.2 Concept Web as a Markov Random Field

Markov Random Field (MRF) [200] is a probabilistic graphical model widely used for defining constraints on and between entities in a problem. The entities are represented as nodes and the constraints between the entities are incorporated by the edges connecting them. MRF follows the Markovian property that only conditions the state of a node only on the neighboring nodes (Figure 4.4a). Due to these representational constraints, all probabilistic functions are defined over *maximal cliques*. A *clique* is a subset of nodes that are connected to each other directly, and a *maximal clique* is a clique with the highest number of nodes possible (Figure 4.4b).

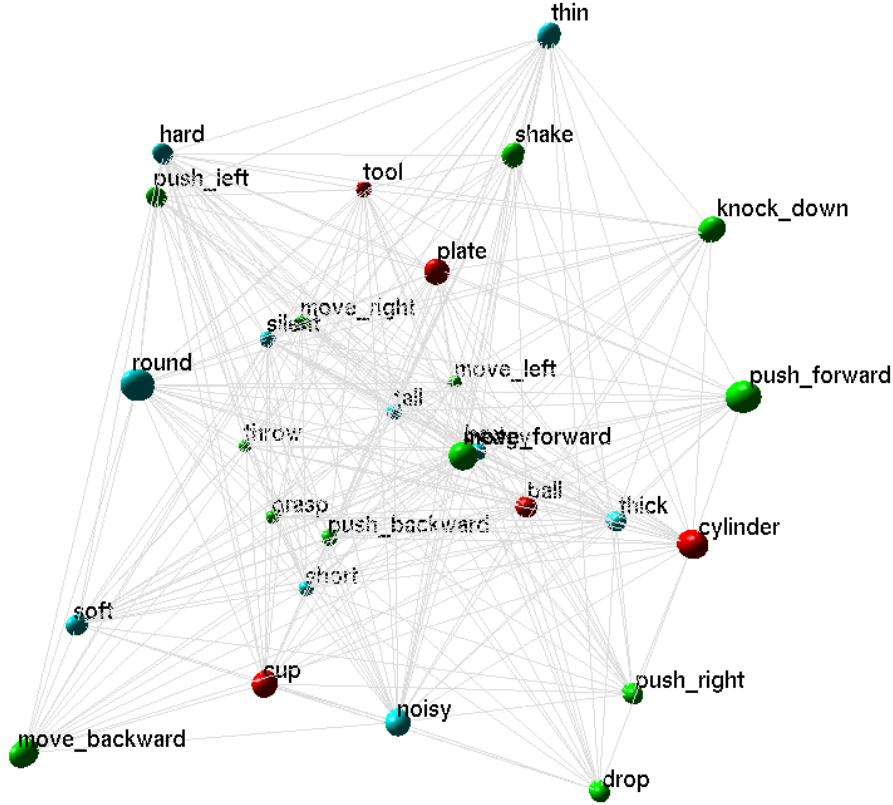


Figure 4.3: A snapshot of the concept web iCub has constructed. Connections between related concepts are denoted with gray links. Noun concepts are indicated with red, adjective concepts with blue, verb concepts with green, and superordinate concepts with cyan. The graph is created using Ubigraph graph visualization library [199]. [Adapted from [6] ©2015 IEEE. Best viewed in color.]

MRF effectively models the following joint probability distribution:

$$P(\omega) := \frac{1}{Z} \exp(-U(\omega)), \quad (4.6)$$

where  $\omega$  is a possible configuration of the web  $W$ , and  $U(\omega)$  is the energy function of the MRF given a configuration  $\omega$ , calculated as:

$$\begin{aligned} U(\omega) &:= U_{data}(\omega) + U_{smooth}(\omega) \\ &:= \sum_{c \in \omega} \psi_c(c) + \sum_{\mathcal{K} \in \mathbb{K}} \psi_{\mathcal{K}}(\mathcal{K}, \omega), \end{aligned} \quad (4.7)$$

with  $\mathbb{K}$  denoting the set of all cliques,  $c$  is the set of all active concepts in the given configuration  $\omega$ , and  $\psi_c$  is the potential of each active concept  $c$ , defined by:

$$\psi_c(c) := D(c, \mathbf{x}), \quad (4.8)$$



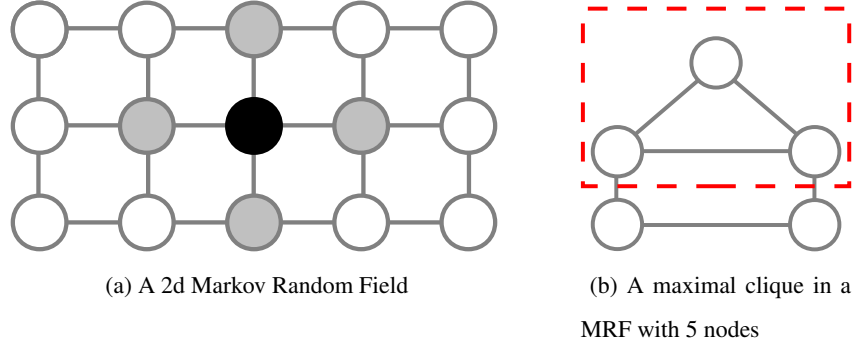


Figure 4.4: (a) A sample 2D Markov Random Field. The Markovian property holds in Markov Random Fields, by which a random variable (*i.e.*, the black node), given its immediate neighbors (the gray nodes), is independent of all other random variables. (b) A maximal clique (with 3 nodes) is indicated in an MRF with 5 nodes.

with  $o$  being the current observation, and  $D(c, o)$  its distance to the active concept  $c$  (Equation 4.3).  $\psi_K(\mathcal{K}, \omega)$  is the potential of clique  $\mathcal{K}$  in configuration  $\omega$ :

$$\psi_K(\mathcal{K}, \omega) := \mathcal{V}_{\mathcal{K}}(\mathbf{x}_{\mathcal{K}}) := \sum_{x_i \in \mathbf{x}_{\mathcal{K}}} |val(x_i) - E(x_i | \mathbf{x}_{\mathcal{K}-i})| \quad (4.9)$$

where  $\mathcal{V}_{\mathcal{K}}(\mathbf{x}_{\mathcal{K}})$  is defined as an (abused) shorthand notation for the potential of a clique node consisting of active variables  $\mathbf{x}_{\mathcal{K}}$ ,  $x_i$  is the  $i^{th}$  variable in the clique,  $\mathbf{x}_{\mathcal{K}-i}$  are the variables in the clique excluding the  $i^{th}$  variable,  $val(x_i)$  is the current value assignment of the variable  $x_i$ ,  $|\cdot|$  is the absolute value function,  $E(\cdot)$  is the expected value function, and  $E(x_i | \mathbf{x}_{\mathcal{K}-i})$  is the expected value of the  $i^{th}$  variable given the values of the remaining variables in the clique.  $\psi_K(\mathcal{K}, \omega)$  therefore tries to minimize the difference between the values of the clique variables and their expected values given the rest of the variables in the clique.  $Z$  denotes the partition function:

$$Z := \sum_{\omega \in \Omega} \exp \left( - \sum_{c \in \omega} \psi_c(c) - \sum_{\mathcal{K} \in \mathbb{K}} \psi_K(\mathcal{K}, \omega) \right), \quad (4.10)$$

where  $\Omega$  is the set of all possible configurations, and  $\mathbb{K}$  is the set of all cliques. Figure 4.5 visualizes the concept and clique potentials schematically.

### 4.2.3 Belief Propagation in MRF

The potential values in MRF model demonstrate the correlation between two connected nodes. In belief propagation methods, this correlation is thought as a *message*

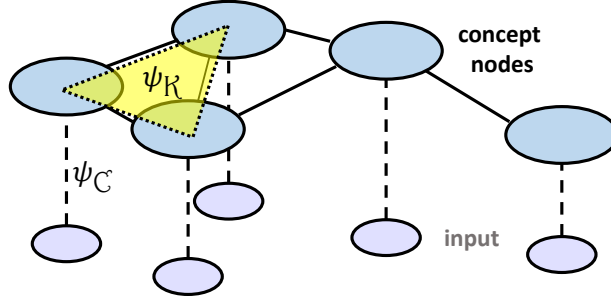


Figure 4.5: A schematic representation of MRF modeling of the concept web. Initial predictions about the concepts are used to initialize concept node probability values. Conformance to initially predicted values are maintained by minimizing the sum of unary potential functions  $\psi_C$ . Meanwhile, clique potentials are initialized from the cooccurrence information from the training data, and conformance to the cooccurrence information is maintained through minimizing the sum of clique potentials  $\psi_K$ . [Adapted from [6] ©2015 IEEE.]

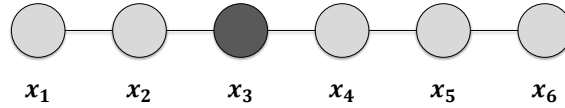


Figure 4.6: Sample Markov Random Field chain of variable nodes

from one node to the an adjacent one. To demonstrate belief propagation, let us calculate the marginal probability distribution over a node ( $x_3$ ) for the MRF in Figure 4.6:

$$p(x_3) = \frac{1}{Z} \sum_{x_2} \psi(x_2, x_3) \sum_{x_1} \psi(x_1, x_2) \sum_{x_4} \psi(x_3, x_4) \sum_{x_5} \psi(x_4, x_5) \sum_{x_6} \psi(x_5, x_6). \quad (4.11)$$

If we treat the terms in Equation 4.11 as messages from the adjacent nodes of query node  $x_3$ , then we can re-formulate it, merging these messages as follows:

$$p(x_3) = \frac{1}{Z} \left[ \sum_{x_1, x_2} \prod_{i=1}^2 \psi(x_i, x_{i+1}) \right] \cdot \left[ \sum_{x_4, x_5, x_6} \prod_{i=3}^5 \psi(x_i, x_{i+1}) \right],$$

$$p(x_3) = \frac{1}{Z} \mu_{x_2}(x_3) \mu_{x_4}(x_3), \quad (4.12)$$

where  $\mu_{x_2}(x_3)$  is the message to  $x_3$  from  $x_2$ , and  $\mu_{x_4}(x_3)$  is the message for  $x_3$  from  $x_4$ .

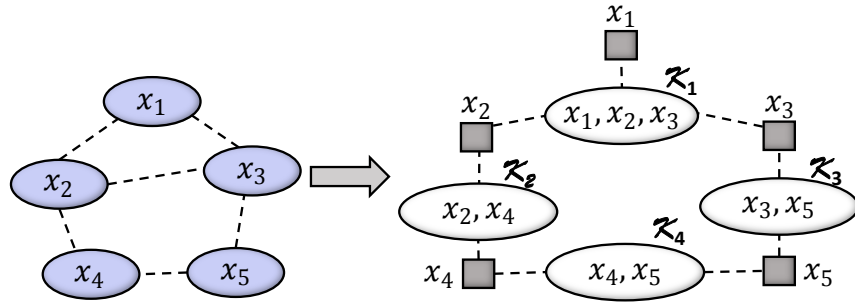


Figure 4.7: The conversion of a MRF graph as a factor graph, as input to the Loopy Belief Propagation. [Adapted from [6] ©2015 IEEE.]

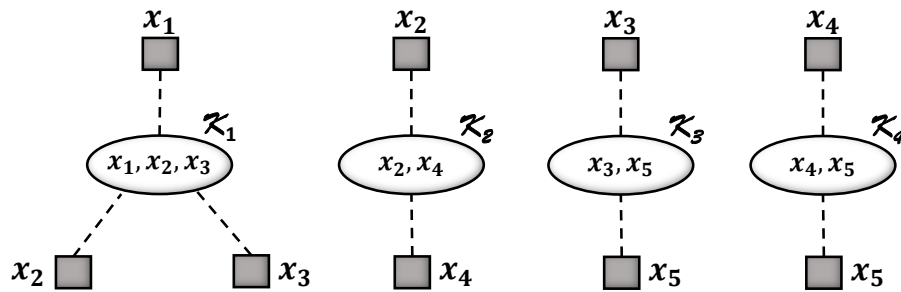


Figure 4.8: Divided subtrees of the graph in Figure 4.7. [Adapted from [6] ©2015 IEEE.]

#### 4.2.4 Inferences in Concept Web Using Loopy Belief Propagation

Our concept web  $W$  is a cyclic graph by definition, and therefore, making exact inferences given observations is not possible. For such problems, approximate solutions are used, and a widely-used method for this task is Loopy Belief Propagation (LBP) [201], which iteratively updates the influence of one variable (*i.e.*, concept) on another until convergence. The influence of one variable on another is called *message*, and this process is called *message passing*.

LBP re-factorizes the graph into separator nodes and clique nodes - see Figure 4.7 for an example. Clique nodes are shown as elliptic nodes, whereas separator nodes are symbolized with square nodes. Separator nodes are in fact the concepts in the web whereas the clique nodes represent the potential of a clique as a single node.

Message passing procedure in LBP differs in many ways when compared with standard belief propagation. For instance, the graph is divided into sub-trees each of which includes one clique node and connected separator nodes to it (Figure 4.8). After extracting the subtrees, LBP performs the following until convergence:

1. **Update Clique Potentials:** Updating the clique potentials can be thought as message passing from connected separator node to the clique node. Therefore, we can compute the new potentials by multiplying the potentials of separator nodes with the previous value of potentials in the clique node:

$$\mathcal{V}_{\mathcal{K}}^*(\mathbf{x}_{\mathcal{K}}) = \mathcal{V}_{\mathcal{K}}(\mathbf{x}_{\mathcal{K}}) \prod_{x_m \in ne(\mathbf{x}_{\mathcal{K}})} \phi_m(x_m), \quad (4.13)$$

where  $\mathbf{x}_{\mathcal{K}}$  is the set of variables in clique node  $\mathcal{K}$ ,  $\mathcal{V}_{\mathcal{K}}(\mathbf{x}_{\mathcal{K}})$  is the previous potential of the clique node, and  $\mathcal{V}_{\mathcal{K}}^*(\mathbf{x}_{\mathcal{K}})$  is its updated potential.

2. **Update Separator Potentials:** After updating the clique potentials, we apply the message passing in the reverse direction. This time, updating the separator potentials is different from updating the clique potentials in that the message from the updated clique node to any one of the connected separator nodes is calculated by summation of the potentials of the clique nodes except the separator node:

$$\mu_{\mathcal{K}^* \rightarrow x_m}(x_m) = \sum_{\mathbf{x}_n \in \mathbf{x}_{\mathcal{K}} \setminus x_m} \mathcal{V}_{\mathcal{K}}^*(\mathbf{x}_n). \quad (4.14)$$

If the potential of separator node  $x_m$  is updated previously, the new potential value is the multiplication of the previous potential of node with the division of new message from clique node to the previous one:

$$\phi_s^*(x_m) = \phi_s(x_m) \frac{\mu_{\mathcal{K}^* \rightarrow x_m}(x_m)}{\mu_{\mathcal{K} \rightarrow x_m}(x_m)}. \quad (4.15)$$

Otherwise, it is directly set to the new value:

$$\phi_s^*(x_m) = \phi_s(x_m) \mu_{\mathcal{K}^* \rightarrow x_m}(x_m). \quad (4.16)$$

where  $\phi_s(x_m)$  is the previous potential of the separator node  $x_m$ , and  $\phi_s^*(x_m)$  is the updated potential.

3. **Iteration:** The previous two steps are iterated for all clique nodes and their separator nodes until the change in the potentials is less than a threshold.

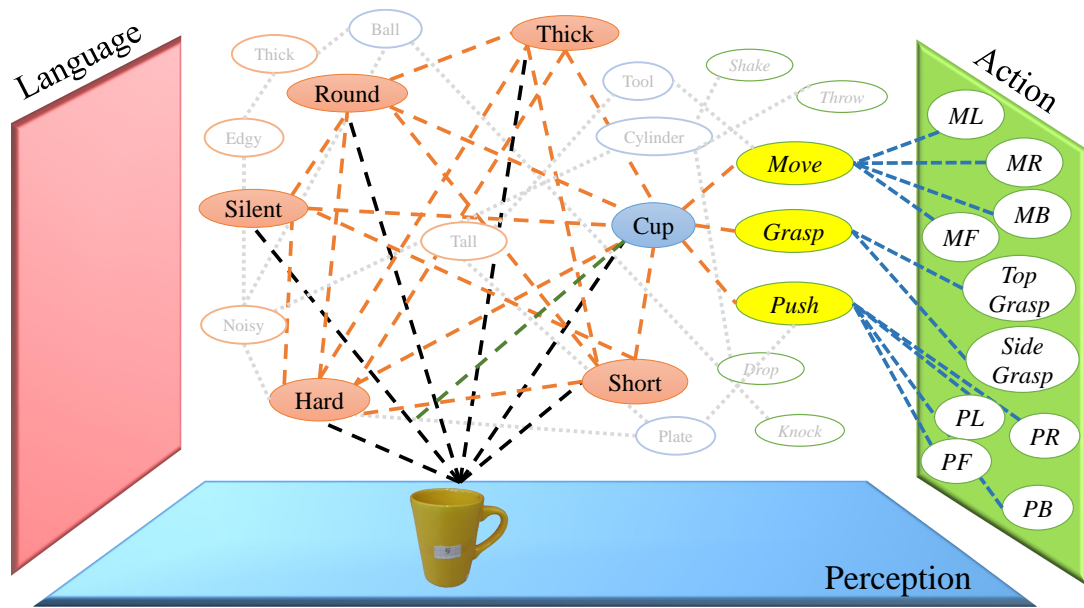


Figure 4.9: Schematic presentation of Scenario 1. iCub is presented with a cup and expected to predict the type and properties of the object, as well as what kind of behaviors can be applied on this object. ML: *Move Left*, MR: *Move Right*, MB: *Move Backward*, MF: *Move Forward*, PL: *Push Left*, PR: *Push Right*, PB: *Push Backward*, PF: *Push Forward*. [Adapted from [6] ©2015 IEEE. Best viewed in color.]

### 4.3 Experiments and Results

The concept web built from the interactions of iCub is provided in Figure 4.3. In the rest of the section, we demonstrate how this concept web can be helpful for a humanoid robot in three different scenarios: (i) a scenario where the relevant concepts in the web are activated based on perception of an object, (ii) a scenario where the relevant concepts in the web are activated based on a partial perception of an object, with an intended action in mind, and (iii) a scenario where the activation is due to only an intended action in mind, without any specific object singled out.

#### 4.3.1 Scenario 1: Perception-driven activation of concepts in the web

In this scenario, iCub is presented with an object, allowed to interact freely with it, and expected to guess what kind of an object it is. It needs to guess both the type of the object (the noun), and its properties (the adjectives). Furthermore, iCub is expected

to predict the verbs (the behaviors) that are possibly applicable on the object.

On perceiving the object, iCub first records its visual data through Kinect, then grasps the object to collect haptic, proprioceptive and auditory data. The related features are combined in the entity feature vector  $\mathbf{e}$ , and compared against the previously extracted prototypes of nouns and adjectives in order to determine the categories of the object. These predictions gives us an a priori guess about the membership probabilities. These a priori probabilities are in turn used to initialize the activations of the nodes in the concept web (Note that the concept web architecture, *i.e.*, the connections between the nodes and their joint probabilities, have been determined previously using the training data). Nodes of the unobserved concepts are initialized to 0.5 probability in an unbiased manner. Afterwards the concept web is allowed to propagate its activations. Once convergence is achieved, we expect iCub to (1) refine its initial guesses about the noun and adjective categories of the object, possibly correcting wrong ones, and (2) determine which behaviors are applicable to the object, by propagating activation through the noun and adjective concepts to connected verb concepts.

A sample scenario is presented in Figure 4.9. A cup is presented to iCub in this case, which iCub correctly detects as a *cup*, and as being *round*, *short*, *hard*, *silent*, and *thick*. It also predicts that it can apply the *grasp* behavior on the object, as well as *move* and *push* behaviors. The rest of the behaviors (*knock down*, *shake*, *throw* and *drop*) are not found applicable on the object.

This scenario depicts the activation of the concept web in a similar fashion to the canonical neurons in the F5 area of monkey brain [202–204]. These “visiomotor” neurons are known to fire selectively to certain actions, as well as to the *presentation* of an object to which this action can be potentially applied. This raw recognition of possible action (without necessary recognition of the object *per se*) has been accepted as one of the neurological mechanisms of affordances. The context web approach also results in a similar predictive activations in the conceptual representations of the applicable behaviors.

We now apply this scenario systematically to present quantitative results in Table 4.2. Six arbitrarily selected objects, one from each noun category, are presented to the iCub, which is then expected to guess its noun category and adjective categories,

and the applicable behaviors. To demonstrate the effectiveness of the approach, the predictions made using the concept web are compared to the prototype-only initial predictions, as well as to that of Support Vector Machines, and Support Vector Machines enhanced with ReliefF [205] feature selection. 6 SVMs are trained separately for both the ReliefF and the no-ReliefF cases, among which 5 of them chooses between one of two dichotomous adjectives (*hard* vs. *soft*, *edgy* vs. *round*), and one is responsible with selecting a noun concept (*ball* vs. *box* vs. *cup* vs. *cylinder* vs. *plate* vs. *tool*). Both SVM and SVM+ReliefF cases achieved more than 90% training accuracy, with the exception of the no-ReliefF *noisy* vs. *silent* case with a training accuracy of 82%. In the ReliefF feature selection case, features with weights  $> 0.1$  are accepted, out of a range of  $[-1, 1]$ .







Table 4.2 shows that the concept web predictions are significantly enhanced for both the nouns and the adjectives, as compared to the baseline methods. It is able to correct the wrong predictions of the baselines; whereas for already correctly predicted cases, the prediction confidences are increased. This result is in line with our previous analysis in [28, 206], in which we conclude that an approach which cannot utilize the dependency information between concepts, such as the SVM, SVM+ReliefF, and individual prototypes approaches, would significantly be outperformed by those which can. Therefore, the effectiveness of the web-based approach is directly due to its ability of capturing second-order conceptual relations, which is ignored by the other methods.

### 4.3.2 Scenario 2: Interaction-driven activation of concepts in the web

In the second scenario (Figure 4.10), the human partner not only presents iCub with an unknown object, but also commands a single, certain action to be performed. This time, the activation spreads to the concept web from two different entry points.

In the first path, iCub looks at the object, and collects its visual features in a partial entity feature vector (composed of features [1-67]). Since it is not allowed to grasp the object to investigate it (grasping may not be the required action), haptic, proprioceptive, or auditory features are not available perceptually. This partial entity feature vector ( $\mathbf{e}_v$ ) is compared against the noun and adjective prototypes to predict the cor-

Table 4.2: A comparison of noun and adjective predictions using the concept web. One object from each noun category is used for demonstration. Images depict RGB-D images from the Kinect sensor. Columns 2 and 3: SVM predictions, Columns 4 and 5: ReliefF feature selection + SVM, Columns 6 and 7: Prototype-only predictions, Columns 8 and 9: Concept web estimations. Parentheses: Prediction confidences. **Bold text: Correct decisions. Stroked text: Wrong decisions.** [Adapted from [6] ©2015 IEEE.]

Object	SVM		SVM+Relieff		Prototypes		Concept Web	
	Nouns	Adjectives	Nouns	Adjectives	Nouns	Adjectives	Nouns	Adjectives
	ball (7%) box (12%) <b>cup (58%)</b> cylinder (14%) plate (4%) tool (5%)	edgy (9%) <b>hard (97%)</b> noisy (12%) <b>short (83%)</b> <b>thick (91%)</b>	ball (3%) box (5%) <b>cup (75%)</b> cylinder (9%) plate (4%) tool (4%)	edgy (4%) <b>hard (98%)</b> noisy (23%) <b>short (94%)</b> <b>thick (91%)</b>	ball (8%) box (13%) <b>cup (43%)</b> cylinder (20%) plate (9%) tool (7%)	edgy (34%) <b>hard (71%)</b> noisy (42%) <b>short (54%)</b> thick (47%)	ball (0%) box (0%) <b>cup (100%)</b> cylinder (0%) plate (0%) tool (0%)	edgy (0%) <b>hard (100%)</b> noisy (0%) <b>short (100%)</b> <b>thick (100%)</b>
	ball (52%) box (10%) cup (12%) cylinder (10%) plate (9%) tool (7%)	edgy (4%) hard (47%) noisy (39%) <b>short (89%)</b> <b>thick (98%)</b>	ball (58%) box (6%) cup (6%) cylinder (10%) plate (12%) tool (8%)	edgy (3%) <b>hard (86%)</b> noisy (19%) <b>short (97%)</b> <b>thick (97%)</b>	ball (26%) box (18%) cup (14%) cylinder (17%) plate (13%) tool (10%)	edgy (42%) <b>hard (53%)</b> noisy (42%) <b>short (64%)</b> <b>thick (57%)</b>	ball (100%) box (0%) cup (0%) cylinder (0%) plate (0%) tool (0%)	edgy (0%) <b>hard (100%)</b> noisy (0%) <b>short (100%)</b> <b>thick (100%)</b>
	ball (6%) box (17%) cup (18%) <b>cylinder (48%)</b> plate (6%) tool (5%)	edgy (7%) <b>hard (97%)</b> noisy (23%) <b>short (36%)</b> thick (33%)	ball (5%) box (9%) cup (13%) <b>cylinder (66%)</b> plate (4%) tool (3%)	edgy (3%) <b>hard (97%)</b> noisy (20%) <b>short (47%)</b> thick (17%)	ball (10%) box (14%) cup (18%) <b>cylinder (41%)</b> plate (9%) tool (8%)	edgy (34%) <b>hard (72%)</b> noisy (30%) <b>short (39%)</b> thick (25%)	ball (0%) box (0%) cup (0%) <b>cylinder (100%)</b> plate (0%) tool (0%)	edgy (0%) <b>hard (100%)</b> noisy (0%) <b>short (100%)</b> thick (0%)
	ball (7%) <b>box (70%)</b> cup (7%) cylinder (10%) plate (3%) tool (3%)	edgy (96%) hard (2%) noisy (16%) <b>short (96%)</b> <b>thick (98%)</b>	ball (4%) <b>box (85%)</b> cup (2%) cylinder (2%) plate (3%) tool (4%)	edgy (97%) hard (0%) noisy (16%) <b>short (97%)</b> <b>thick (100%)</b>	ball (42%) <b>box (42%)</b> cup (11%) cylinder (12%) plate (10%) tool (8%)	edgy (64%) hard (34%) noisy (30%) <b>short (59%)</b> <b>thick (63%)</b>	ball (0%) <b>box (100%)</b> cup (0%) cylinder (0%) plate (0%) tool (0%)	edgy (100%) hard (0%) noisy (0%) <b>short (100%)</b> <b>thick (100%)</b>
	ball (7%) box (18%) cup (13%) <b>cylinder (28%)</b> plate (23%) tool (11%)	edgy (14%) <b>hard (94%)</b> <b>noisy (100%)</b> <b>short (2%)</b> <b>thick (98%)</b>	ball (4%) box (13%) cup (10%) cylinder (17%) <b>plate (46%)</b> tool (10%)	edgy (20%) <b>hard (93%)</b> <b>noisy (100%)</b> <b>short (4%)</b> <b>thick (98%)</b>	ball (11%) box (13%) cup (15%) cylinder (18%) plate (11%) tool (32%)	edgy (48%) <b>hard (55%)</b> <b>noisy (61%)</b> <b>short (39%)</b> <b>thick (57%)</b>	ball (0%) box (0%) cup (0%) cylinder (0%) plate (0%) tool (100%)	edgy (0%) <b>hard (100%)</b> <b>noisy (100%)</b> <b>short (0%)</b> <b>thick (100%)</b>
	ball (12%) box (18%) cup (14%) cylinder (10%) <b>plate (39%)</b> tool (7%)	edgy (13%) <b>hard (62%)</b> noisy (33%) <b>short (52%)</b> <b>thick (99%)</b>	ball (10%) box (15%) cup (8%) cylinder (6%) <b>plate (45%)</b> tool (16%)	edgy (4%) <b>hard (74%)</b> noisy (20%) <b>short (78%)</b> <b>thick (100%)</b>	ball (15%) box (18%) cylinder (17%) <b>plate (21%)</b> tool (13%)	edgy (44%) <b>hard (51%)</b> noisy (44%) <b>short (42%)</b> <b>thick (53%)</b>	ball (0%) box (0%) cup (0%) cylinder (0%) <b>plate (100%)</b> tool (0%)	edgy (0%) <b>hard (100%)</b> noisy (0%) <b>short (0%)</b> <b>thick (100%)</b>



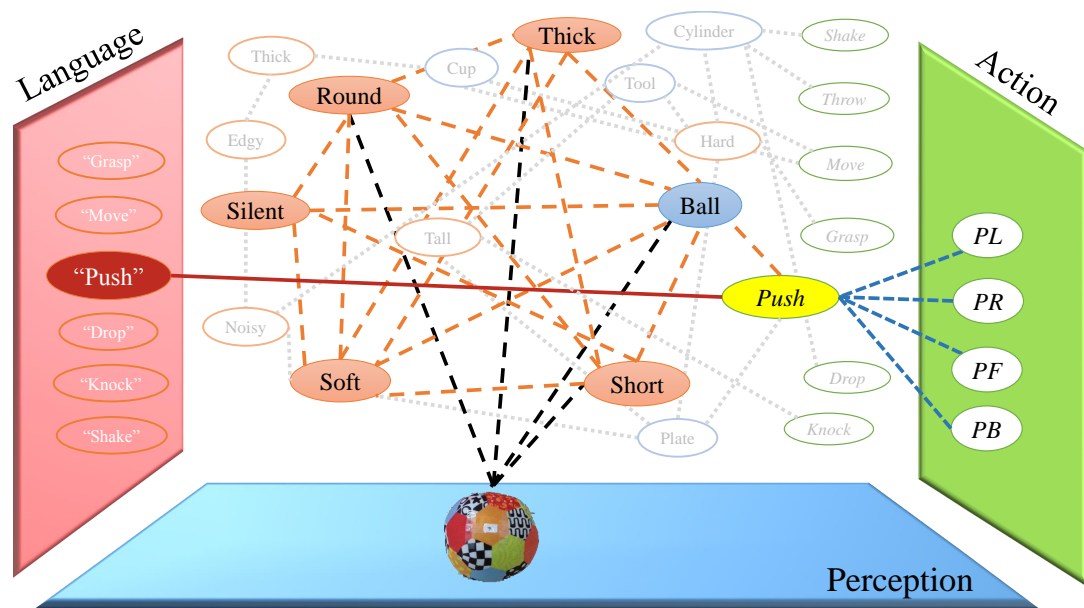



Figure 4.10: Schematic presentation of Scenario 2. iCub is presented with a ball, and is expected to guess its properties, as well as applying the *push* behavior on it. It predicts the category and the properties of the object correctly, even though it is not allowed to touch the object beforehand, and is therefore unaware of its proprioceptive, haptic (*soft*), and auditory (*silent*) properties beforehand. PL: *Push Left*, PR: *Push Right*, PB: *Push Backward*, PF: *Push Forward*. [Adapted from [6] ©2015 IEEE. Best viewed in color.]

responding categories for the object. These predictions are used in turn to activate related concepts in the web. Meanwhile, over a second path, the issued command word (*e.g.*, *grasp*, *push left*, etc.) activates the necessary verb concept through the language space. When the concept web is allowed to propagate activation, knowledge (belief) oscillates between the verb concept and the initially predicted noun and adjective concepts until convergence.

In Figure 4.10, an example scenario is shown in which iCub is given a ball, and told to apply a “push” behavior on it. Although initially the haptic and auditory information are not available to iCub, these concepts are also active in the converged concept web. The quantitative predictions with and without the concept web are depicted in Table 4.3.

Table 4.3: The predictions as corrected by the activation on the concept web, when there is no direct perceptual access to certain features of the object. The iCub is not allowed to grasp the ball object, and therefore makes initial predictions using only the available visual features (columns 2 and 3). The visual parts of the concept prototypes (*i.e.*, features [1-67]) are used for this comparison. This initial activations are then allowed to spread on the concept web, which converges to the significantly more accurate a posteriori predictions displayed in columns 4 and 5. The haptic and audio predictions are corrected through the spreading of activation. Columns 8 and 9: Concept web estimations. Parentheses: Prediction confidences. Bold text: Correct decisions. Stroked text: Wrong decisions. [Adapted from [6] ©2015 IEEE.]

Object	Without Concept Web		With Concept Web	
	Nouns	Adjectives	Nouns	Adjectives
	<b>ball (37%)</b>	edgy (37%) <b>round (63%)</b>	<b>ball (100%)</b>	edgy (0%) <b>round (100%)</b>
	box (14%)	hard (45%) <del>soft (55%)</del>	box (0%)	<b>hard (100%)</b> soft (0%)
	cup (12%)	<del>noisy (54%)</del> silent(46%)	cup (0%)	noisy (0%) <b>silent(100%)</b>
	cylinder (14%)	<b>short (59%)</b> tall (41%)	cylinder (0%)	<b>short (100%)</b> tall (0%)
	plate (11%)	<b>thick (54%)</b> thin (46%)	plate (0%)	<b>thick (100%)</b> thin (0%)
	tool (12%)		tool (0%)	

### 4.3.3 Scenario 3: Action-driven activation of concepts in the web

The final scenario demonstrates how iCub is commanded to perform a certain action in an environment populated with multiple objects (Figure 4.11). The command does not specify on which object to apply the behavior, therefore iCub must itself choose the appropriate object on which to act. Here we must remember that certain behaviors cannot be applied to all objects. Therefore, activation must not spread from these verb concepts to inappropriate noun types. After convergence, iCub will pick up a properly activated noun to act upon. If there are more than one appropriate objects, a random decision will be made among them.

The entry point of activation in this scenario is from the commanded verb concept. In the sample case in Figure 4.11, iCub is presented with three objects: a *cup*, a *plate*, and a *ball*. It is then commanded to apply *drop* behavior. Since *drop* verb is not connected to *cup* and *plate* nouns, activation cannot spread to *cup* and *plate*. On the other hand, *ball* noun is activated through its connection to *drop*. As a result, iCub

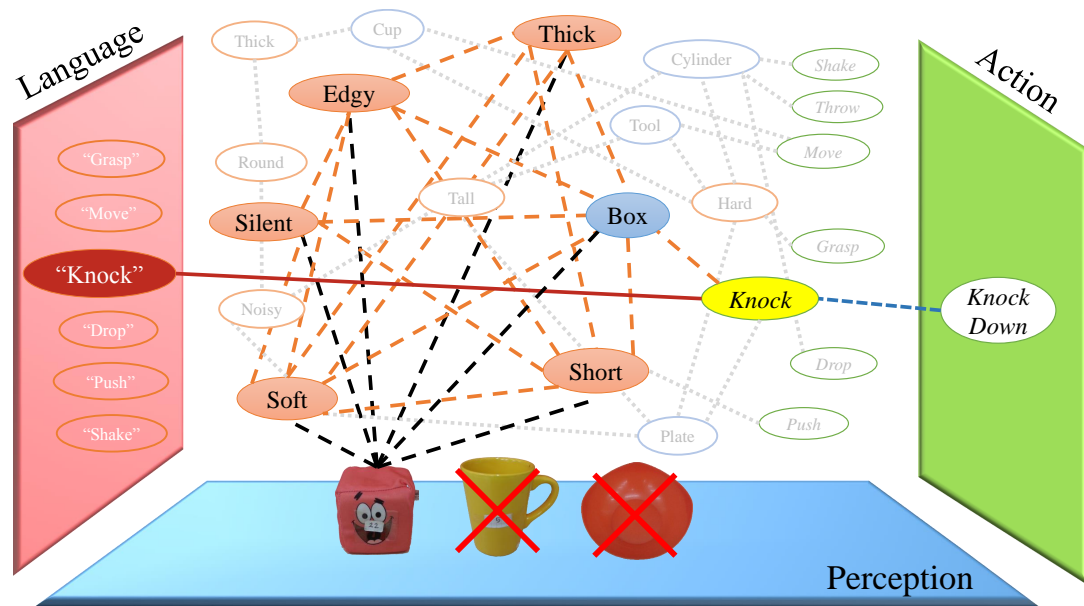


Figure 4.11: Schematic representation of Scenario 3. The sample *box*, *cup*, and *plate* objects are given to the system and *knock down* behavior is commanded to iCub. iCub selects any one of these objects if the commanded behavior is applicable. In this scenario, the *box* object is selected and its activated concepts are shown. The action space and verb concepts are contoured with green, whereas blue and orange colors represent the noun and adjective categories for the object, respectively. The gray and smaller fonts show inactive concepts in the web, while bigger fonts and colored nodes represent activated concepts. There are other concepts that are not shown for clarity. [Adapted from [6] ©2015 IEEE. Best viewed in color.]


decides to apply the action to this object. Table 4.4 presents quantitative selection percentages for two sample cases.

This scenario serves as a proof of concept that behaviors can activate related noun concepts, but not unrelated ones. This kind of “reverse” activation spreading can guide the robot’s actions in the world.

#### 4.4 Summary

In this part, we have discussed how a connected representation of concepts can contribute cognition, and why such a representation would be more biologically plausi-

Table 4.4: The selection of objects to which sample commands are applicable. The selection is performed by the spreading activation on the web, which disperses to the related verb concepts as well. *throw* verb concept activates selectively non-cup and non-plate objects. Selection confidences are indicated in parentheses. Images depict RGB-D images from the Kinect sensor. [Adapted from [6] ©2015 IEEE. Best viewed in color.]

Command	Scene	Existing Objects	Selected Objects
“throw”		<del>cup (25%)</del> <del>box (25%)</del> <del>yellow plate (25%)</del> <del>red plate (25%)</del>	<b>box (100%)</b>
“push”		box (16.67%) green cup (16.67%) white cup (16.67%) yellow plate (16.67%) red plate (16.67%) ball (16.67%)	<b>box (16.67%)</b> <b>green cup (16.67%)</b> <b>white cup (16.67%)</b> <b>yellow plate (16.67%)</b> <b>red plate (16.67%)</b> <b>ball (16.67%)</b>

ble. Analyzing the existing studies in computational modeling has showed the current lack of such a model. Therefore, we have proposed a Markov Random Field based, connected concept web model, that is grounded on the co-occurrences of concepts from the interactions of the robot. We have showed that in spite of the highly cyclic nature of the resulting graph, we can effectively conduct inference on it using Loopy Belief Propagation, as widely performed in the literature.

We demonstrated that, given an observation of an object, our robot can activate in its “brain” the relevant noun concepts, adjective concepts, verb concepts (describing what behaviors can be applied on the object) as well as the words that can be used for the object. Moreover, given an interaction on an object or in fact, an interaction without an object (that would normally take an object), the robot can activate the necessary concepts in the web as well. Being linked to language, perception and motor (action) spaces, the concept web allows activation of relevant information from

and to any modality. Moreover, we showed that such a web allows the robot to make a better interpretation of the environment. By using the co-occurrences from other concepts, the results demonstrate that wrongly predicted concepts can be corrected, and confidences of correct predictions can be increased.



## CHAPTER 5

### SPATIAL CONCEPTS AND THE REPRESENTATION OF WHOLE SCENES

The concept web we described in the previous chapter provides a biologically plausible and computationally robust scheme for conceptualization. However, it is also lacking in important parts. The first is the lack of spatial relations, which results in a deficiency of representing scenes with multiple objects. Such lack of spatial knowledge is potentially crucial, for instance, in planning. In addition, the concept web is composed of fully associative links, which enforces all active concepts to be semantically linked to each other. This is problematic for instance when there are more than one objects in the world, since a single concept web is unable to represent a ball (which is round) cannot be active together with a box (which is edgy) at the same time.

We rectify these limitations with two improvements: (1) The capability to represent spatial relations, and (2) instantiating the concept web separately for individual objects. Combining these two features, the concept web can effectively represent whole scenes seamlessly. Our working hypothesis is that prepositional spatial relations should also be regarded as *bona fide* concepts themselves, in relation to other concepts, just like a regular noun or adjective concept. Given their binary *and* directed nature (since, e.g., a ball can stand on a box, but not vice versa, since the ball is round, thus the spatial relations have *order*), we propose a Hybrid Markov Random Field variant, with *directed* connections between spatially-related nodes. These contributions has been published in the proceedings of International Conference on Advanced Robotics (ICAR) 2015 [67].

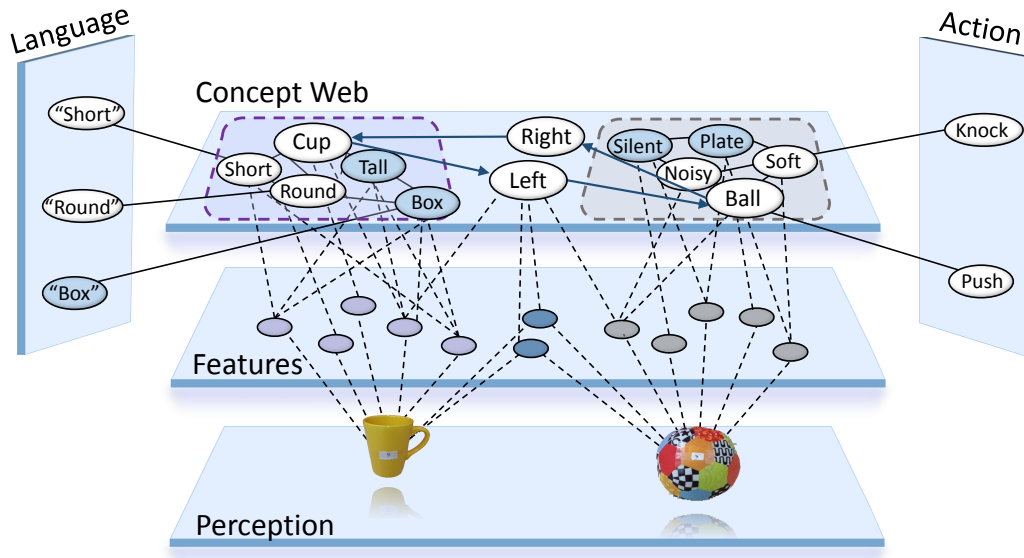


Figure 5.1: The concept web combining concept web instantiations of perceived objects, and their spatial relations. Shaded areas correspond to concept webs of the individual objects. These are fed by the extracted features of the objects, as well as by the language and action planes. The spatial relations between the objects combine these individual object representations, and are fed by the relative and the individual features of the two objects, and again by the language and action planes. [Adapted from [67] ©2015 IEEE.]

## 5.1 Data Collection and Feature Extraction

For developing the spatial concepts ( $\mathcal{S} = \{on, below, left, right, in\ front\ of, behind\}$ ), binary features also are collected from couples of objects in the scene during training and testing. Following Landau and Jackendorf [44] and Golland *et al.* [64], we employ binary *projective* features between two objects, which define the relative  $x, y, z$  positions of the two objects with respect to each other (Table 5.1). The projective features are adequate for the projective spatial concepts we deal with in this study. It should be noted that for developing other spatial concepts, such as  $\{in, out, near, far\}$ , that we do not include in this study, other specialized topological and proximity features, such as the containment and relative distance features would be beneficial, as mentioned in [64].



Table 5.1: The features used for extracting spatial concepts, including (1) binary projective features extracted from two objects in relation to each other, and (2) the individual visual features of the two objects. [Adapted from [67] ©2015 IEEE.]

Feature Type	Feature	Position
Projective ( $\mathbf{e}_{proj}$ )	Relative $x$ position	1
	Relative $y$ position	2
	Relative $z$ position	3
Visual - Object 1 ( $\mathbf{e}_v^1$ )	Object dimensions:( $width, height, depth$ )	4-6
	Normal zenith histogram bins	7-26
	Normal azimuth histogram bins	27-46
	Shape index histogram bins	47-66
Visual - Object 2 ( $\mathbf{e}_v^2$ )	Object dimensions:( $width, height, depth$ )	67-69
	Normal zenith histogram bins	70-89
	Normal azimuth histogram bins	90-109
	Shape index histogram bins	110-129

## 5.2 Representing Spatial Concepts with Prototypes

Spatial concepts are binary, and are calculated from couples of objects. Therefore, first of all, spatial prototype includes the binary projective features  $\mathbf{e}_{proj}$ , extracted from the relative positioning of the two objects. Moreover, they also contain implicit relation between the shape of the objects and the physically possible spatial positions: It is difficult to balance something on top a round object, therefore the *on* and *below* relations have semantic ties to the object shapes. Therefore, visual features of the two objects  $\mathbf{e}_v^1$  and  $\mathbf{e}_v^2$  are also included in the prototypes. The resulting prototypes are of length 129. The features used for the prototypes are given in Table 5.1, and the complete spatial prototypes extracted are depicted in Table 5.2.

## 5.3 Hybrid Markov Random Field

The standard Markov Random field is an undirected graph, which is suitable for our representation of noun, adjective, and verb concepts. However, the spatial concepts require a different scheme. That is because these spatial concepts are *directed* in nature: When *object-1* is on the left of *object-2*, *object-2* is *not* on the left of *object-1*, but on the right of it. Therefore, we propose a variant of Markov Random Field representation, which we call the *Hybrid Markov Random Field*, to model such relations.

Table 5.2: Prototypes extracted for spatial concepts. [Adapted from [67] ©2015 IEEE.]

Spatial Concepts	Projective Features	Visual Features (First Object)	Visual Features (Second Object)
Left	0-0	*****000000000000*+0000*****0000000000000+*****	*****000000000000*0+0000*****00000000000000*****
Right	0+0	*****+*****000000000000*+00000*****0000000000000+*****	*****000000000000*00+00*****00000000000000*****
On	00+	*****000000000000*+00000*****00000000000000*****	*****000000000000*0000000*****00000000000000+*****
Below	00-	*****000000000000*+00000*****00000000000000+*****	*****000000000000*+0+000*****00000000000000*****
Front	+00	*****000000000000*+00000*****00000000000000*****	*****+*****000000000000*0+0000*****00000000000000*****
Behind	-00	*****+*****000000000000*+0000*****00000000000000*****	*****+*****000000000000*+0000*****00000000000000*****

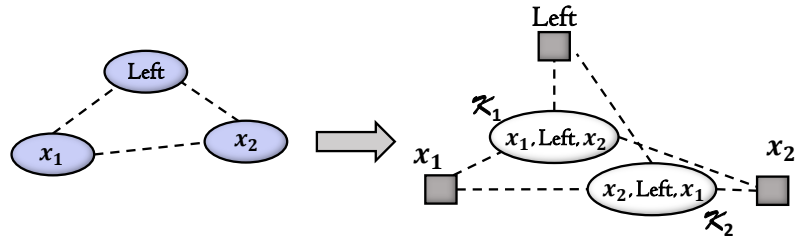


Figure 5.2: Extraction of directed cliques in a hybrid Markov Random Field, and which is converted into a factor graph with two clique nodes. [Adapted from [67] ©2015 IEEE.]

Figure 5.2 depicts a hybrid Markov Random Field schematically, as compared to the standard undirected Markov Random Field (Figure 4.7). The difference is in encoding a directed connection via two separate clique nodes in the factor graph. The first clique node denotes information flow in one direction (from concept  $x_1$  to  $x_2$  in the figure), and the second clique node denotes information flowing in the opposite direction (from concept  $x_1$  to  $x_2$  here). The potentials of the two clique nodes are calculated separately, resulting in two “Left” concepts here, each one representing Left of one of the related two objects.

## 5.4 Scene Representation

The representation of an encountered scene in the system is handled in two levels (Figure 5.1). On the one side, the attention of the system focuses on each object, and extracts a concept web of the related concepts with the object - this is akin to creating instantiations of previously learned concepts for representing the specific object. Object-specific concept web instantiations are modeled using standard (undirected) Markov Random Field representation, since as shown in [6], undirected connections are not only intuitive but also effective in capturing co-dependence between noun, adjective, and verb concepts semantically related together. The rationale for considering objects one by one is capturing human-like reasoning: Humans are able to understand and reason on single objects situated in an environment, that is, the *identity* of the object is easily extracted and generates its own related concept activations.

Simultaneously still, the whole scene is considered together, and spatial relations between couples of objects are analyzed, by additional MRF links created between the concept webs of each object. The hybrid Markov Random Field representation is used for modeling the spatial relations due to their directed nature. Moreover, the spatial relations are modeled between the *noun* components of the object representations, since it is natural for humans to address unnamed objects by their nouns, instead of their adjectives, since nouns are more discriminative in communication (*e.g.*, “Pass me the cup next to the kettle”, instead of “Pass me the small noisy object next to the tall object.”).

## 5.5 Experiments and Results

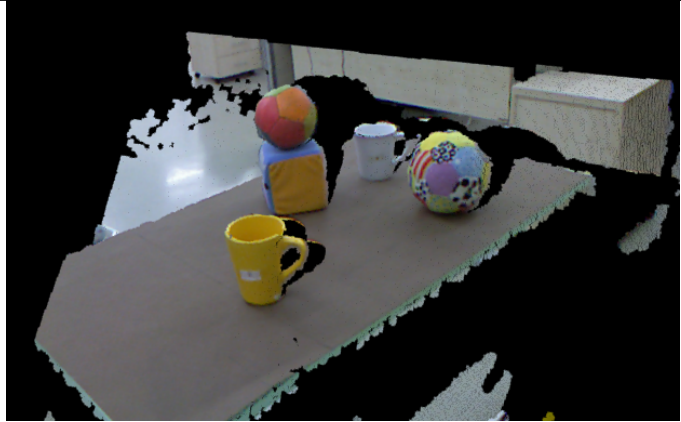
We demonstrate the extended concept web model and the effectiveness of representing spatial relations in a concept web via three different scenarios: (1) Semantic interpretation of an encountered scene as activations in the concept web, (2) correction of wrong predictions through the co-occurrence information coded in the concept web, and (3) using spatial relations to guide object search for human-robot interaction. The training set is composed of 600 arbitrary binary formations of the training objects, which are designed with a priori information: For instance, since it is very difficult to balance an object over a ball, “X-ON-BALL” combination does not exist in the training set, and therefore not represented in the concept web. On the other hand, any object can be found to the left of another object, which resulted in a large number of “X-LEFT-Y” formations.

### 5.5.1 Scene Interpretation

The interpretation of a sample scene through the proposed system is depicted in Table 5.3. In the 3D scene view acquired by the Kinect, there are two cups, two balls, and one box. iCub attends to all the objects in the scene one by one and extracts their concept webs. Then it examines their spatial relations in the hybrid MRF, resulting in the presented concepts.

Systematically, we have run the system on 5 different world views with 3 to 6 objects

Table 5.3: A sample scenario of scene interpretation. Some of the extracted relations for the presented 3D view are indicated. [Adapted from [67] ©2015 IEEE. Best viewed in color.]

Scene	Extracted Relations
	<p>A ball <b>on</b> a box  A box <b>below</b> a ball  A ball <b>on the right of</b> a cup  A cup <b>on the left of</b> a ball  A ball <b>in front of</b> a cup  A cup <b>behind</b> a ball  A box <b>on the left of</b> a cup  A cup <b>on the right of</b> a box  A cup <b>on the right of</b> a cup  A cup <b>on the left of</b> a cup  ...</p>

in the scene at one moment, resulting in 37 binary relations between them. In this setup, the system has achieved a noun concept detection rate of 95.2% and a spatial relation detection rate of 91.8% (Table 5.5).

### 5.5.2 Correcting Wrong Interpretations

The main strength of a probabilistic concept web is keeping a priori information about the world, whether due to physical laws, or canonical object utilizations of everyday life. Our expectations guide our reasoning hugely in every day life, even when our sensors may fail. In the second scenario, we show how the concept web may fulfill a similar function for iCub.

In the scene in Table 5.4, the situation of ball A with respect to ball B is slightly ambiguous: The initial predictions using only the prototypes deduce ball A can be on ball B, on the right of it, and also in front of it. Since the items are too close to each other, the contribution of the (noisy) relative-x/y/z distance predictions becomes minimal in the calculation of the concept-wise Euclidean distances (Equation 4.3), resulting the initial predictions become too close to each other and end up as non-distinctive. In fact, ball A is on another box that is on the right of ball B. Indeed it is not possible to stack balls on top of each other, since they tend to roll easily. Therefore, there is no “BALL-ON-BALL” example in the training set, and such a

Table 5.4: A sample scenario of correcting wrong interpretations of spatial relations. Prediction confidences with and without concept web are indicated. The spatial estimation of ball A is corrected through the concept web. Bold text: Correct decisions. Stroked text: Wrong decisions. [Adapted from [67] ©2015 IEEE. Best viewed in color.]

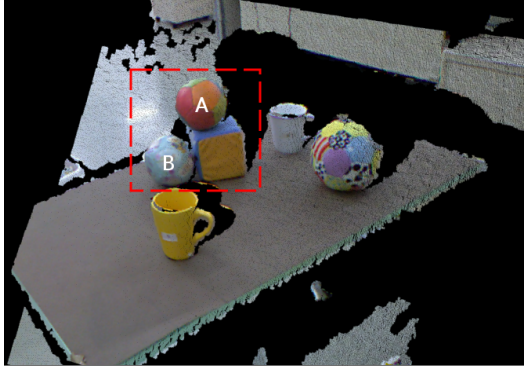
Scene	Spatial Relations	Without Concept Web	With Concept Web
	Ball A <b>on</b> ball B Ball A <b>below</b> ball B Ball A <b>left of</b> ball B Ball A <b>right of</b> ball B Ball A <b>in front of</b> ball B Ball A <b>behind</b> ball B	<del>18%</del> 15% 15% 18% <del>18%</del> 16%	0% 0% 0% <b>100%</b> 0% 0%

Table 5.5: Accuracies of concept web estimations of the noun concepts and spatial relations in the scene, and the accuracies of spatial-direction based human-robot communication. [Adapted from [67] ©2015 IEEE. Best viewed in color.]

Noun Concept Detection Accuracy	Spatial Relation Detection Accuracy	Communication Accuracy
95.2%	91.8%	96%

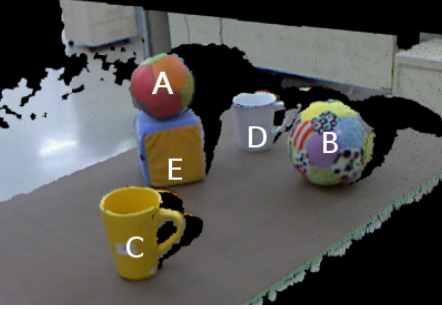
clique has not formed in the hybrid MRF. On the contrary, a large number of “BALL-RIGHT-BALL” instances has formed this clique structure, biasing the concept web towards dismissing the wrong “BALL-ON-BALL” prediction, in favor of the “BALL-RIGHT-BALL” prediction.

### 5.5.3 Human-Robot Interaction

In the final scenario, we try to communicate with iCub using spatial descriptions. iCub instantaneously evaluates any encountered scene, extracting concept webs of the objects and the spatial relations. Then, the human partner tries to guide iCub to certain object(s) using one of the following patterns:

**Fixed Noun - Fixed Relation - Fixed Noun:** Examples are BALL-ON-BOX, CUP-BEHIND-CUP, etc. A sample query is “The cup that is behind the box.”

Table 5.6: A sample scenario of human-robot interaction based on spatial-directions. Objects found by iCub in response to sample queries on the given 3D scene are indicated. [Adapted from [67] ©2015 IEEE. Best viewed in color.]

Scene	Queries	Found Objects
	<p>Object(s) <b>on the right</b> of the box?</p> <p>Object(s) <b>behind</b> the box?</p> <p>Object(s) <b>on</b> the box?</p> <p>Box is <b>on the right</b> of what?</p> <p>Box is <b>in front of</b> what?</p> <p>Box is <b>below</b> of what?</p> <p>Cup that is <b>behind</b> the box?</p> <p>Cup that is <b>on</b> another object?</p> <p>Cup that is <b>on the right</b> of another object?</p> <p>Ball that is <b>behind</b> the box?</p> <p>Ball that is <b>on</b> the box?</p>	<p>Cup D</p> <p>No such object</p> <p>Ball A</p> <p>Cup C</p> <p>None</p> <p>Ball A</p> <p>No such object</p> <p>No such object</p> <p>Cup D (to the right of ball A, box E, and cup C)</p> <p>No such object</p> <p>Ball A</p>

**Variable - Fixed Relation - Fixed Noun:** Examples are X-LEFT-BOX, X-RIGHT-CUP, etc. A sample query is “Object that is on the left of the box.”

**Fixed Noun - Fixed Relation - Variable:** Examples are BALL-BE-HIND-X, PLATE-LEFT-X, etc. A query is “The box is on the right of which object(s)?”

Through the language space, commanded concepts are allowed to stay active in the hybrid MRF level, while the separator node activations of the not-mentioned concepts are reset. Since the hybrid MRF is directional, the separator node activations are reset according to whether the fixed noun(s) in the command are in the first or second noun position. The whole system is then allowed to reiterate until convergence, at the end of which only the concepts that are relevant to *both* the visual scene and the command remain active. A sample case is presented in Table 5.6. Tests of 100 queries performed over 5 real-world scenes demonstrated a performance of 94% for this scenario (Table 5.5).

## 5.6 Summary

In this part, we proposed a method for integrating prepositional spatial concepts into the concept web model. We have also showed how whole scenes can be represented using these spatial relations. For capturing the inherently directed nature of the spatial concepts, we extended the standard Markov Random Field model to a hybrid MRF variant, which can have both undirected and directed relations between concepts. In several scenarios, we demonstrated that iCub can extract a concept-based representation of a scene and use the concepts for various reasoning tasks.



## CHAPTER 6

### A FORMALISM AND COMPUTATIONAL MODEL FOR CONTEXT

In our framework, context is defined as emerging from set of concepts that the robot perceives from its immediate environment. We use Latent Dirichlet Allocation (LDA) to detect the latent (unobserved) context of each encountered scene, as well as the individual contexts of the concepts that comprise the scene. The concepts that exist in an encountered scene are extracted through the robot's interaction with the objects. The robot makes an initial guess about the identity and the nature of each object, thereby predicting the noun and adjectives that describe this entity. Any expectation or command of an action is also added to this initial guess as a verb concept. These individual guesses, however, are to some degree related to each other, for instance certain objects commonly have certain properties (cups being round and short), or certain actions being not applicable on certain objects (dropping is not good for cups); therefore the individually predicted concepts are represented in a web structure to exploit this relatedness. This web representation of the scene is finally used for determining the context(s) of (1) the scene in general and (2) the objects individually in particular. The detected context(s) are in turn used to guide the reasoning on the web of concepts, thereby providing a feedback loop from a high-level cognitive function back to a more primitive level of perception. In this chapter, we provide the details of each of these steps.

This part of the thesis is partially published in [207] and accepted for publication in [195].

## 6.1 A Formalism of Context

Our postulate is that context is tightly related to concepts. When we see, *e.g.*, an environment with a sink, a dishwasher, and a table with cups and plates, we interpret the setting as a “kitchen”. In this case, what triggers the interpretation of “kitchen”ness, the kitchen context, is the concepts of sink, dishwasher, etc [143]. A context can be triggered by object-related concepts (nouns, adjectives), as in this example, but also by verb concepts (*e.g.*, pouring), spatial concepts (*e.g.*, cups being on the table), temporal concepts (*e.g.*, morning, noon) or social concepts (*e.g.*, family, date). Let us denote all these types of concepts by  $\mathbb{T}$  and define  $\mathbb{T}$  as follows:

$$\mathbb{T} = \{t^{noun}, t^{adj}, t^{verb}, t^{adverb}, t^{spatial}, t^{temporal}, t^{social}\}. \quad (6.1)$$

Then, let us use  $\mathbb{C}^t$  to denote the set of concepts of type  $t$ , with  $t \in \mathbb{T}$ . From this definition, it follows, for example, that the set of noun concepts,  $\mathbb{N}$ , is the same set as  $\mathbb{C}^{t^{noun}}$ . Moreover, let the set of all concepts be denoted with  $\bigcup_{t \in \mathbb{T}} \mathbb{C}^t$ , and its power set with  $\mathbb{P} = \mathbb{P}(\bigcup_{t \in \mathbb{T}} \mathbb{C}^t)$ .

The link between contexts and concepts might be of different types. For example, there are certain concepts related to a context specifically, such that their existence in a scene automatically invokes the related context. A dishwasher is a typical example, whose activation alone is enough to activate the kitchen context. In other cases, a *set* of concepts may need to be active together in order to invoke the context, such as water and boiling, which separately do not necessarily invoke the kitchen idea, but together do. This leads to the following definition:

**Definition 1.** *There exist concept sets  $\mathbb{S} \in \mathbb{P}$  that sufficiently imply a context  $\chi_k$ , that is,*

$$\exists \mathbb{S} (\text{Active}(\mathbb{S}) \implies \text{Active}(\chi_k)). \quad (6.2)$$

*These sets are minimal in that any proper subset of them does not necessarily trigger context  $\chi_k$ :*

$$\forall \mathbb{S}_s \subset \mathbb{S} (\mathbb{S}_s \not\Rightarrow \text{Active}(\chi_k)). \quad (6.3)$$

*We call any such set an enforcing concept set of context  $\chi_k$ , since it enforces the activation of context  $\chi_k$ , and denote it with  $\mathbb{S}^{+k}$ . Since there are more than one such*

sets, let us use  $\mathbb{P}^{+k}$  to denote the set of all such sets:

$$\mathbb{P}^{+k} = \{\mathbb{S} \mid (\text{Active}(\mathbb{S}) \implies \text{Active}(\chi_k))\}. \quad (6.4)$$

Not all concepts trigger a context. There exist concept sets that are in conflict with a specific context. A pool, for instance, is in conflict with the kitchen context. Such conflicting concept sets enforce the activation of an alternative context as defined below:

**Definition 2.** *There may exist concept sets  $\mathbb{S} \in \mathbb{P}$  which are in conflict with context  $\chi_k$ , and therefore enforce the activation of an alternative context  $\chi_{\bar{k}}$ :*

$$\text{Active}(\mathbb{S}) \implies \text{Active}(\chi_{\bar{k}}), \quad \chi_{\bar{k}} \neq \chi_k. \quad (6.5)$$

We call these conflicting concept sets of  $\chi_k$ , and denote them with  $\mathbb{S}^{-k}$ . Since there are more than one such sets, let us use  $\mathbb{P}^{-k}$  to denote the set of all such sets:

$$\mathbb{P}^{-k} = \{\mathbb{S} \mid \text{Active}(\mathbb{S}) \implies \text{Active}(\chi_{\bar{k}}), \chi_{\bar{k}} \neq \chi_k\}. \quad (6.6)$$

Real scenes might contain several contexts simultaneously. We may find ourselves in a studio flat with a combined kitchen-living room. Or in an outdoor bar next to a pool. In such cases, more than one context can be activated simultaneously in our minds, with all the implications due, such as the possibility of preparing a drink in the outdoor bar, together with the danger of falling into the pool. Therefore, contexts are not mutually-exclusive. This kind of multiple contextual activation is possible if *enforcing* concept sets of a context co-occur with its *conflicting* concept sets:

**Property 1.** If enforcing concept sets  $\mathbb{S}_k^+$  of a context  $\chi_k$  co-occur with its conflicting concept sets  $\mathbb{S}_k^-$ , both  $\chi_k$  and an alternative context  $\chi_{\bar{k}}$  are activated,  $\chi_k$  due to  $\mathbb{S}_k^+$ , and  $\chi_{\bar{k}}$  due to  $\mathbb{S}_k^-$ .

**Definition 3.** *If at least two different contexts are active in a scene ( $\text{Active}(\chi_k) \wedge \text{Active}(\chi_{\bar{k}}) \wedge \bar{k} \neq k$ ), the scene is called a mixed-context scene.*

Not all concepts related to a context are enforcing in the sense given in Definition 1. For instance, a cup concept is *consistent*, i.e., meaningful, in a kitchen context, but it alone cannot trigger the kitchen context. It can as well exist in a living room context,

or in an office context. However, when surrounded with a sink and a dishwasher, a cup will also be thought as part of a kitchen context. This distinction yields the following definition:

**Definition 4.** *The remaining concept sets  $\mathbb{S} \in P \setminus (\mathbb{S}^{+k} \cup \mathbb{S}^{-k})$  do not enforce context  $\chi_k$ ,*

$$\text{Active}(\mathbb{S}) \not\Rightarrow \text{Active}(\chi_k), \quad (6.7)$$

*however, when considered together with enforcing sets  $\mathbb{S}_k^+$ , they are consistent with the activation of context  $\chi_k$ ,*

$$\text{Active}(\mathbb{S}_k^+) \wedge \text{Active}(\mathbb{S}) \implies \text{Active}(\chi_k). \quad (6.8)$$

*We call these consistent concept sets of  $\chi_k$ , and denote them with  $\mathbb{S}^{*k}$ . Since there are more than one such sets, let us use  $\mathbb{P}^{*k}$  to denote the set of all such sets:*

$$\mathbb{P}^{*k} = \mathbb{P} \setminus (\mathbb{P}^{+k} \cup \mathbb{P}^{-k}). \quad (6.9)$$

From the definitions of the different types of concept sets that might be related to a context, we can now formally define a context as follows:

**Definition 5.** *A context  $\chi_k$ , indexed by  $k$ , is a latent variable, which becomes activated if an enforcing concept set  $\mathbb{S}^{+k} \in \mathbb{P}^{+k}$  is active.*

In summary, we deduce that a context has three different relations with concepts: (1) The set of enforcing sets of concepts, which necessarily invoke the activation of the concept, (2) The set of consistent sets of concepts, which do not necessarily invoke its activation, but are also meaningful in it and do not necessarily invoke the activation of an alternative context either, and (3) The set of conflicting sets of concepts, which are not meaningful in the context, and therefore necessarily invoke the activation of an alternative context.

Any attempt for modeling context must therefore be able to incorporate these properties of context. We present such a modeling of context, using Latent Dirichlet Allocation, which can explicitly handle all these properties.

## 6.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [208] is a method for modeling topics of documents in large text corpora. Assuming a document  $d \in \mathbb{D}$  is a set of words  $\{w_1, \dots, w_N\}$  drawn from a fixed vocabulary ( $w_i \in \mathbb{W}$ , vocabulary size is  $|\mathbb{W}|$ ,  $|\cdot|$  denotes set cardinality), LDA posits a finite mixture over a fixed set of topics  $\{z_1, \dots, z_k\}$  ( $z_t \in \mathbb{Z}$ ,  $|\mathbb{Z}| = K$  is the topic count). Then, a document can be described by its probabilities of being related to each of these topics,  $P(z_t|d_i)$ . Meanwhile, a topic is modeled by its probability of producing each word in the vocabulary,  $P(w_j|z_t)$ . LDA aims to infer these document and topic probability distributions, given a corpus  $\mathbb{D}$ .

Being a generative model, LDA assumes that the corpus had previously been generated by choosing a Dirichlet prior  $\alpha$ , and a  $K \times |\mathbb{W}|$  matrix, called  $\beta$ , that contains the probabilities of each word given each topic, *i.e.*, with entries  $\beta_{jk} = P(w_j|z_k)$ . Furthermore, it assumes that every document  $d \in \mathbb{D}$  had been generated by first choosing a probability distribution of topics for this document,  $\theta \sim \text{Dir}(\alpha)$ , followed by, for each word location  $n$  in the document, choosing a topic  $z_n \sim \text{Discrete}(\theta)$ , and eventually a word  $w_n$ , given the chosen topic  $z_n$  and the  $\beta$  matrix denoting  $P(w_n|z_n, \beta)$ .

LDA effectively tries to estimate the unknown  $\alpha$  and  $\beta$  parameters from the given corpus, through which it is possible to infer any other parameter. This problem, however, is infamously intractable [208]. There are various solutions though, including a variational inference method [208], a collapsed Gibbs sampling solution [3] and collapsed variational inference approach [209]. The Gibbs sampling based solution algorithm is summarized below:

**Gibbs Sampling Algorithm for Solving LDA** Introduced by Griffiths and Steyvers [3], the Batch Gibbs Sampling Approach (Algorithm 2) is a “collapsed” method for solving the LDA problem, because it integrates out the Dirichlet parameters and instead directly samples the topic variables  $\vec{z} = [z_1, \dots, z_N]$  for every word position  $n \in \{1, \dots, N\}$ . The algorithm starts by randomly assigning  $\vec{z}$ , and then until convergence samples the topic assignment  $z_j$  for the word  $w_j$  in document  $d$ , according

---

**Algorithm 2:** Batch Gibbs sampling algorithm. [Adapted from [3, 210].]initialize  $\vec{z} = [z_1, \dots, z_N]$  randomly from the set  $\{1, 2, \dots, K\}$ **while** not converged **do**    choose a word index  $j$  from  $\{1, 2, \dots, N\}$     sample  $z_j$  according to  $P(z_j | \vec{z}_{\setminus j}, \vec{w}_N)$  (Equation 6.10)**end while**

---

to the instantaneous state:

$$P(z_j | \vec{z}_{\setminus j}, \vec{w}_N) \propto \frac{n_{z_j, \setminus j}^{w_j} + \xi}{n_{z_j, \setminus j} + |\mathbb{W}| \xi} \times \frac{n_{z_j, \setminus j}^d + \alpha}{N_{\setminus j}^d + K \alpha}, \quad (6.10)$$

where  $(\cdot)_{\setminus j}$  notation stands for all items excluding the currently considered index  $j$ , therefore letting  $\vec{z}_{\setminus j}$ : the vector of all topics except  $z_j$ ,  $\vec{w}_N$ : the vector of all words,  $n_{z_j, \setminus j}^{w_j}$ : the number of times that word  $w_j$  has been assigned to topic  $z_j$ , except at index  $j$ ,  $n_{z_j, \setminus j}$ : the number of times that any word has been assigned to topic  $z_j$ , except at index  $j$ ,  $n_{z_j, \setminus j}^d$ : the number of times that any word in document  $d$  has been assigned to topic  $z_j$ ,  $N_{\setminus j}^d$ : the total number of all words in document  $d$  except at index  $j$ , with  $|\mathbb{W}|$  denoting the size of the vocabulary set, and  $K$  denoting the topic count. The approach assumes symmetric Dirichlet priors  $\alpha$  and  $\xi$ , *i.e.*, that they are vectors with the same value in all entries. The  $\alpha$  vs.  $\xi$  trade-off controls the compromise between having few topics per document, vs. having few topics per word.

The strengths of LDA are two-fold: First, it is a generative model. There exists other powerful, non-generative models for topic analysis (for instance, see [211]), however, being a generative model, LDA can assign probabilities to documents that have not been seen before. Second, it allows non-strict memberships of words to topics: A word may be generated by multiple topics, and according to which document it occurs in, considering the topic probability distribution of the document, a different topic might be assigned to the different occurrences of the word. As a robust, unsupervised Bayesian method, it has been utilized recently in a variety of applications ranging from detecting “hot topics” in science [3], to fraud detection [212], activity profiling [213], and identifying functional regulatory networks of miRNAs and mRNAs [214]. Since the method provides the statistical tools for discovering hidden topics in unsupervised data, we propose that it can also be used for modeling context. In fact, ours is not the first attempt to use LDA formulation in robotics:

Table 6.1: The correspondence between the LDA terms and the notation used in this work. [Adapted from [195] ©2015 IEEE.]

LDA	Our Notation
document $d \in \mathbb{D}$	a single scene ( <i>i.e.</i> , the set of active concepts in the scene)
corpus $\mathbb{D}$	all scenes encountered during training phase
word $w_i \in \mathbb{W}$	an active concept $c_{act}$ in the concept webs (can be a noun, adjective, or verb: $c_{act} \in \mathbb{C} = \mathbb{N} \cup \mathbb{A} \cup \mathbb{V}$ )
topic	a ‘context’, either Kitchen, Playroom, or Workshop

It has been utilized successfully for object categorization from multi-modal sensory data [215–217], and for autonomous drive annotation [218]. However, our work is the first attempt to use LDA for modeling context in robotics.

### 6.3 Modeling Contextual Information with LDA

We now describe how we model our robotics scenario within the Latent Dirichlet Allocation framework. The components of our system correspond to the specific LDA terms as follows (Table 6.1):

1. Each scene the robot encounters is represented as an LDA document. In our concept web-based model, this scene/document is then a set of active concepts.
2. The sum of all the encountered scenes is analogous to the corpus  $\mathbb{D}$ .
3. Each active concept  $c_{act}$  in this scene corresponds to a word  $w_i$  in the document.
4. Finally, the “context”s that we are trying to discover correspond to the latent topics of LDA.

Eventually, we try to associate each scene we encounter with its relevant context(s).

### 6.4 An Incremental and Online Version: Incremental-LDA

Since a robot operates in a dynamic world, it needs to be able to discover newly emerging contexts with new interactions. To truly comply with developmental principles, the robot not only needs to estimate itself the ideal number of contexts, but

also to validate its own prediction continuously and revise and update it if necessary; we cannot foresee this for it (for a very good discussion on what makes a system developmental, see [25]).

One limitation of LDA is that it requires a fixed number of topics. This requirement is characteristic of the parametric approaches, where the parameters of the solution are defined a priori and do not change no matter how many training examples are encountered. Although they are very widely used and successful in general (among well-known examples are regression, Fisher’s discriminant analysis, Bayesian graphical methods), the necessity of predefining parameters can be restrictive. In latent feature models case, different methods have been proposed for dealing with an unknown number of clusters, focusing specifically on Dirichlet-process and Bayesian solutions [219–221]. Targeting specifically the LDA problem, Teh *et al.* [222] proposed a Hierarchical Dirichlet Process framework which can start with infinitely many possible topics, and settle on the likeliest number of topics itself. Wang *et al.* [223] developed an online solution for this hierarchical setting.

Since the previously proposed variations are either batch or parametrically dependent on the number of topics  $K$ , we enhance the original LDA methodology with a simple mechanism that allows both online learning, and dynamic updating of the ideal  $K$  value over time. This new variant, henceforth called Incremental-LDA, does not need the number of contexts to have been predefined, starting instead with the most general case of  $K = 1$ , and increasing the context count as necessary.

**Incremental LDA** Incremental-LDA (Algorithm 3) decides on  $K$  dynamically, starting the with most general case,  $K = 1$ , and incrementing the context count as necessary. For deciding when to increase  $K$ , we define and use  $\mathbb{C}_{low}$ , the set of words whose confidence values for contextual assignments are lower than a threshold value  $\tau$ . If there exists such words with low confidences, *i.e.*,  $\mathbb{C}_{low} \neq \emptyset$ , Incremental-LDA attempts to increase their confidences by incrementing the context count.

**K-Incremental Gibbs Sampling** Standard batch Gibbs sampler due to Griffiths and Steyvers [3] is not suitable for use with Incremental-LDA, because it needs to start from scratch each time the context count  $K$  is incremented. The previous solution



---

**Algorithm 3:** The proposed Incremental-LDA algorithm. [Adapted from [195]

©2015 IEEE.]

---

```
initialize context count  $K \leftarrow 1$ .
for all encountered scenes do
    run K-Incremental Gibbs sampler with  $K$ 
    while  $\mathbb{C}_{low} \neq \emptyset$  do
        increment context count  $K \leftarrow K + 1$ 
        run K-Incremental Gibbs sampler with  $K$ 
    end while
    output converged context assignments  $\vec{z}_N$  for the scene
end for
```

---

is forgotten completely, whereas parts of it would still be applicable. This is especially true for the parts of the previous solution that exhibited high enough confidence. Therefore we introduce K-Incremental Gibbs Sampling (Algorithm 4) as an incremental variant: When the context count is incremented to  $K$ , K-Incremental Gibbs Sampling resumes its search from the previously converged solution for  $K - 1$  contexts, conducting a local search in the close vicinity. This is done by retaining the previous assignments of the high-confidence terms, while initializing low-confidence terms ( $\mathbb{C}_{low}$ ) to the newest context id  $K$ . Effectively, the highly confident part of the solution is reused. Note that for escaping possible local minima, a high-confidence term can also be reassigned to the new context with a low probability  $\delta \ll 1$ .

---

**Algorithm 4:** The K-Incremental Gibbs sampling approach we propose as a companion to Incremental-LDA. [Adapted from [195] ©2015 IEEE.]

---

```
initialize  $\vec{z}_N$  from the previous solution for  $K - 1$  contexts
 $\forall$  context  $t \mid c_t \in \mathbb{C}_{low}$ , initialize  $z_t \leftarrow K$ 
 $\forall$  context  $t' \mid c_{t'} \notin \mathbb{C}_{low}$ , reassign  $z_{t'} \leftarrow K$  with prob.  $\delta \ll 1$ 
while not converged do
    choose a concept index  $j$  from  $\{1, 2, \dots, N\}$ 
    sample  $z_j$  according to  $P(z_j \mid \vec{z}_{N \setminus j}, \vec{w}_N)$  (Equation 6.10)
end while
```

---

## 6.5 LDA versus the Requirements of Contextual Modeling

In Section 6.1, we provide an explicit formalization of context. We propose LDA formulation is a particularly appropriate method for modeling this formalization, given its following properties:

- Due to the probabilistic nature of LDA, it allows non-strict assignment of words and documents to topics. For instance, if a certain word  $w$  occurs within the vicinity of group A of words in one kind of document, and group B of words in another, LDA can assign  $w$  to topic  $\chi_A$  in the first case, and topic  $\chi_B$  in the second case. This scenario corresponds to *consistent* concepts in our formalization, where  $w$  is consistent with both topics, with  $w \in \mathbb{P}^{*A} \wedge w \in \mathbb{P}^{*B}$ .
- If, on the other hand, a word  $w$  occurs within the vicinity of group C of words only, it is strongly associated with topic  $\chi_C$ , such that its probability of belonging to other topics  $\mathbb{Z} \setminus \chi_C$  diminishes to 0. This scenario corresponds to *enforcing* concepts in the formalization, with  $w \in \mathbb{P}^{+C}$ .
- If a word  $w$  never occurs within the vicinity of group D, its probability of belonging to D approaches to 0. In this case,  $w$  is an *conflicting* concept of topic  $\chi_D$ , with  $w \in \mathbb{P}^{-D}$ .
- If two words  $w_i \in \mathbb{P}^{+A}$  and  $w_j \in \mathbb{P}^{*A}$  occur in a document together, due to the enforcing nature of  $w_i$  and consistent nature of  $w_j$ , LDA determines this document as of topic  $\chi_A$ .  $w_i$  and  $w_j$  themselves are also associated with topic  $\chi_A$  in this document.
- In contrast, if  $w_i \in \mathbb{P}^{+A}$  and  $w_m \in \mathbb{P}^{-A}$  occur in a document together, due to the conflict of the enforcing nature of  $w_i$  with conflicting nature of  $w_m$ , LDA assigns *two* topics to the document, both  $\chi_A$  and  $\chi_{\bar{A}}$ , with  $\bar{A} \neq A$ ,  $\chi_A$  due to  $w_i$  and  $\chi_{\bar{A}}$  due to  $w_m$ . This is a common scenario in real life, where items related to different topics can also be found together occasionally, which we called above a *mixed-context* scenario. In such a case,  $w_i$  is associated with topic  $\chi_A$  and  $w_m$  is associated with topic  $\chi_{\bar{A}}$ .
- LDA works on the bag-of-words assumption that the order of the words in a

document is not important, which is compatible with the unordered set formalization of context. Indeed, concepts exist or do not exist in a scene, there is no ordering between them. In other words, the probabilities of concepts are not dependent on their order of appearance, in contrast to certain NLP scenarios. On the other hand, it does take into account the cardinality of concepts, the more instances of the same concept exists in a scene, the more strongly it affects the context.

- As detailed above, LDA can be made to operate in an online and incremental manner, which is consistent with our aim of lifelong development of robots in a changing world.

## 6.6 Making Use of Context: Feeding the Contextual Information back to the Concept Web

Since the system does not employ an attentional mechanism, it focuses on each object in the scene one by one, identifying the concepts related to each one with a concept web. The set of all these active concepts for all objects is then used for deducing the context of the scene. After determining the context, the probabilities of concepts are updated with the conditional likelihood of concepts in that context:

$$P(c)^* = \sigma \times P(c) + (1 - \sigma) \times P(c|\chi), \quad (6.11)$$

where  $c \in \mathbb{C} = \mathbb{N} \cup \mathbb{A} \cup \mathbb{V}$  is a concept,  $P(c)$  is the MRF-decided probability of the concept  $c$ ,  $\chi$  is the context,  $P(c|\chi)$  is the probability of the concept given the context (decided by Incremental-LDA), and  $P(c)^*$  is the updated value of the concept probability. The whole system, which consists of (1) reiteration of the object concept webs, (2) context deduction, and (3) probabilistic update of concept webs according to the context, and (3) reiteration of MRF loop, is then repeated until the convergence of the individual concept webs and context analysis.

$\sigma$  in Equation 6.11 is responsible with regulating the strength of contextual feedback in our world, with  $\sigma = 0$  corresponding to using only contextual information, and  $\sigma = 1$  corresponding to pure concept web decision. An average log likelihood  $\hat{l}$  is calculated over the test set as follows and depicted in Figure 6.1:

$$\hat{l} = \frac{1}{N|\mathbb{C}^{n+}|} \sum_{i=1}^N \sum_{c \in \mathbb{C}^{n+}} \log P(c|x_n, \sigma), \quad (6.12)$$

with  $N$  denoting the observation count,  $x_n$  being the  $n^{\text{th}}$  observation,  $\mathbb{C}^{n+}$  with cardinality  $|\mathbb{C}^{n+}|$  being the set of concepts *related with* the  $n^{\text{th}}$  observation, and  $P(c|x_n, \sigma)$  denoting the probability of obtaining the related concept  $c$  given observation  $x_n$ , under the setting  $\sigma$ . The results estimate a reasonable interval between  $[0.4, 0.5]$ ; from this interval, we select  $\sigma$  as 0.5. Note that the convergence of  $\hat{l}$  for  $\sigma \geq 0.7$  corresponds to the contextual feedback being too weak to affect concept web decision at all, therefore the average log-likelihood does not vary in this region.

## 6.7 Entropy-Based Evaluation of the System

We define an entropy-based metric of disorder to evaluate the performance of the system, with two terms:

$$\tilde{H} = \rho \times H(C|X) + (1 - \rho) \times H(X|S), \quad (6.13)$$

where  $H(\cdot)$  is the entropy function,  $C$ ,  $X$ ,  $S$  are random variables denoting concepts, contexts, and scenes respectively,  $H(C|X)$  is the conditional entropy of concepts given the context,  $H(X|S)$  is the conditional entropy of contexts given the scene, and  $\rho$  is a parameter determining the relative importance of the two terms (set to 0.25 experimentally). These two terms stem from two possibly opposing targets: We would like as few contexts as possible assigned to a scene, giving us more specific documents; and at the same time as few concepts as possible associated with a context. A combination of the two terms is expected to give us the most specific contextualization of the scene<sup>1</sup>.

## 6.8 Experiments and Results

We evaluate our framework and assumptions from three different aspects:

---

<sup>1</sup> Similar multi-objective optimization of these two metrics can be found in the literature, for instance see [3].

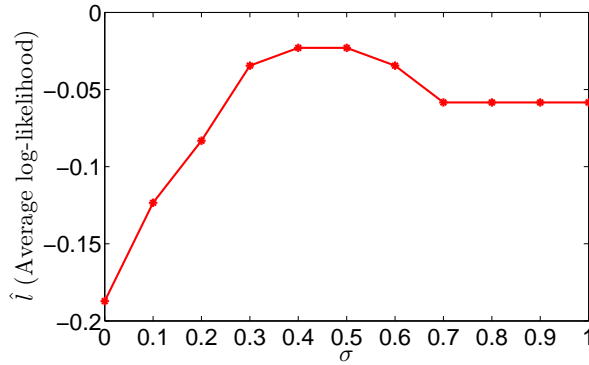


Figure 6.1: Average log likelihood  $\hat{l}$  for varying  $\sigma$  (Equation 6.11,  $\sigma = 0$ : Pure contextual information,  $\sigma = 1$ : Pure concept web decision). The interval  $[0.4, 0.5]$  is depicted as maximizing  $\hat{l}$ . [Adapted from [195] ©2015 IEEE.]

1. We first test whether Incremental-LDA can determine the optimal number of contexts; *e.g.*, if it stops adding new contexts at the optimal point. We also test if reusing partial solutions in K-Incremental Gibbs sampler leads to better performance.
2. Then we compare extracting context *directly* from raw features of the scene, against modeling it on top of the concept web.
3. Finally, we demonstrate how contextual information can improve reasoning, in three different scenarios: (1) scene interpretation, (2) object recognition, and (3) planning.

The training and test scenes in the experiments can belong to 3 different contexts (*Kitchen*, *Playroom*, and *Workshop*). Unless explicitly mentioned, a scene is a *pure* context scene, *i.e.*, contains elements of a single context. A scene can also contain elements from multiple contexts, in which case it is denoted as a *mixed* context scene. For generating each scene in the set, a context is decided randomly and then the scene is populated with randomly chosen objects that have the noun, adjective, and verb attributes related to the selected context.

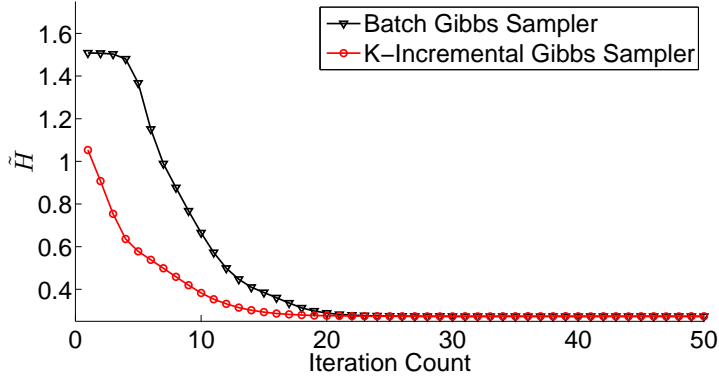


Figure 6.2: A comparison of the entropy ( $\tilde{H}$ ) evolution (Equation 6.13) of K-Incremental Gibbs solver, versus the standard batch Gibbs solver. The K-Incremental Gibbs solver is fed a partial solution for 2 contexts and then run for  $K = 3$  contexts. The batch Gibbs sampler is directly run for  $K = 3$  contexts. [Adapted from [195] ©2015 IEEE.]

### 6.8.1 Performance of Incremental-LDA and K-Incremental Gibbs Sampling

First, we analyze the dynamics of Incremental-LDA as: (1) the system encounters more and more scenes, and (2) the number of contexts  $K$  is increased. In the first case, we wish it to detect the correct context count as soon as possible, *i.e.*, with the smallest number of scenes possible. In the second case, we would like it to converge on the correct number of contexts, which is  $K = 3$  for our experimental scenario.

Figure 6.3 depicts two different evaluations of the outcome: The top two figures, Figure 6.3(a) and (b) present the number of highly uncertain concepts ( $|\mathcal{C}_{low}|$ ) on the  $y$ -axis. The bottom figures, Figure 6.3(c) and (d) demonstrate the change in the combined entropy ( $\tilde{H}$ , Equation 6.13), for the same scenarios. For both cases, lower values are more desirable.

For each configuration, we use 10 test sets of  $|\mathbb{D}|$  scenes with random contexts. The number of encountered scenes in each test set,  $|\mathbb{D}|$ , is one of the free variables. Each scene  $d \in \mathbb{D}$  is populated with 3-5 random objects of the randomly selected context. In the figures, the mean and standard deviation values for the 10 test sets are indicated with error bars.

Figure 6.3(a) shows that, for test sets of varying scene counts, the system would tend to increase the context count  $K$  until it reaches  $K = 3$ , at which point  $|\mathbb{C}_{low}|$  eventually hits 0. (Remember that while the set  $\mathbb{C}_{low}$  is not empty, a new context will be introduced in an attempt to reduce it, see Section 6.4.) For the trivial case of 1 scene, this is already the case with a single context ( $K = 1$ ). This is expected, since a single context is perfectly capable of describing a single scene. For the boundary case of 2 scenes, 2 contexts seems to be enough. However for 3 and more scenes, it seems that 3 contexts is necessary to clear out the set of uncertain concepts. Left to its own, the system would naturally stop adding new contexts at this point, since  $|\mathbb{C}_{low}| = 0$  by now. However, for the sake of comparison in these graphs, the system is enforced to try out the extra context counts of  $K \geq 4$ . It should be noted that this behavior is artificial, and as shown by Figure 6.3(a), the preferred context count by the system is  $K = 3$ , where  $|\mathbb{C}_{low}| = 0$  for the nontrivial scene counts of  $|\mathbb{D}| \geq 3$ .

Figure 6.3(b) depicts the very same results from a different point of view, with the encountered scene counts  $|\mathbb{D}|$  on the  $x$ -axis this time. Here it is also visible that context counts of 1 and 2 are not enough for maintaining  $|\mathbb{C}_{low}| = 0$  for more scenes than 2. (While  $K = 1$  is enough for a single scene, and  $K = 2$  is enough for 2 scenes.) 3 and more contexts are effective in keeping  $|\mathbb{C}_{low}| = 0$  for any number of scenes. However, as mentioned above,  $K = 3$  being the first setting that achieves  $|\mathbb{C}_{low}| = 0$ , the system opts to maintain  $K = 3$  naturally, and the extra values  $K \geq 4$  are obtained by artificially forcing the system to increase its context count, for the sake of comparison.

For proving the sanity of the behavior, we also have a look at the system dynamics by considering the entropy change in these scenarios, using the combined entropy  $\tilde{H}$  developed in Section 6.7. Figures 6.3(c) and (d) depict the change in entropy of the system for the exact scenarios discussed above. In Figure 6.3(c), we show how the entropy of the system changes as  $K$  is increased in the system, note again that the values  $K \geq 4$  are not natural but artificially forced. The minimum values of entropy are encountered at  $K = 1$  context for a single scene ( $|\mathbb{D}| = 1$ ), at  $K = 2$  contexts for 2 scenes, and at  $K = 4$  contexts for any non-trivial scene count  $|\mathbb{D}| \geq 3$ . It can be seen that, if the system is artificially forced to increase the context count  $K$  beyond that, the entropy actually *increases*, which is not desired. Therefore, the proposed

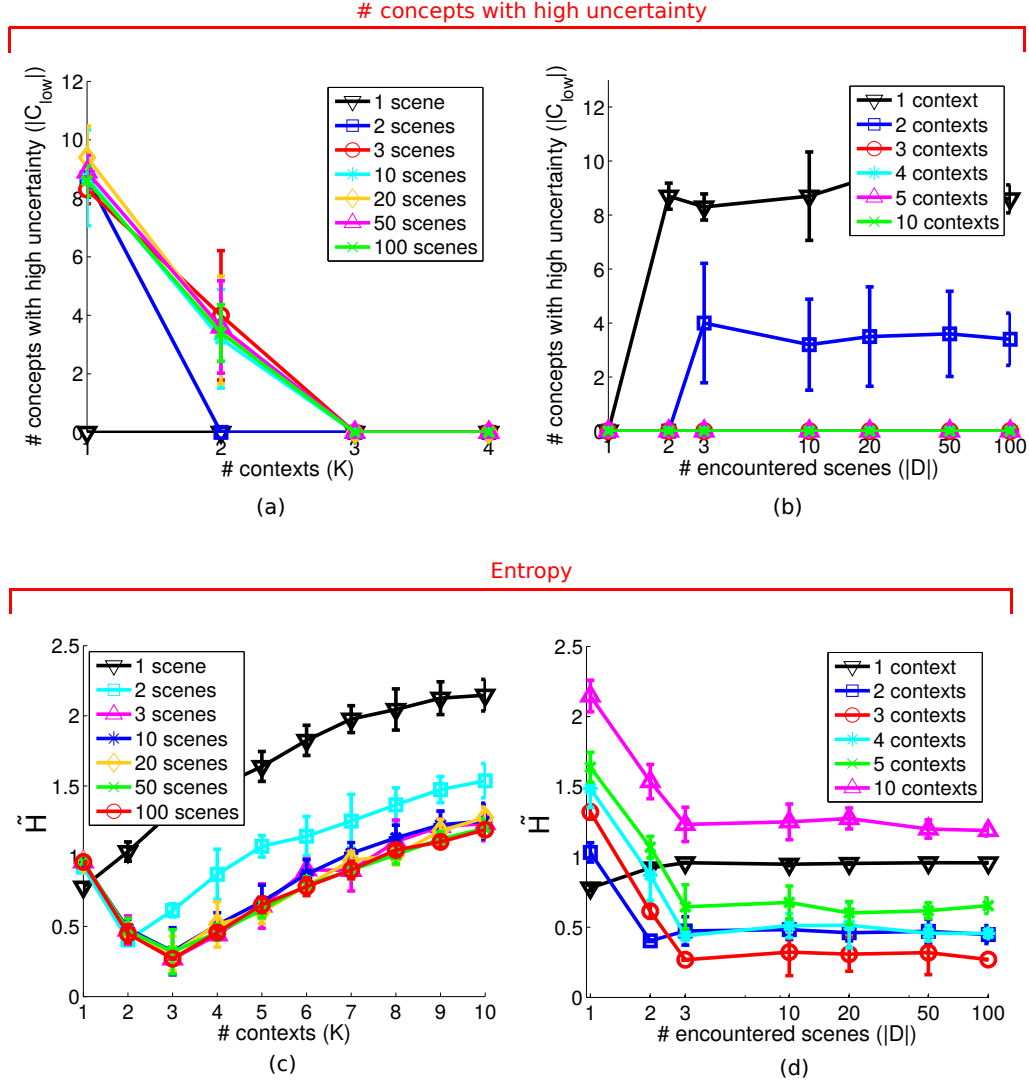


Figure 6.3: The effect of encountered scene counts and varying context counts  $K$ . Note that Incremental-LDA would itself stop at  $K = 3$ , however we force increasing  $K$  for the sake of comparison. (a) Effect of increasing  $K$  on the number of uncertain concepts,  $|\mathcal{C}_{low}|$ , for varying number of scenes. By  $K = 3$  contexts  $|\mathcal{C}_{low}|$  diminishes to 0, therefore Incremental-LDA would stop adding new contexts at this point. (b) Effect of encountered scenes on the number of uncertain concepts,  $|\mathcal{C}_{low}|$ , for different context counts. (c) Effect of increasing  $K$  on the entropy of the system,  $\tilde{H}$ , for varying number of scenes. (d) Effect of encountered scenes on the entropy of the system,  $\tilde{H}$ , for different context counts. In all the experiments, 10 test sets of  $|\mathbb{D}|$  scenes each are used. The mean values for the 10 test sets are plotted, while the standard deviations are indicated with error bars. In (b) and (d), the  $x$ -axis is in log-scale. [Adapted from [195] ©2015 IEEE. Best viewed in color.]



method, which increases the context count only until  $|C_{low}|$  diminishes to 0, stopping therefore at  $K = 3$ , is also effective at correctly catching the minimum entropy point of the system, again at  $K = 3$  (for the non-trivial case  $|\mathbb{D}| \geq 3$ ).

Finally, Figure 6.3(d) demonstrates the same results, this time by displaying the varying scene count  $|\mathbb{D}|$  on the  $x$ -axis. It is possible to see that, for non-trivial scene counts of  $|\mathbb{D}| \geq 3$ , if 3 or more contexts are used ( $K \geq 3$ ), the system is able to converge to the minimum entropy value of the setting as soon as it has encountered  $|\mathbb{D}| = 3$  scenes, which is an optimistic result that shows the system is able to converge quickly. Again comparing the possible  $K$  settings, we see that, except for the trivial settings of  $|\mathbb{D}| \in \{1, 2\}$ ,  $K = 3$  results in the lowest possible entropy values, as expected.

Next, we compare the performance of K-Incremental Gibbs sampling with batch Gibbs sampling. The question is whether reusing the previous partial solution leads to faster convergence times for K-Incremental Gibbs sampler. Our test set includes 100 scenes.

Figure 6.2 presents the results over this test set that conform with our expectations: Using a partial solution for 2 contexts, K-Incremental Gibbs sampler converges faster compared to the batch solver. We measure the convergence of the system in terms of its entropy<sup>2</sup>.

Note that K-Incremental Gibbs Sampling is fundamentally a variant of the standard Gibbs sampler employing an informed initialization, which has been successful in various challenging problems with very high number of contexts (topics), e.g., in [3], which extracts  $\approx 300$  “hot” scientific topics over 28,154 abstracts published in PNAS between 1991 and 2001. Therefore, in spite of the physical limitations on the data set used in this study, resulting in a modest number of concepts and contexts, it is reasonable to expect that K-Incremental Gibbs sampler will also be able to scale up for a high number of contexts as well.

---

<sup>2</sup> Also note that the entropy value eventually reached by the two solvers is indeed the expected minimum entropy value for these environmental conditions.

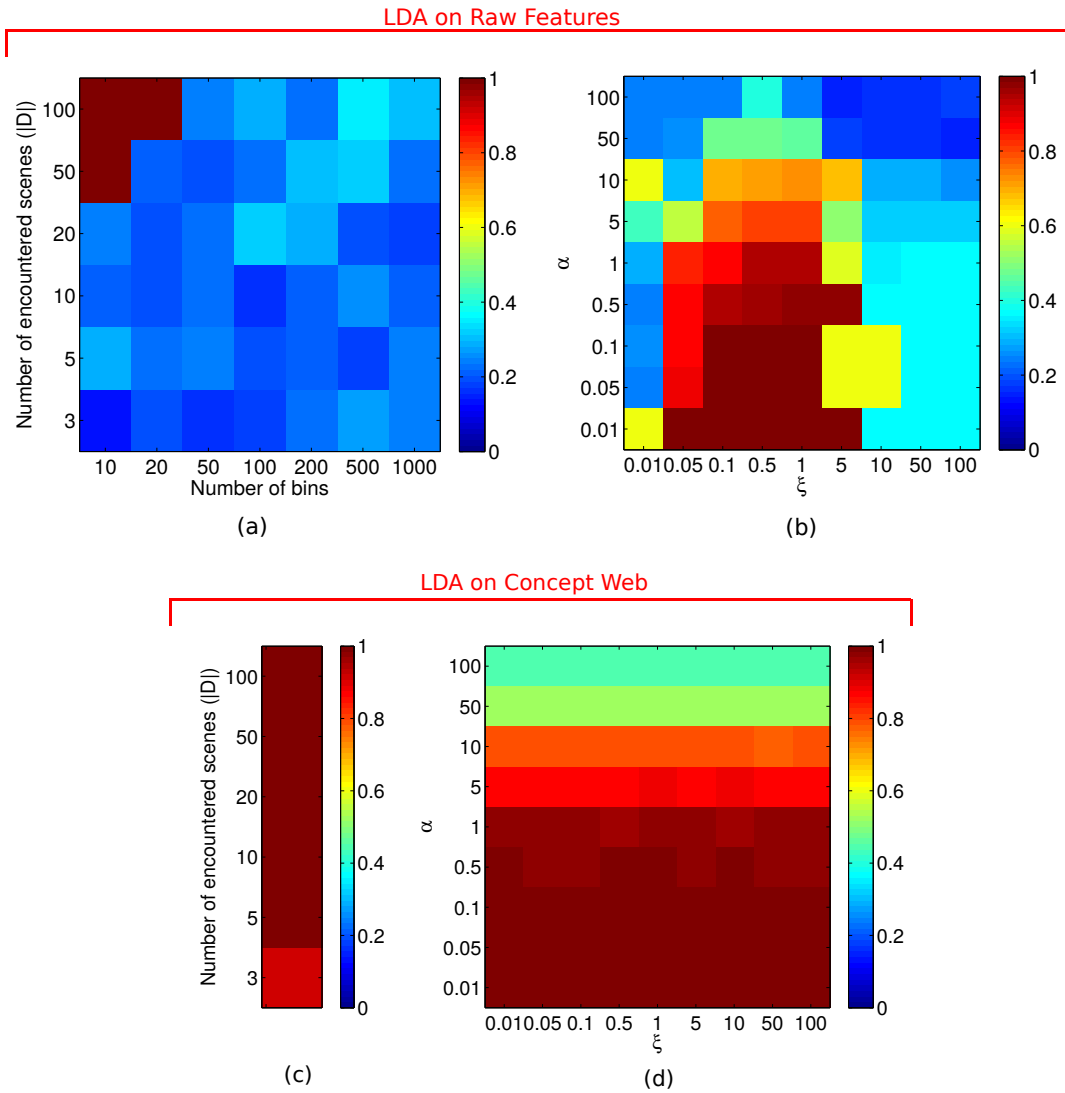


Figure 6.4: The performances of LDA over raw features only, versus of LDA over MRF-based concept web, presented as prediction accuracies scaled to [0,1]. Presented values are the predicted likelihoods of “correct contexts” in each corresponding case. For evaluation, the ground truth data of the expected contexts were extracted via supervision.  $\alpha$  and  $\xi$  are the trade-off parameters from Equation 6.10. (a) Using only the raw features as input to LDA, for varying discretization bin counts and increasing numbers of encountered scenes ( $\alpha = 0.1$ ,  $\xi = 0.1$ ). (b) Using only the raw features as input to LDA, for varying settings of  $\alpha$  and  $\xi$  (50 scenes, 10 bins). (c) Using the concept web as input to LDA, for increasing numbers of encountered scenes. ( $\alpha = 0.1$ ,  $\xi = 0.1$ . Discretization is not necessary, therefore the result vector is 1-dimensional.) (d) Using the concept web as input to LDA, for varying settings of  $\alpha$  and  $\xi$  (50 scenes). [Adapted from [195] ©2015 IEEE. Best viewed in color.]

## 6.8.2 Context from the Concept Web against Context from Raw Features

Next we evaluate how useful the concept web is in guiding contextualization. Figure 6.4 shows the comparison of LDA on concept web versus LDA on raw-features-only. First, we contrast how the two schemes fare in case of insufficient scene encounters. Concurrently, we also investigate to what degree the discretization of the raw-features is necessary, if at all. In the second type of tests, we conduct a grid parameter search in the LDA space, to decide the best parameter settings for the two algorithms, as well as their sensitivity level to the changes in these parameters. Note that these two sets of experiments must be thought of in unison, in the sense that we have iteratively updated the parameters used in one set according to the best results of the other set, therefore we hope to present meaningful results in both sets. In the figures, we present the predicted likelihoods assigned by these algorithms to the contexts that we “know” to be true. The correct contexts have been decided through supervision for evaluation purposes only.

Figure 6.4(a) versus Figure 6.4(b) depicts the results of the first set, *i.e.*, the effects of scene count and discretization (with the trade-off parameters  $\alpha$  and  $\xi$  from Equation 6.10 both set to 0.1) An important result that pops out is that the raw-features approach needs 50 scenes to settle on a meaningful partitioning, while the concept web method manages to converge with an impressive speed at as few as 3-5 scenes. Even at 50 scenes, the raw-features approach needs to be supported by coarse discretization of the features (*i.e.*, being divided into 10 bins at most), since LDA is unable to locate statistically significant co-occurrences otherwise. For other settings, the decisions of the raw-features approach are at chance level: 33.3% for a 3-way decision.

Figures 6.4(c) and 6.4(d), on the other hand, present the results of the grid search in the  $\alpha$ - $\xi$  space (with 50 scenes, 10-bins of discretization). Once again, we see that LDA-on-raw-features is more fragile against parameter changes, while the concept web method proves robust under most settings. Indeed, even for the worst parameter settings, notice that the concept-web case provides confidences of over 50%, which are sufficient for correct decision making, and are well over the chance level of 33.3%.

The results confirm that learning context from concepts is better than learning them

from raw features in two aspects: (i) Learning converges faster, and is therefore more reliable even after as few as 3-5 scene encounters, and (ii) It is less sensitive to the model parameters, which increases the robustness of learning without needing a careful tuning of parameters.

### **6.8.3 Using Context, Part 1: Making Sense of Pure- and Mixed-Context Environments**

Now we demonstrate how our context model can be utilized in reasoning and decision making. The first scenario is designed for assessing how successful our model is in recognizing contexts of scenes. The robot encounters six different scenes, three of which are composed of items of a single context, and the remaining three of multiple contexts. Table 6.2 demonstrates the predicted context(s), showing that the robot can distinguish between pure and mixed-context scenes correctly, and decide on the correct components in case of a mixed-context scene. These results are important, because they demonstrate that our interpretation of the scene context is correct, regardless of the scene being composed of a single context or multiple contexts. Therefore, we obtain justification for our next step of using this contextual interpretation for guiding reasoning in other cognitive tasks.

### **6.8.4 Using Context, Part 2: Object Recognition in Context**

The second scenario considers the effect of context on object recognition. Table 6.3 demonstrates the recognition results for seven sample objects that are either (i) individually perceived (columns 2-3), (ii) assessed in an individual concept web (columns 4-5), or (iii) evaluated in context<sup>3</sup> (columns 7-8).

The results show that concept web itself can correct certain mistakes of the perception-only assessment, while also boosting confidences of guesses to 100% certainty. However, it is not flawless and is also prone, albeit in a lesser amount, to errors (see the 2<sup>nd</sup> and 3<sup>rd</sup> rows in the table). In such cases, it is especially difficult to correct these errors, due to the initial high confidence associated with them. Contextual information

---

<sup>3</sup> The objects are given in pure-context environments, for the sake of easy analysis.

Table 6.2: Prediction confidences are indicated in parentheses. Bold text indicates correct decisions whereas stroked text indicates wrong decisions. [Adapted from [195] ©2015 IEEE. Best viewed in color.]



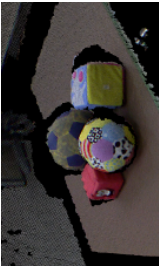



Pure Contexts			Mixed Contexts		
Scene	Existing Objects	Predicted Context (% contribution)	Scene	Existing Objects	Predicted Context (% contribution)
	2 cups, 2 plates	<b>Kitchen (100%)</b>		3 boxes, 1 ball 1 cylinder, 1 tool	<b>Playroom (72.59%)</b> <b>Workshop (26.23%)</b> Kitchen (1.18%)
	2 boxes, 2 balls	<b>Playroom (100%)</b>		2 plates, 2 cup, 1 ball, 1 box	<b>Kitchen (62.04%)</b> <b>Playroom (37.14%)</b> Workshop (0.82%)
	2 tools, 3 cylinders	<b>Workshop (100%)</b>		1 tool, 1 cylinder, 1 plate, 1 cup	<b>Kitchen (46.67%)</b> <b>Workshop (51.56%)</b> Playroom (1.77%)

Table 6.3: Object recognition in context. Prediction confidences are indicated in parentheses. Bold text indicates correct decisions whereas stroked text indicates wrong decisions. [Adapted from [195] ©2015 IEEE. Best viewed in color.]

Objects	Perception only		Concept Web		In Context		
	Predicted Nouns (% confidence)	Predicted Adjectives (% confidence)	Predicted Nouns (% confidence)	Predicted Adjectives (% confidence)	Context	Predicted Nouns (% confidence)	Predicted Adjectives (% confidence)
	ball (8%) box (13%) <b>cup (43%)</b> cylinder (20%) plate (9%) tool (7%)	edgy (34%) <b>hard (71%)</b> noisy (42%) <b>short (54%)</b> thick (47%)	ball (0%) box (0%) <b>cup (100%)</b> cylinder (0%) plate (0%) tool (0%)	edgy (0%) <b>hard (100%)</b> noisy (0%) <b>short (100%)</b> <b>thick (100%)</b>		ball (0%) box (0%) <b>cup (100%)</b> cylinder (0%) plate (0%) tool (0%)	edgy (0%) <b>hard (100%)</b> noisy (0%) <b>short (100%)</b> <b>thick (100%)</b>
	<b>ball (33%)</b> box (16%) cup (13%) cylinder (13%) plate (14%) tool (11%)	edgy (42%) hard (39%) <b>noisy (62%)</b> <b>short (61%)</b> <b>thick (56%)</b>	<b>ball (100%)</b> box (0%) cup (0%) cylinder (0%) plate (0%) tool (0%)	edgy (0%) hard (0%) <b>noisy (100%)</b> <b>short (100%)</b> <b>thick (100%)</b>		<b>ball (100%)</b> box (0%) cup (0%) cylinder (0%) plate (0%) tool (0%)	edgy (0%) hard (0%) noisy (0%) <b>short (100%)</b> <b>thick (100%)</b>
	ball (12%) box (13%) cup (17%) <b>cylinder (29%)</b> plate (12%) tool (17%)	edgy (45%) <b>hard (56%)</b> <b>noisy (58%)</b> short (41%) thick (40%)	ball (0%) box (0%) cup (0%) <b>cylinder (100%)</b> plate (0%) tool (0%)	edgy (0%) <b>hard (100%)</b> <b>noisy (100%)</b> short (0%) thick (0%)		ball (0%) box (0%) cup (0%) <b>cylinder (100%)</b> plate (0%) tool (0%)	edgy (0%) <b>hard (100%)</b> <b>noisy (100%)</b> short (0%) <b>thick (100%)</b>
	ball (14%) <b>box (43%)</b> cup (12%) cylinder (12%) plate (11%) tool (8%)	<b>edgy (64%)</b> hard (34%) noisy (30%) <b>short (59%)</b> <b>thick (63%)</b>	ball (0%) <b>box (100%)</b> cup (0%) cylinder (0%) plate (0%) tool (0%)	<b>edgy (100%)</b> hard (0%) noisy (0%) <b>short (100%)</b> <b>thick (100%)</b>		ball (0%) <b>box (100%)</b> cup (0%) cylinder (0%) plate (0%) tool (0%)	<b>edgy (100%)</b> hard (0%) noisy (0%) <b>short (100%)</b> <b>thick (100%)</b>
	ball (12%) box (13%) cup (12%) cylinder (14%) <b>plate (40%)</b> tool (9%)	edgy (46%) <b>hard (53%)</b> noisy (44%) short (45%) <b>thick (59%)</b>	ball (0%) box (0%) cup (0%) cylinder (0%) <b>plate (100%)</b> tool (0%)	edgy (0%) <b>hard (100%)</b> noisy (0%) short (0%) <b>thick (100%)</b>		ball (0%) box (0%) cup (0%) cylinder (0%) <b>plate (100%)</b> tool (0%)	edgy (0%) <b>hard (100%)</b> noisy (0%) short (0%) <b>thick (100%)</b>
	ball (11%) box (13%) cup (15%) cylinder (18%) plate (11%) <b>tool (32%)</b>	edgy (48%) <b>hard (55%)</b> <b>noisy (61%)</b> short (39%) <b>thick (57%)</b>	ball (0%) box (0%) cup (0%) cylinder (0%) plate (0%) <b>tool (100%)</b>	edgy (0%) <b>hard (100%)</b> <b>noisy (100%)</b> short (0%) <b>thick (100%)</b>		ball (0%) box (0%) cup (0%) cylinder (0%) plate (0%) <b>tool (100%)</b>	edgy (0%) <b>hard (100%)</b> <b>noisy (100%)</b> short (0%) <b>thick (100%)</b>
	ball (14%) box (17%) cup (19%) <b>cylinder (26%)</b> plate (13%) tool (11%)	edgy (42%) <b>hard (60%)</b> noisy (42%) short (45%) thick (40%)	ball (0%) box (0%) cup (0%) cylinder (0%) plate (0%) tool (0%)	edgy (0%) <b>hard (100%)</b> noisy (0%) <b>short (100%)</b> thick (0%)		ball (0%) box (0%) cup (0%) <b>cylinder (100%)</b> plate (0%) tool (0%)	edgy (0%) <b>hard (100%)</b> noisy (0%) <b>short (100%)</b> thick (0%)

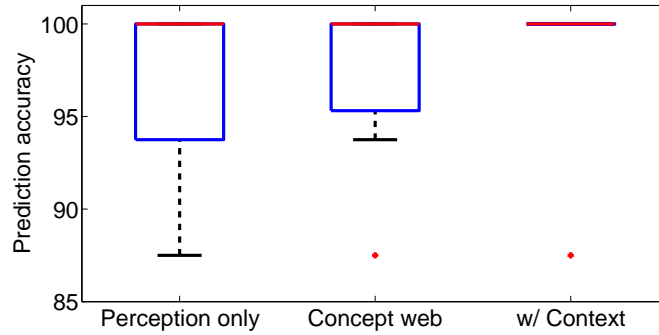


Figure 6.5: The combined results of object recognition in context, over all 15 objects in the test set. The prediction accuracies over all determined noun and adjective concepts, using (i) only perceptual features, (ii) the concept web, and (iii) contextual information are compared. In the plot, the red lines denote the median values, the boxes denote the data that fall between the 25<sup>th</sup> and 75<sup>th</sup> percentiles, the whiskers cover the extreme data that are not outliers, and stars indicate the outliers. [Adapted from [195] ©2015 IEEE.]

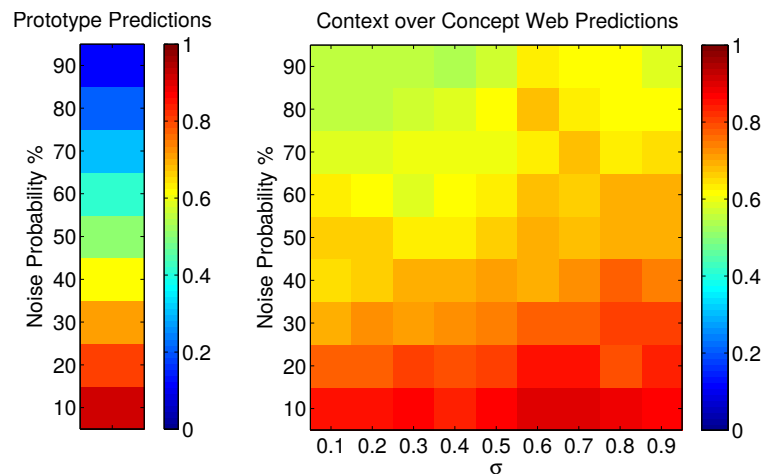


Figure 6.6: The performance of the individual prototype-based predictions, versus context enhanced concept web predictions, under artificially added noise, presented as prediction accuracies scaled to [0,1]. The noise probability denotes the probability of artificial noise being added to each single concept, via reversing its prototype-predicted probability from  $p\%$  to *reversed* to  $(100 - p)\%$ .  $\sigma$  refers to the trade-off parameter in Equation 6.11. [Adapted from [195] ©2015 IEEE. Best viewed in color.]

can be beneficial in these settings.

Remembering our fundamental assumption that related objects occur together in context (which allowed us to develop an LDA-based model in the first place), the system can use context to revise and correct its previous judgments. The loop of (a) context deduction, (b) probabilistic update of concept web, and (c) reiteration of MRF, as described in Section 6.6 and Equation 6.11, is utilized for refining predictions in context. Combined results for all 15 test objects are demonstrated in Figure 6.5, which also show an improvement of performance for the context-guided recognition.

In all these results, however, the individual predictions made solely using prototypes are quite good already, thereby making it difficult to adequately estimate the benefits of using context. Hence we have conducted an additional set of experiments, depicted in Figure 6.6, under artificial noise specifically added to the prototype predictions. An average of the prediction accuracies (scaled to  $[0, 1]$ ) over 15 sets of experiments are shown. For each set of experiments, a noise probability is determined in the range  $[10\%, 90\%]$ , and each concept's prototype-predicted probability  $p\%$  is *reversed* to  $(100-p)\%$  with the specified noise probability. The  $\sigma$  trade-off parameter of Equation 6.11 is varied in the range  $[0.1, 0.9]$ . Figure 6.6 demonstrates that the system is quite resilient under increasing artificial noise: Combining information from many sources all of which contributes to the contextual analysis, the system is able to detect the context correctly and thereby correct individual wrong predictions using the majority vote.

### 6.8.5 Using Context, Part 3: Planning in Context

Finally, we show how contextual information can be useful in a planning task. It is known that humans hugely rely on contextual information for planning their actions [224], possibly due to a severely restricted working memory capacity [225,226], which results in efficient day-to-day planning, but maybe less-than-favorable performances in chess. The robots would also benefit from similar contextual guidance in planning.

To show how context can be used similarly in a robotic planning scenario, we provide



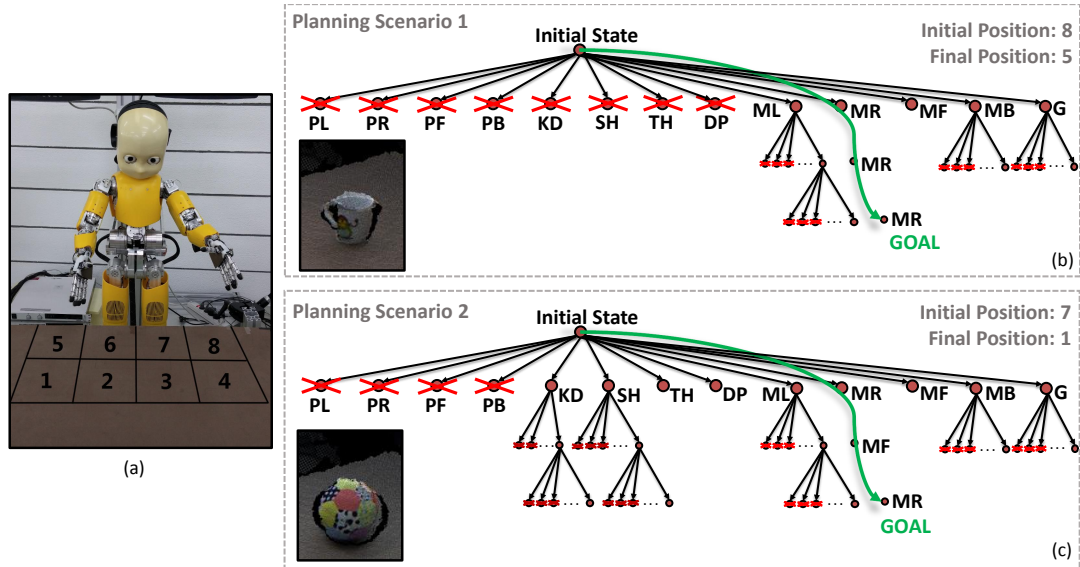


Figure 6.7: Pruning of forward planning trees by integrating contextual information. (a) iCub’s workspace. (b) First planning scenario. iCub is expected to move a cup from position 8 to position 5. Since pushing and knocking actions are dangerous in the kitchen context, these nodes are pruned without further expansion. Pruned branches are indicated with crosses. (c) Second scenario. iCub must bring a ball from position 7 to 1. Pushes are pruned, since pushing a ball causes it to roll down from the table. PX: Push left/right/forward/backward, MX: Move left/right/forward/backward, KD: Knock down, SH: Shake, TH: Throw, DP: Drop, G: Grasp. [Adapted from [195] ©2015 IEEE.]

two simple situations as proof-of-concept: The robot has to move two objects over a table (Figure 6.7(a)) from an initial to a goal position. Since the robot has learned the effect features of behaviors on training objects, it is theoretically able to expand a planning tree starting from the initial state and expanding behavior nodes until the goal condition is reached. These scenarios are simulated; however, the decisions of the robot are based on real world data: The robot plans according to the expected results of actions as learned by the verb prototypes. Although it does not physically *move* to perform the plan, theoretically the plans are executable. In other words, we are interested not in the physical success of the plans, but in the computational efficiency of producing these plans. We use the breadth-first forward planning approach depicted in Algorithm 5.

---

**Algorithm 5:** Breadth-first forward planning with context-dependent pruning.

[Adapted from [195] ©2015 IEEE.]

---

```
if goal position  $p_g =$  initial position  $p_i$  then
    return empty plan []
end if
QUEUE  $\leftarrow$   $[[b_1], \dots, [b_I]]$ ,  $\forall b_i \in \mathcal{B}_A$ ,  $\mathcal{B}_A$ : the set of applicable behaviors in the
current context
while QUEUE is not empty do
    pop PLAN from QUEUE
    - Predict the outcome of the behaviors in the PLAN:
    current position  $p_c \leftarrow$  initial position  $p_i$ 
    for all behavior  $b_i$  in PLAN do
        update current position:  $p_c \leftarrow b_i[p_c]$ 
    end for
    - Check whether we have reached the goal:
    if current position  $p_c =$  goal position  $p_g$  then
        return PLAN
    end if
    - Add possible behaviors in the current context as alternative plans:
    for all behavior  $b_j \in \mathcal{B}_A$  do
        if next position due to  $b_j$  ( $p_n \leftarrow b_j[p_c]$ ) is within table boundaries then
            push PLAN.append( $[b_j]$ ) to QUEUE
        end if
    end for
end while
return empty plan []
```

---

In the first scenario, Figure 6.7(b), the robot is asked to move a cup from position 8 to position 5. This goal can be achieved with three consecutive *move right* actions in our setting. A fully-expanded tree, therefore, would consist of three levels, and with a branching factor of 13, it will consist of  $13^0 + 13^1 + 13^2 + 13^3 = 2380$  nodes. However, given the contextual information of the scene, which is the *Kitchen* context, the robot can refrain from expanding the inappropriate behaviors in a Kitchen<sup>4</sup>, leaving only the *move left*, *move right*, *move forward*, *move backward* and *grasp* as possible actions to be expanded. Such an elimination gives a drastic reduction in the size of the planning tree, resulting in  $5^0 + 5^1 + 5^2 + 5^3 = 156$  nodes instead of 2380.

Figure 6.7(c) shows another scenario in the *Playroom* context. This time, the robot refrains from applying the *push* actions on associated objects, since balls, which are also in this context, tend to roll down and fall from the table when pushed. Therefore, the *push* nodes are pruned, leaving  $9^0 + 9^1 + 9^2 + 9^3 = 820$  nodes in the tree. We use a breadth-first forward planning scheme subject to context-dependent pruning.

Figure 6.8 compares un-pruned and pruned node counts for 10000 random scenarios in the move-over-the-table scenario presented above, presented for the three contexts

---

<sup>4</sup> Assuming we do not want to, for instance, *shake* a full cup.

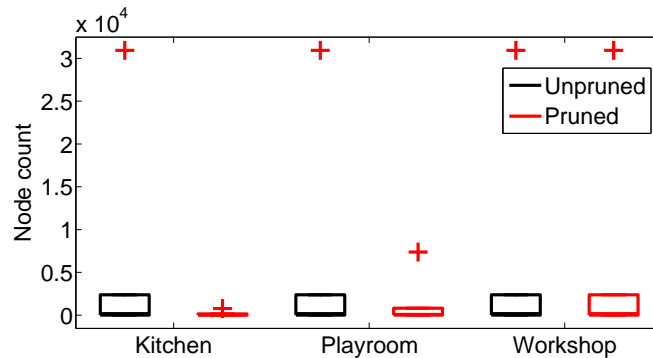


Figure 6.8: The node counts of *unpruned* vs. *pruned* planning trees of 10000 random scenarios, grouped by their contexts. The Kitchen context is subject to more pruning, as expected, due to a large number of *NA* behaviors. The Workshop context, on the other hand, is not subject to any pruning, since all behaviors are potentially applicable. In the plot, the boxes denote the data that fall between the 25<sup>th</sup> and 75<sup>th</sup> percentiles, and stars indicate the outliers. [Adapted from [195] ©2015 IEEE.]

separately. Each scenario is prepared by randomly determining a context, as well as initial and goal positions on the table environment, and then asking the robot to plan a behavior sequence from the initial to the goal position in this contextual background. Note that the amount of node reduction in these experiments depend on the randomly chosen target position. If the goal position is very close to the initial position, then relatively little reduction is possible, since the height of the planning tree will already be fairly shallow even in the unpruned case. However, if the random target is chosen sufficiently far from the initial position, which would normally require a very deep and wide planning tree, significant pruning is possible. The outliers in the graph correspond to such points. Note that the amount of pruning in the Kitchen case is greater than the Playroom case, since potentially greater number of actions are non-applicable in the Kitchen case. In the Workshop case, where all actions are applicable, there are no possible reductions.

The reductions shown here are only provided as proof-of-concepts, but it is clear how important it is for a robot to learn to prune its search trees in a real world setting. For a very limited robot of a small, or maybe even intermediate set of actions, considering each action for every situation might be an option, but for any robot who aims to operate in the real world, the actions will be so varied and planning chains will necessarily be so long that even most basic reductions (*i.e.*, no need to consider opening the kitchen door for heating a glass of milk) will be of critical importance.

### **6.8.6 Running Time Performance of the System**

The whole system is able to work close to real-time: 10 test runs with non-optimized code on a standard desktop PC (i5 core, 8GB RAM) provided an average running time of  $209.82ms \pm 3.38ms$  for the detection of context with Incremental LDA, and  $1395.22ms \pm 15.31ms$  for the convergence of the concept web.

## **6.9 Summary**

In this part, we proposed a method for formalizing, learning, and using context. For modeling context, we employed and extended Latent Dirichlet Allocation (LDA), a

widely-used topic model in the computational linguistics literature. Unlike the existing applications of LDA in robotics for, *e.g.*, word learning, where LDA is directly applied onto low-level sensorimotor data, we were motivated by the concept web hypotheses in humans and its computational advantages to apply LDA onto a concept web model that we developed in our previous work using Markov Random Fields.

We demonstrated the following important aspects:

- In an unsupervised fashion, the robot can learn context even if the number of contexts is not given. By using an online version of the Gibbs sampler proposed in the article, the robot can work online to process new observations and can tackle new contexts. By a systematic analysis, we show that the model finds the *correct* number of contexts in different settings.
- The robot can use the learned contexts to improve its performance in cognitive tasks. In the article, we showed this aspect for object recognition and planning.
- Finally we show how learning context over a web of abstracted concepts is *easier* and provides better performance for an LDA-based architecture, which deals with the sensorimotor complexity of real world better than raw features themselves.



## CHAPTER 7

### CONCLUSION

In this thesis, we have addressed an important problem in cognitive systems, that of modeling a concept web in a similar fashion to us, humans. The web is constructed based on the co-occurrences of concepts from the interactions of the robot, and modeled using Markov Random Fields. Since the resulting web is a cyclic graph, inferences are made using Loopy Belief Propagation, as is widely done in the literature.

We have demonstrated that, given an observation of an object, our robot can activate in its “brain” the relevant noun concepts, adjective concepts, verb concepts (describing what behaviors can be applied on the object) as well as the words that can be used for the object. Moreover, given an interaction on an object or in fact, an interaction without an object (that would normally take an object), the robot can activate the necessary concepts in the web as well. Being linked to language, perception and motor (action) spaces, the concept web allows activation of relevant information from and to any modality. As we reviewed in detail in Section 2.2.1, such a concept web is very much in line with findings from neuroscience.

Moreover, we showed that such a web allows the robot to make a better interpretation of the environment. By using the co-occurrences from other concepts, wrongly predicted concepts can be corrected, and confidences of correct predictions can be increased.

We have also showed how a humanoid robot can model, learn and use context. For modeling context, we employed and extended Latent Dirichlet Allocation (LDA), a widely-used topic model in the computational linguistics literature. Unlike the exist-

ing applications of LDA in robotics for, *e.g.*, word learning, where LDA is directly applied onto low-level sensorimotor data, we were motivated by the concept web hypotheses in humans and its computational advantages to apply LDA onto a concept web model that we developed in our previous work using Markov Random Fields.

We demonstrated the robot can learn context in an unsupervised manner, even if the number of contexts is not given. By using an online version of the Gibbs sampler proposed in the article, the robot can work online to process new observations and can tackle new contexts. By a systematic analysis, we show that the model finds the *correct* number of contexts in different settings.

We have further demonstrated how the robot can use the learned contexts to improve its performance in cognitive tasks. We have selected the object recognition and planning tasks for showcasing this ability.

Finally we show how learning context over a web of abstracted concepts is *easier* and provides better performance for an LDA-based architecture, which deals with the sensorimotor complexity of real world better than raw features themselves.

## 7.1 Discussion

There are naturally many design choices in any architecture of reasonable size. The distinguishing features of the presented model are:

- Basing the conceptualization on a graph-like structure, in which every concept is connected to all other related ones, as compared to a tree structure with only parent-child connections
- Representing spatial concepts as nodes in the concept web, rather than simple links connecting other nodes
- Building contextual understanding on top of the concept web
- Aiming for a lifelong and developmental learning of context
- Proposing a context-based pruning for real-time planning



Below, we discuss these features in the light of the state-of-the-art findings.

### 7.1.1 A Web of Concepts

There are several ontology-based studies in the literature, *e.g.* [118–120, 151], where concepts are represented in tree-structures. These ontologies are effective in representing structured information partially, but there are many missing semantic links that are pragmatic and useful for us, humans. For instance, “cup” and “glass” would be related, coming from a single ancestor, which might be “container”. However, “cup” and “coffee” will not be connected in such a tree-shaped representation. Yet, the cup-and-coffee connection is very real for us. Whenever we want some coffee, the first thing we will be looking for will be a cup - without a cup, coffee is not drinkable. Similarly, the connection between “cup” and “coffee” leads also to a third concept, namely that of “drinking” affordance. And to many more as well, coffee and cup together afford “waking up” when it is “early” in the morning, “chit chat” when “together with friends”, or “nausea” when one’s stomach is “empty”. Often times, these are not strictly structural, but rather *functional* and pragmatic links that enable us to behave rapidly and efficiently in the real world. Therefore, we posit that it is important to represent these multitude of connections efficiently and natively. As hypothesized by Bar [5], we also build our system on the observation that commonly co-occurring concepts are generally linked. Thus, we propose a densely-connected web structure that enables efficient spreading of information is of primary importance for adequately representing our understanding of the world.

### 7.1.2 Spatial Concepts as Nodes

Another feature of our system is regarding the spatial relations, or prepositional concepts, as nodes themselves within the graph. Representing spatial concepts as nodes is synonymous with making them *first-order* members of the graph. They are similar to the noun, adjective, and verb concepts in that (1) They are initialized to naive predictions, (2) but then iterated over and regulated according to other perceived concepts, (3) all the while regulating the activations of the other concepts themselves.

This approach of representing spatial relations as first-order concepts is more in line with the proposals of Coventry *et al.* [46–49] and Garrod *et al.* [50,51], rather than for instance Anand *et al.* [68] and Misra *et al.* [69], in that spatial relations are viewed as complex and prone to personal and functional interpretation, rather than being fully defined by simple geometric relations. The support for our hypothesis is the fact that humans also accumulate significant physical understanding of the world, which they implicitly and perpetually use for assessing the scene. The real world is an incredibly consistent place in terms of physical rules, that is why detecting and reusing physical patterns is a feasible shortcut. As demonstrated by the above studies, this absolute reliance on learned patterns of physicality manifest itself explicitly even in our language use. We therefore propose that spatial concepts deserve to be explicit nodes in our densely-connected web structure, being distilled themselves from a rich physical understanding and carrying experience-based connections to other concepts.

### **7.1.3 Basing Context on the Concept Web**

We base our context representation on the proposed densely-connected concept web. Our motivation comes from the hypothesis that human cognition is mostly based on concepts [1, 2], and that concepts commonly occurring together is what gives rise to the context [5, 143] (also see [78] for statistical relations of concepts directly affecting cortical representations). This approach also boasts computational advantages: In Chapter 4, we have showed how concept web enables a superior performance of object recognition and conceptualization as compared to a raw-feature based scheme. In Chapter 6, we go on to conduct an explicit evaluation of the concept-web based formulation against raw-features based modeling, which provides further evidence regarding the benefits of utilizing structured information from the concept web. We demonstrate how the concept web provides better performance with significantly fewer training examples, as well as reduced sensitivity against system parameters. These advantages are due to its abstraction capability: The real world presents an overwhelming amount of complex information, which needs some structure to be imposed before statistically significant relations can be discovered. This is argued to be the driving reason of conceptualization in humans as well (*e.g.*, [1, 2, 8, 227–231], for a slightly different but interesting argument, see also [232].)

#### **7.1.4 Lifelong and Developmental Learning in Robots**

Lifelong learning in robots (*e.g.*, [233]) aims to overcome important limitations that prevent the robots from operating in real-life environments: Specifically, [233] warns that (1) a human designer cannot always accurately predict the robot's world, (2) even if such a prediction were possible, a corresponding model would be tedious, possibly infeasible to hand-code, and (3) even if such a model were to be constructed, planning on it would likely be intractable except in severely restricted cases [164, 165]. [22, 25, 30–32, 234–238] envision a paradigm shift as the solution: The robots need to go through a prolonged developmental phase [237], in which they will discover the world through their own interactions, and as they gradually encounter more complex problems [235, 238], they adapt their previous knowledge to gain new insight [239]. There are two facets vital to this paradigm: First, *learning can never end* [233]: The robot needs to gain knowledge in an incremental manner, building more complex skills on previously gained ones. Otherwise, the real world would simply be too complex to learn. Second, *the developer can never know* [25, 240]: It is not possible to predict the numerous environments a robot can encounter with. Therefore, as [240] mentions, an artificial system has to have the means of checking its own knowledge constantly. Any priori bias hard-coded by the designer can become false at any time, for instance, assuming a sensory or motor failure of the robot. [25] outlines a roadmap by suggesting that each and every assumption of a robot must be made explicit, so that it will have a chance of verifying these assumptions in a constantly changing environment. We try to adopt this guideline, by endowing the robot with a means of checking its own context model in a changing environment. Not only do we want it to discover the “correct” setting of contexts, but we also want it to monitor this knowledge continuously. Should a new context appear in its world, it will then be able to react accordingly.

#### **7.1.5 Planning in the Real World**

Bylander [164] and Chapman [165] show that planning is intractable in the general sense, unless it is restricted severely, for instance, to propositional planning with strictly positive preconditions and exactly one postcondition. Such restricted cases

can be defined to reduce the planning problem to a polynomial-time subset; however, small deviations make the problem intractable again: *e.g.*, the NP-hard problem of allowing two postconditions along with one precondition, or the NP-complete problem of one strictly positive postcondition along with one precondition. As Bylander [164] and Hendler [166] note, it is difficult to describe any interesting world in propositional logic, let alone such restrictions for the sake of tractability. We have to find a workaround. We propose that this workaround can be, and for humans is, context [167–170].

Also supporting our hypothesis is the work of Siegler and colleagues, *e.g.*, [241], who, from a developmental point of view, stresses how important context is in helping children choose which skill or problem solving strategy to apply in a certain situation. So important is this process of choosing, he claims, that the question is not “whether children ‘have’ a concept or strategy or theory at a given age”, but it is rather “the set of conceptualizations and strategies and theories that children know and the mechanisms by which they *choose* among them” [241] (emphasis added).

## 7.2 Limitations and Future Work

Overall, we provide promising results that a learning scheme which *includes* background information, instead of leaving it out, is feasible *and* useful for a robot when dealing with the real world. Our work can be extended in several directions.

The experiments were performed on real objects, although the settings are not realistic. This limitation was due to the interaction capabilities of iCub: iCub cannot walk and is confined to a table-top environment. Moreover, due to its delicate hands and the limited precision of the touch sensors on the hands, the range of objects that can be interacted with was limited to light-weight and convex objects. This also restricted us in the varieties of contexts. However, LDA is shown to scale up extremely well in natural language processing settings, where it could be tested with huge corpora (*e.g.*, [3, 208]) as well in a number of other complicated real-life scenarios including functional miRNA–mRNA regulatory modules identification [214] and fraud detection [212]; therefore, we believe that our framework will scale well in realistic

robotics settings. For quantitatively evaluating the system’s robustness against larger and noisier datasets, though, a sensitivity analysis needs to be conducted systematically. In particular, the amount of complexity, which can be added to the system without significantly disturbing its performance, needs to be identified. The more insensitive to increasing complexity the system is, the more robust it will prove to be in real life situations. During such a sensitivity analysis, the system can also be more precisely evaluated by a combination of success/fail metrics, such as precision/recall graphics, in order to quantitatively distinguish the tendency to false negatives from false positives.

In Incremental-LDA, we assumed that the number of contexts can only *increase* in the environment, and therefore it is not necessary to check if the context count  $K$  can go down. We observe similar assumptions in the literature, *e.g.*, [222], where the number of topics can only *increase* in time. We believe that there is no reason for a biological cognitive agent to remove learned contexts from its system; although they might be merged as new contexts or split into sub-contexts, the only case where the number of contexts might decrease is when the agent forgets learned associations.

It should also be noted that, although our current concept web is composed of noun, adjective, and verb concepts, a cognitive model should include spatial, temporal, adverb, and social concepts as well. With the incorporation of these types of concepts in our concept web, contexts related to their semantics will also be able to manifest themselves in our model.

Moreover, a more realistic model would need to account for super-ordinate, or “higher-order” concepts as well, such as “animal”, or “utensil”. Similarly, there can be “sub-ordinate” concepts, such as “terrier dog”, or “cup with a handle”. Currently, the concept web is designed to include a single level, called the *basic level or conceptualization* [242]. Incorporating a hierarchical conceptualization mechanism which takes other concepts as its input would extend this single level, and allow a richer semantic structure.

For the addition of such hierarchical organization, the links within the concept web need to be extended to have different *types*. There are currently two explicit link types in the system: There are undirected, fully associative links between the related con-

cept instantiations, associated with an object, and directed links between the objects and spatial relations. However, there can be more types: A fully grown concept web needs to be able to represent *is-a* and *has-a* type relations as well, which will enrich the semantics, by allowing a hierarchical organization of the concepts.

Moreover, a similar hierarchical organization would also be beneficial for the contextual modeling. Contexts can be hierarchical as well, there can be super-concepts or sub-concepts, known contexts can split in time into more specific ones as more experience is gained, or previously separate contexts can merge into one. Therefore, the contextual representation would also benefit from a hierarchical extension.

Another requirement that is currently missing from the system is temporal knowledge. The dimension of temporality needs to be introduced, which will serve to distinguish between different scenes. In addition, old information might be reconsidered in time under the light of new evidence, provided that the system has an explicit understanding of the passing time.

Although we have strived for online and fully-unsupervised discovering the *contexts*, a similar flexibility is ultimately necessary for the discovery of the *concepts* as well: Currently we are employing a supervised learning scheme of previously determined concepts, which is not very feasible from a developmental point of view. Finally, our proof-of-concept results should be investigated in real-time scenarios: Since there is reason to believe that context can be advantageous in recognition and planning scenarios, it should be able to ease the tasks in more real-life settings, possibly with more contexts, more realistic objects, and more complicated tasks. Such real-life settings may even be too difficult to operate at all without contextual aid, and will naturally demonstrate the benefits of using context more clearly.

In such an online setting, it would also be more naturalistic to start with a fully connected concept web, and let the system learn and evolve its own connection weights with experience in an online manner, instead of imposing to the web the required list of connections through a batch set of examples. The current version of the concept web learns offline, during a specially designated training time. At the end of this period, it emerges with a set of connectedness rules, which are crystallized in the concept web, not to be modified afterwards. Instead, a more naturalistic setting would

be the lifelong learning and adaptation of connections. Take a infant, for instance, who begins her life with the idea that everything is gnaw-able, only to discover when she grows up that some things are better off not being gnawed. Moreover, the current concept web has the notion of binary connections: Two concepts are either connected to each other, or not. However, more flexible, possibly condition-dependent connections could also be more consistent with the idea of contextual interpretation. *Cup* and *throw*, for instance, might be connected in the playroom context, consisting of plastic cups, and unconnected in the kitchen context with “real” cups.

Finally, it is without doubt that in this study, we have merely scratched the surface of the possible effects of context. There remains so much to be investigated. An interesting direction, for instance, is how context naturally affects the perception of certain adjectives, whose meanings can be extremely flexible. The adjective *big*, for instance, can be used to describe a house, a football stadium, or the universe, or even the human heart, in different contexts. Thus, we cannot say that *big* has a certain, fixed meaning; it simply means vastly different scales in different contexts, and we are very well capable of figuring out the right connotation. Other adjectives, on the other hand, can be rather independent of the context; take *sticky*, for example, or *brittle*. Finally, there are also certain adjectives that seem to have enforcing bonds with certain contexts: A toy in a playroom had better not be *sharp*, or it can be dangerous. When we hear of a toy described as sharp, such as a *sharp pirate sword*, we immediately reason it to be mock-sharpness: Literal sharpness is banned from the playroom. This flexible interplay between the adjectives and context seems to be a crucial part of their understanding, after all, one which needs further investigation.

Moreover, the susceptibility to context-dependent interpretations is not restricted to adjective concepts. We have already shown proof-of-concept results of the context-dependent pruning of irrelevant actions, for the sake of “real-time” planning. However, the effect of context on action selection goes much further than that. As pointed out by Siegler and colleagues [241], there is a developmental step between children leaning a skill as-is, and their internalizing the skill so much that they can adapt it to the necessities of the context. The stirring motion, which is initially learned with a toy spoon, can as well be transferred to mixing mud with a twig. Context can easily arise different affordances of objects, or trigger the transfer of previously internalized

skills for completely different goals. Even from these observations only, it is obvious that we are still a long way from a complete utilization of context.

Our results demonstrate promising potential on the path to competent understanding of conceptualization and context. Further work will continue to bring more light, until the time when these mysterious cognitive faculties will hopefully be understood decently, and imitated successfully.



## REFERENCES

- [1] George Lakoff. *Women, fire, and dangerous things: What categories reveal about the mind*. Cambridge Univ Press, 1990.
- [2] Terrence Deacon. *The symbolic species: the co-evolution of language and the human brain*, 1997.
- [3] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.
- [4] John A. Bargh, Mark Chen, and Lara Burrows. Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, 71(2):230, 1996.
- [5] Moshe Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004.
- [6] Hande Çelikkanat, Güner Orhan, and Sinan Kalkan. A probabilistic concept web on a humanoid robot. *IEEE Transactions on Autonomous Mental Development*, 7(2):92–106, 2015.
- [7] Jerome Seymour Bruner, Jacqueline J. Goodnow, and George A. Austin. *A study of thinking*. RE Krieger Publishing Company, 1977.
- [8] Ulrike Hahn and Nick Chater. Concepts and similarity. *Knowledge, concepts and categories*, pages 43–92, 1997.
- [9] Eleanor H. Rosch. Natural categories. *Cognitive psychology*, 4(3):328–350, 1973.
- [10] Peter Gärdenfors. *Conceptual spaces: The geometry of thought*. MIT press, 2004.
- [11] Liane Gabora, Eleanor Rosch, and Diederik Aerts. Toward an ecological theory of concepts. *Ecological Psychology*, 20(1):84–116, 2008.
- [12] John K. Kruschke and Mark K. Johansen. A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(5):1083, 1999.
- [13] Yves Rosseel. Mixture models of categorization. *Journal of Mathematical Psychology*, 46(2):178–210, 2002.

- [14] Teuvo Kohonen. *Learning vector quantization*. Springer, 1997.
- [15] Stevan Harnad. The symbol grounding problem. *Physica, D*, 42:335–346, 1990.
- [16] Lawrence W. Barsalou. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22:577–609, 1999.
- [17] Luc Steels. Evolving grounded communication for robots. *Trends in Cognitive Science*, 7(7):308–312, 2003.
- [18] Angelo Cangelosi and Thomas Riga. An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. *Cognitive Science*, 30(4):673–689, 2006.
- [19] Angelo Cangelosi. Grounding language in action and perception: From cognitive agents to humanoid robots. *Physics of Life Reviews*, 7(2):139–151, 2010.
- [20] Arthur M. Glenberg and Michael P. Kaschak. Grounding language in action. *Psychonomic Bulletin & Review*, 9(3):558–565, 2002.
- [21] Doreen Jirak, Mareike M. Menz, Giovanni Buccino, Anna M. Borghi, and Ferdinand Binkofski. Grasping language—a short story on embodiment. *Consciousness and cognition*, 19(3):711–720, 2010.
- [22] Angelo Cangelosi, Giorgio Metta, Gerhard Sagerer, Stefano Nolfi, Christopher Nehaniv, Kerstin Fischer, Jun Tani, Tony Belpaeme, Giulio Sandini, and Francesco Nori. Integration of action and language knowledge: A roadmap for developmental robotics. *IEEE Transactions on Autonomous Mental Development*, 2(3):167–195, 2010.
- [23] Anna M. Borghi, Arthur M. Glenberg, and Michael P. Kaschak. Putting words in perspective. *Memory & Cognition*, 32(6):863–873, 2004.
- [24] Anna M. Borghi. Object concepts and embodiment: Why sensorimotor and cognitive processes cannot be separated. *La nuova critica*, 15(4):447–472, 2007.
- [25] Alexander Stoytchev. Some basic principles of developmental robotics. *IEEE Transactions on Autonomous Mental Development*, 1(2):122–130, 2009.
- [26] Joanna J. Bryson. Embodiment versus memetics. *Mind & Society*, 7(1):77–94, 2008.
- [27] Nikolaos Mavridis and Deb Roy. Grounded situation models for robots: Where words and percepts meet. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2006)*, pages 4690–4697. IEEE, 2006.

- [28] Onur Yürüten, Erol Şahin, and Sinan Kalkan. The learning of adjectives and nouns from affordance and appearance features. *Adaptive Behavior*, 21(6):437–451, 2013.
- [29] Luc Steels. The recruitment theory of language origins. In *Emergence of Communication and Language*, pages 129–150. Springer, 2007.
- [30] Angelo Cangelosi. The grounding and sharing of symbols. *Pragmatics & Cognition*, 14(2):275–285, 2006.
- [31] Angelo Cangelosi and Stevan Harnad. The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of communication*, 4(1):117–142, 2000.
- [32] Angelo Cangelosi. Evolution of communication and language using signals, symbols, and words. *IEEE Transactions on Evolutionary Computation*, 5(2):93–101, 2001.
- [33] Luc Steels. The talking heads experiment. 1999.
- [34] Tony Belpaeme and Anthony Morse. Word and category learning in a continuous semantic domain: Comparing cross-situational and interactive learning. *Advances in Complex Systems*, 15(03n04), 2012.
- [35] Takashi Hashimoto and Akira Masumi. Learning and transition of symbols: Towards a dynamical model of a symbolic individual. In C. L. Nehaniv, C. Lyon, and A. Cangelosi, editors, *Emergence of Communication and Language*, page 223–236. Springer, 2007.
- [36] Jivko Sinapov, Connor Schenck, Kerrick Staley, Vladimir Sukhoy, and Alexander Stoytchev. Grounding semantic categories in behavioral interactions: Experiments with 100 objects. *Robotics and Autonomous Systems*, 2012.
- [37] Giorgio Metta, Giulio Sandini, David Vernon, Lorenzo Natale, and Francesco 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*, pages 50–56. ACM, 2008.
- [38] Sinan Kalkan, Nilgün Dağ, Onur Yürüten, Anna M. Borghi, and Erol Şahin. Verb concepts from affordances. *Interaction Studies Journal*, 2013.
- [39] Mathias Rudolph, Manuel Mühlig, Michael Gienger, and H-J Bohme. Learning the consequences of actions: Representing effects as feature changes. In *IEEE International Conference on Emerging Security Technologies (EST)*, pages 124–129, 2010.
- [40] Leslie G. Ungerleider. Two cortical visual systems. *Analysis of Visual Behavior*, pages 549–586, 1982.

- [41] Hanna Damasio, Thomas J. Grabowski, Daniel Tranel, Laura L.B. Ponto, Richard D. Hichwa, and Antonio R. Damasio. Neural correlates of naming actions and of naming spatial relations. *Neuroimage*, 13(6):1053–1064, 2001.
- [42] A. David Milner and Melvyn A. Goodale. *The visual brain in action*. 1995.
- [43] Benjamin Kuipers. The spatial semantic hierarchy. *Artificial Intelligence*, 119(1):191–233, 2000.
- [44] Barbara Landau and Ray Jackendoff. ‘What’ and ‘where’ in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16, 1993.
- [45] Kenny R. Coventry, Merce Prat-Sala, and Lynn Richards. The interplay between geometry and function in the comprehension of over, under, above, and below. *Journal of Memory and Language*, 44(3):376–398, 2001.
- [46] Kenneth R. Coventry. *Spatial prepositions and functional relations: The case for minimally specified lexical entries*. PhD thesis, University of Edinburgh, 1993.
- [47] Kenny R. Coventry. Function, geometry and spatial prepositions: Three experiments. *Spatial Cognition and Computation*, 1(2):145–154, 1999.
- [48] Kenny R. Coventry, Richard Carmichael, and Simon C. Garrod. Spatial prepositions, object-specific function, and task requirements. *Journal of Semantics*, 11(4):289–311, 1994.
- [49] Kenny R. Coventry and Mercè Prat-Sala. Object-specific function, geometry, and the comprehension of in and on. *European Journal of Cognitive Psychology*, 13(4):509–528, 2001.
- [50] Simon C. Garrod and Anthony J. Sanford. Discourse models as interfaces between language and the spatial world. *Journal of Semantics*, 6(1):147–160, 1988.
- [51] Simon Garrod, Gillian Ferrier, and Siobhan Campbell. In and on: investigating the functional geometry of spatial prepositions. *Cognition*, 72(2):167–189, 1999.
- [52] Kerstin Fischer. The role of users’ concepts of the robot in human-robot spatial instruction. In *Spatial Cognition V Reasoning, Action, Interaction*, pages 76–89. Springer, 2007.
- [53] Eva Stopp, Klaus-Peter Gapp, Gerd Herzog, Thomas Laengle, and Tim C. Lueth. Utilizing spatial relations for natural language access to an autonomous mobile robot. volume 861, page 39. Springer Science & Business Media, 1994.

- [54] Kevin Gold, Marek Doniec, and Brian Scassellati. Learning grounded semantics with word trees: Prepositions and pronouns. In *IEEE International Conference on Development and Learning*, pages 25–30, 2007.
- [55] Reinhard Moratz and Thora Tenbrink. Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. *Spatial cognition and computation*, 6, 2006.
- [56] Deb Roy, Kai-Yuh Hsiao, and Nikolaos Mavridis. Mental imagery for a conversational robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 34(3):1374–1383, 2004.
- [57] Nico Van de Weghe. *Representing and reasoning about moving objects: A qualitative approach*. PhD thesis, Ghent University, 2004.
- [58] Nico Van de Weghe, Anthony G. Cohn, Guy De Tre, and Philippe De Maeyer. A qualitative trajectory calculus as a basis for representing moving objects in geographical information systems. *Control and Cybernetics*, 35(1):97, 2006.
- [59] Marc Hanheide, Annika Peters, and Nicola Bellotto. Analysis of human-robot spatial behaviour applying a qualitative trajectory calculus. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2012)*, pages 689–694, 2012.
- [60] Nicola Bellotto. Robot control based on qualitative representation of human trajectories. 2012.
- [61] Nicola Bellotto, Marc Hanheide, and Nico Van de Weghe. Qualitative design and implementation of human-robot spatial interactions. In *Social Robotics*, pages 331–340. Springer, 2013.
- [62] Konstantinos Iliopoulos, Nicola Bellotto, and Nikolaos Mavridis. From sequence to trajectory and vice versa: solving the inverse qtc problem and coping with real-world trajectories. In *2014 AAAI Spring Symposium Series*, 2014.
- [63] Chuho Yi, Il Hong Suh, Gi Hyun Lim, and Byung-Uk Choi. Bayesian robot localization using spatial object contexts. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2009)*, pages 3467–3473, 2009.
- [64] Dave Golland, Percy Liang, and Dan Klein. A game-theoretic approach to generating spatial descriptions. In *Conference on empirical methods in natural language processing*, pages 410–419, 2010.
- [65] Sergio Guadarrama, Lorenzo Riano, Dave Golland, Daniel Gouhring, Yangqing Jia, Dan Klein, Pieter Abbeel, and Trevor Darrell. Grounding spatial relations for human-robot interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2013)*, pages 1640–1647, 2013.

- [66] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth J. Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*, 2011.
- [67] Hande Çelikkanat, Erol Şahin, and Sinan Kalkan. Integrating spatial concepts into a probabilistic concept web. In *IEEE International Conference on Advanced Robotics*, 2015.
- [68] Abhishek Anand, Hema Swetha Koppula, Thorsten Joachims, and Ashutosh Saxena. Contextually guided semantic labeling and search for three-dimensional point clouds. *The International Journal of Robotics Research*, 2012.
- [69] Dipendra K. Misra, Jaeyong Sung, Kevin Lee, and Ashutosh Saxena. Tell me Dave: Context-sensitive grounding of natural language to manipulation instructions. In *Robotics Science and Systems (RSS 2014)*, 2014.
- [70] Gertrude H. Eggert. *Wernicke’s works on aphasia: A sourcebook and review*, volume 1. Mouton The Hague, 1977.
- [71] Matthew A. Lambon Ralph. Neurocognitive insights on conceptual knowledge and its breakdown. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1634):20120392, 2014.
- [72] Robert F. Goldberg, Charles A. Perfetti, and Walter Schneider. Perceptual knowledge retrieval activates sensory brain regions. *The Journal of Neuroscience*, 26(18):4917–4921, 2006.
- [73] Marion L. Kellenbach, Matthew Brett, and Karalyn Patterson. Large, colorful, or noisy? attribute-and modality-specific activations during retrieval of perceptual attribute knowledge. *Cognitive, Affective, & Behavioral Neuroscience*, 1(3):207–221, 2001.
- [74] Friedemann Pulvermüller. *The neuroscience of language: on brain circuits of words and serial order*. Cambridge University Press, 2002.
- [75] Olaf Hauk, Ingrid Johnsrude, and Friedemann Pulvermüller. Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, 41(2):301–307, 2004.
- [76] Friedemann Pulvermüller. Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, 6(7):576–582, 2005.
- [77] Linda L Chao and Alex Martin. Representation of manipulable man-made objects in the dorsal stream. *Neuroimage*, 12(4):478–484, 2000.

- [78] Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel A. Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, 2008.
- [79] Alice J O’Toole, Fang Jiang, Hervé Abdi, and James V. Haxby. Partially distributed representations of objects and faces in ventral temporal cortex. *Cognitive Neuroscience, Journal of*, 17(4):580–590, 2005.
- [80] Antonio R. Damasio. Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, 33(1):25–62, 1989.
- [81] Alexander G. Huth, Shinji Nishimoto, An T. Vu, and Jack L. Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012.
- [82] Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, 17(11):4302–4311, 1997.
- [83] Gregory McCarthy, Aina Puce, John C. Gore, and Truett Allison. Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience*, 9(5):605–610, 1997.
- [84] Galia Avidan, Uri Hasson, Rafael Malach, and Marlene Behrmann. Detailed exploration of face-related processing in congenital prosopagnosia: 2. functional neuroimaging findings. *Journal of Cognitive Neuroscience*, 17(7):1150–1167, 2005.
- [85] V.P. Clark, K. Keil, J. Ma Maisog, Susan Courtney, Leslie G. Ungerleider, and James V. Haxby. Functional magnetic resonance imaging of human visual cortex during face matching: a comparison with positron emission tomography. *Neuroimage*, 4(1):1–15, 1996.
- [86] Eric Halgren, Anders M. Dale, Martin I. Sereno, Roger B.H. Tootell, Ksenija Marinkovic, and Bruce R. Rosen. Location of human face-selective cortex with respect to retinotopic areas. *Human Brain Mapping*, 7(1):29–37, 1999.
- [87] Reza Rajimehr, Jeremy C. Young, and Roger B.H. Tootell. An anterior temporal face patch in human cortex, predicted by macaque maps. *Proceedings of the National Academy of Sciences*, 106(6):1995–2000, 2009.
- [88] Doris Y. Tsao, Sebastian Moeller, and Winrich A. Freiwald. Comparing face patch systems in macaques and humans. *Proceedings of the National Academy of Sciences*, 105(49):19514–19519, 2008.

- [89] Paul E. Downing, Yuhong Jiang, Miles Shuman, and Nancy Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473, 2001.
- [90] Marius V. Peelen and Paul E. Downing. Selectivity for the human body in the fusiform gyrus. *Journal of Neurophysiology*, 93(1):603–608, 2005.
- [91] Rebecca F. Schwarzlose, Chris I. Baker, and Nancy Kanwisher. Separate face and body selectivity on the fusiform gyrus. *The Journal of Neuroscience*, 25(47):11055–11059, 2005.
- [92] Geoffrey K. Aguirre, E. Zarahn, and M. D’Esposito. An area within human ventral cortex sensitive to “building” stimuli: evidence and implications. *Neuron*, 21(2):373–383, 1998.
- [93] Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998.
- [94] Marius V. Peelen, Alison J. Wiggett, and Paul E. Downing. Patterns of fmri activity dissociate overlapping functional brain areas that respond to biological motion. *Neuron*, 49(6):815–822, 2006.
- [95] Kevin A. Pelphrey, James P. Morris, Charles R. Michelich, Truett Allison, and Gregory McCarthy. Functional anatomy of biological motion perception in posterior temporal cortex: an fmri study of eye, mouth and hand movements. *Cerebral cortex*, 15(12):1866–1876, 2005.
- [96] Asim Roy. On findings of category and other concept cells in the brain: Some theoretical perspectives on mental representation. *Cognitive Computation*, pages 1–6, 2014.
- [97] Itzhak Fried, Katherine A. MacDonald, and Charles L. Wilson. Single neuron activity in human hippocampus and amygdala during recognition of faces and objects. *Neuron*, 18(5):753–765, 1997.
- [98] Gabriel Kreiman, Christof Koch, and Itzhak Fried. Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience*, 3(9):946–953, 2000.
- [99] Hiroto Kawasaki, Ralph Adolphs, Hiroyuki Oya, Christopher Kovach, Hanna Damasio, Olaf Kaufman, and Matthew Howard Iii. Analysis of single-unit responses to emotional scenes in human ventromedial prefrontal cortex. *Journal of Cognitive Neuroscience*, 17(10):1509–1518, 2005.
- [100] Katalin M. Gothard, Francesco P. Battaglia, Cynthia A. Erickson, Kevin M. Spitzer, and David G. Amaral. Neural responses to facial expression and face identity in the monkey amygdala. *Journal of Neurophysiology*, 97(2):1671–1683, 2007.



- [101] Rodrigo Quian Quiroga. Concept cells: the building blocks of declarative memory functions. *Nature Reviews Neuroscience*, 13(8):587–597, 2012.
- [102] Rodrigo Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005.
- [103] Rodrigo Quian Quiroga, Alexander Kraskov, Christof Koch, and Itzhak Fried. Explicit encoding of multimodal percepts by single neurons in the human brain. *Current Biology*, 19(15):1308–1313, 2009.
- [104] Nanthia Suthana and Itzhak Fried. Percepts to recollections: insights from single neuron recordings in the human brain. *Trends in Cognitive Sciences*, 16(8):427–436, 2012.
- [105] Longnian Lin, Guifen Chen, Hui Kuang, Dong Wang, and Joe Z. Tsien. Neural encoding of the concept of nest in the mouse brain. *Proceedings of the National Academy of Sciences*, 104(14):6066–6071, 2007.
- [106] Ikue Yoshida and Kensaku Mori. Odorant category profile selectivity of olfactory cortex neurons. *The Journal of Neuroscience*, 27(34):9105–9114, 2007.
- [107] Yasuko Sugase, Shigeru Yamane, Shoogo Ueno, and Kenji Kawano. Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, 400(6747):869–873, 1999.
- [108] Chou P. Hung, Gabriel Kreiman, Tomaso Poggio, and James J. DiCarlo. Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–866, 2005.
- [109] Gabriel Kreiman, Itzhak Fried, and Christof Koch. Single-neuron correlates of subjective vision in the human medial temporal lobe. *Proceedings of the national academy of sciences*, 99(12):8378–8383, 2002.
- [110] Rodrigo Quian Quiroga, Roy Mukamel, Eve A. Isham, Rafael Malach, and Itzhak Fried. Human single-neuron responses at the threshold of conscious recognition. *Proceedings of the National Academy of Sciences*, 105(9):3599–3604, 2008.
- [111] Karalyn Patterson, Peter J. Nestor, and Timothy T. Rogers. Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, 8(12):976–987, 2007.
- [112] Alex Martin. The representation of object concepts in the brain. *Annu. Rev. Psychol.*, 58:25–45, 2007.
- [113] Matthew A. Lambon Ralph, Karen Sage, Roy W. Jones, and Emily J. Mayberry. Coherent concepts are computed in the anterior temporal lobes. *Proceedings of the National Academy of Sciences*, 107(6):2717–2722, 2010.

- [114] Hanna Damasio, Daniel Tranel, Thomas Grabowski, Ralph Adolphs, and Antonio R. Damasio. Neural systems behind word and concept retrieval. *Cognition*, 92(1):179–229, 2004.
- [115] Catherine J. Mummery, Karalyn Patterson, Cathy J. Price, John Ashburner, Richard S.J. Frackowiak, and John R. Hodges. A voxel-based morphometry study of semantic dementia: Relationship between temporal lobe atrophy and semantic memory. *Annals of Neurology*, 47(1):36–45, 2000.
- [116] Holly Robson, Roland Zahn, James L. Keidel, Richard J. Binney, Karen Sage, and Matthew A. Lambon Ralph. The anterior temporal lobes support residual comprehension in wernicke’s aphasia. *Brain*, 137(3):931–943, 2014.
- [117] W. Kyle Simmons and Alex Martin. The anterior temporal lobes and the functional architecture of semantic memory. *Journal of the International Neuropsychological Society*, 15(05):645–649, 2009.
- [118] Moritz Tenorth and Michael Beetz. KnowRob: A knowledge processing infrastructure for cognition-enabled robots. *The International Journal of Robotics Research*, 32(5):566–590, 2013.
- [119] Moritz Tenorth and Michael Beetz. Knowrob: Knowledge processing for autonomous personal robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2009)*, pages 4261–4266. IEEE, 2009.
- [120] Michele Palmia. Design and implementation of a system for mutual knowledge among cognition-enabled robots. Master’s thesis, 2013.
- [121] Miniija Tamosiunaite, Irene Markelic, Tomas Kulvicius, and F Worgotter. Generalizing objects by analyzing language. In *11th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2011)*, pages 557–563, 2011.
- [122] John C. Trueswell and Lila R. Gleitman. Learning to parse and its implications for language acquisition. 2009.
- [123] Paul Baxter, Joachim de Greeff, Rachel Wood, and Tony Belpaeme. Modelling concept prototype competencies using a developmental memory model. *Paladyn*, 3(4):200–208, 2012.
- [124] Anthony F. Morse, Joachim de Greeff, Tony Belpaeme, and Angelo Cangelosi. Epigenetic robotics architecture (ERA). *IEEE Transactions on Autonomous Mental Development*, 2(4):325–339, 2010.
- [125] Emre Uğur, Erhan Öztop, and Erol Şahin. Goal emulation and planning in perceptual space using learned affordances. *Robotics and Autonomous Systems*, 59(7):580–595, 2011.
- [126] Emre Uğur and Erol Şahin. Traversability: A case study for learning and perceiving affordances in robots. *Adaptive Behavior*, 18(3-4):258–284, 2010.

- [127] Luis Montesano, Manuel Lopes, Alexandre Bernardino, and José Santos-Victor. Learning object affordances: From sensory–motor coordination to imitation. *IEEE Transactions on Robotics*, 24(1):15–26, 2008.
- [128] Edward A. Feigenbaum and Herbert A. Simon. EPAM-like models of recognition and learning. *Cognitive Science*, 8(4):305–336, 1984.
- [129] John H. Gennari, Pat Langley, and Doug Fisher. Models of incremental concept formation. *Artificial Intelligence*, 40(1–3):11 – 61, 1989.
- [130] Michael Lebowitz. Experiments with incremental concept formation: UNIMEM. *Machine Learning*, 2(2):103–138, 1987.
- [131] Douglas H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2(2):139–172, 1987.
- [132] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.
- [133] Janet L. Kolodner. Maintaining organization in a dynamic long-term memory. *Cognitive Science*, 7(4):243–280, 1983.
- [134] John C. Turner, Penelope J. Oakes, S. Alexander Haslam, and Craig McGarty. Self and collective: Cognition and social context. *Personality and Social Psychology Bulletin*, 20:454–454, 1994.
- [135] Madan M. Pillutla and Xiao-Ping Chen. Social norms and cooperation in social dilemmas: The effects of context and feedback. *Organizational Behavior and Human Decision Processes*, 78(2):81–103, 1999.
- [136] Thomas C. Brown and Terry C. Daniel. Context effects in perceived environmental quality assessment: scene selection and landscape quality ratings. *Journal of Environmental Psychology*, 7(3):233–250, 1987.
- [137] Monica Biernat and Theresa K. Vescio. Categorization and stereotyping: Effects of group context on memory and social judgment. *Journal of Experimental Social Psychology*, 29(2):166–202, 1993.
- [138] Karen Dion, Ellen Berscheid, and Elaine Walster. What is beautiful is good. *Journal of personality and social psychology*, 24(3):285, 1972.
- [139] Hanan Shteingart, Tal Neiman, and Yonatan Loewenstein. The role of first impression in operant learning. *Journal of Experimental Psychology: General*, 142(2):476, 2013.
- [140] Scott Plous. *The psychology of judgment and decision making*. Mcgraw-Hill Book Company, 1993.

- [141] Lawrence W. Barsalou. Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1281–1289, 2009.
- [142] Lawrence W. Barsalou. Situated simulation in the human conceptual system. *Language and cognitive processes*, 18(5-6):513–562, 2003.
- [143] Wenchi Yeh and Lawrence W. Barsalou. The situated nature of concepts. *The American journal of psychology*, pages 349–384, 2006.
- [144] George A. Bekey. *Robotics: State of the art and future challenges*. Imperial College Press, 2008.
- [145] Benoit Larochelle, G-JM Kruijff, Nanja Smets, Tina Mioch, and Peter Groenewegen. Establishing human situation awareness using a multi-modal operator control unit in an urban search & rescue human-robot team. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2011)*, pages 229–234, 2011.
- [146] Andreas Gregoriades, Samuel Obadan, Harris Michail, Vicky Papadopoulou, and Despina Michael. A robotic system for home security enhancement. *Ag-ing Friendly Technology for Health and Independence*, pages 43–52, 2010.
- [147] Goushi Tsuruma, Hideaki Kanai, Toyohisa Nakada, and Susumu Kunifuji. Dangerous situation awareness support system for elderly people with dementia. In *Int. Conf. on Human Computer Interaction*, pages 62–67. ACTA Press, 2007.
- [148] Daniel Nyga, Ferenc Balint-Benczedi, and Michael Beetz. PR2 looking at things: Ensemble learning for unstructured information processing with markov logic networks. In *ICRA*, 2014.
- [149] Hyun Kim, Minkyong Kim, Kang-Woo Lee, Young-Ho Suh, Joonmyun Cho, and Young-jo Cho. Context-aware server framework for network-based service robots. In *IEEE SICE-ICASE*, pages 2084–2089, 2006.
- [150] Waskitho Wibisono, Arkady Zaslavsky, and Sea Ling. Improving situation awareness for intelligent on-board vehicle management system using context middleware. In *IEEE Intelligent Vehicles Symposium*, pages 1109–1114, 2009.
- [151] Eric Wang, Yong Se Kim, Hak Soo Kim, Jin Hyun Son, Sanghoon Lee, and Il Hong Suh. Ontology modeling and storage system for robot context understanding. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 922–929. Springer, 2005.
- [152] Fulvio Mastrogiovanni, Antonio Sgorbissa, and Renato Zaccaria. Context assessment strategies for ubiquitous robots. In *IEEE/RSJ International Conference on Robotics and Automation (ICRA)*, pages 2717–2722, 2009.

- [153] Amir Padovitz, Seng W. Loke, and Arkady Zaslavsky. Multiple-agent perspectives in reasoning about situations for context-aware pervasive computing systems. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 38(4):729–742, 2008.
- [154] Fábio Miranda, Tiago Cabral Ferreira, João Paulo Pimentão, and Pedro Sousa. Review on context classification in robotics. In *Rough Sets and Intelligent Systems Paradigms*, pages 269–276. Springer, 2014.
- [155] Greg B. Simpson. Context and the processing of ambiguous words. *Handbook of psycholinguistics*, 22:359–374, 1994.
- [156] Teun A. Van Dijk. Discourse and context. *A Sociocognitive Approach*. Cambridge, 2008.
- [157] Teun A. Van Dijk. Society and discourse. *How Social Contexts Influence Text and Talk*. Cambridge, 8, 2009.
- [158] Kenny R. Coventry, Angelo Cangelosi, Stephen N. Newstead, and Davi Bugmann. Talking about quantities in space: Vague quantifiers, context and similarity. *Language and Cognition*, 2(2):221–241, 2010.
- [159] Roger C. Schank and Robert P. Abelson. Scripts, plans, goals and understanding. *Hillsdale, NJ: Lawrence Erlbaum*, 1977.
- [160] Solène Kalénine, Françoise Bonthoux, and Anna M. Borghi. How action and context priming influence categorization: a developmental study. *British Journal of Developmental Psychology*, 27(3):717–730, 2009.
- [161] Eun E. Yoon, Glyn W. Humphreys, and M. Jane Riddoch. The paired-object affordance effect. *Journal of Experimental Psychology: Human Perception and Performance*, 36(4):812, 2010.
- [162] Anna M. Borghi, Andrea Flumini, Felice Cimatti, Davide Marocco, and Claudia Scorolli. Manipulating objects and telling words: a study on concrete and abstract words acquisition. *Frontiers in psychology*, 2, 2011.
- [163] Nikhilesh Natraj, Victoria Poole, J.C. Mizelle, Andrea Flumini, Anna M. Borghi, and Lewis A. Wheaton. Context and hand posture modulate the neural dynamics of tool–object perception. *Neuropsychologia*, 51:506–519, 2012.
- [164] Tom Bylander. Complexity results for planning. In *IJCAI*, volume 10, pages 274–279, 1991.
- [165] David Chapman. Planning for conjunctive goals. *Artificial intelligence*, 32(3):333–377, 1987.
- [166] James A. Hendler, Austin Tate, and Mark Drummond. Ai planning: Systems and techniques. *AI magazine*, 11(2):61, 1990.

- [167] Oliver Lindemann, Prisca Stenneken, Hein T. Van Schie, and Harold Bekkering. Semantic activation in action planning. *Journal of Experimental Psychology: Human Perception and Performance*, 32(3):633, 2006.
- [168] Michiel van Elk, Hein van Schie, and Harold Bekkering. Action semantics: a unifying conceptual framework for the selective use of multimodal and modality-specific object knowledge. *Physics of life reviews*, 2013.
- [169] Sarah H. Creem and Dennis R. Proffitt. Grasping objects by their handles: a necessary interaction between cognition and action. *Journal of Experimental Psychology: Human Perception and Performance*, 27(1):218, 2001.
- [170] Sarah L. Friedman and Ellin Kofsky Scholnick. *The developmental psychology of planning: Why, how, and when do we plan?* Psychology Press, 2014.
- [171] Irving Biederman, Robert J. Mezzanotte, and Jan C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology*, 14(2):143–177, 1982.
- [172] Antonio Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 53(2):169–191, 2003.
- [173] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie. Objects in context. In *Computer vision, 2007. ICCV 2007. IEEE 11th international conference on*, pages 1–8. IEEE, 2007.
- [174] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2929–2936. IEEE, 2009.
- [175] John McCarthy. Artificial intelligence, logic and formalizing common sense. *Philosophical logic and artificial intelligence*, 1989.
- [176] John McCarthy. From here to human-level ai. *Artificial Intelligence*, 171(18):1174 – 1182, 2007.
- [177] Irving Biederman, Jan C. Rabinowitz, Arnold L. Glass, and E. Webb Stacy. On the information extracted from a glance at a scene. *Journal of Experimental Psychology*, 103(3):597, 1974.
- [178] Mary C. Potter. Short-term conceptual memory for pictures. *Journal of experimental psychology: human learning and memory*, 2(5):509, 1976.
- [179] Helene Intraub. Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology: Human Perception and Performance*, 7(3):604, 1981.
- [180] Jodi L. Davenport and Mary C. Potter. Scene consistency in object and background perception. *Psychological Science*, 15(8):559–564, 2004.

- [181] Marvin M. Chun and Yuhong Jiang. Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive psychology*, 36(1):28–71, 1998.
- [182] Marvin M. Chun and Elizabeth A. Phelps. Memory deficits for implicit contextual information in amnesic subjects with hippocampal damage. *Nature neuroscience*, 2(9):844–847, 1999.
- [183] Mark Good, Livia De Hoz, and Richard G.M. Morris. Contingent versus incidental context processing during conditioning: dissociation after excitotoxic hippocampal plus dentate gyrus lesions. *Hippocampus*, 8(2):147–159, 1998.
- [184] Fei Fei Li, Rufin VanRullen, Christof Koch, and Pietro Perona. Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, 99(14):9596–9601, 2002.
- [185] Marta Kutas and Steven A. Hillyard. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205, 1980.
- [186] Giorgio Ganis and Marta Kutas. An electrophysiological study of scene effects on object identification. *Cognitive Brain Research*, 16(2):123–144, 2003.
- [187] E. Halgren, P. Baudena, G. Heit, M. Clarke, K. Marinkovic, and P. Chauvel. Spatio-temporal stages in face and word processing. 2. depth-recorded potentials in the human frontal and rolandic cortices. *Journal of Physiology-Paris*, 88(1):51–80, 1994.
- [188] Anders M. Dale, Arthur K. Liu, Bruce R. Fischl, Randy L. Buckner, John W. Belliveau, Jeffrey D. Lewine, and Eric Halgren. Dynamic statistical parametric mapping: combining fmri and meg for high-resolution imaging of cortical activity. *Neuron*, 26(1):55–67, 2000.
- [189] Gina Kuperberg, Phillip J. Holcomb, Tatiana Sitnikova, Douglas Greve, Anders M. Dale, and David Caplan. Distinct patterns of neural modulation during the processing of conceptual and syntactic anomalies. *Cognitive Neuroscience, Journal of*, 15(2):272–293, 2003.
- [190] Neil Burgess, Eleanor A. Maguire, Hugo J. Spiers, and John O’Keefe. A temporoparietal and prefrontal network for retrieving the spatial context of lifelike events. *Neuroimage*, 14(2):439–453, 2001.
- [191] Jon S. Simons and Hugo J. Spiers. Prefrontal and medial temporal lobe interactions in long-term memory. *Nature Reviews Neuroscience*, 4(8):637–648, 2003.
- [192] Eleanor Maguire. The retrosplenial contribution to human navigation: a review of lesion and neuroimaging findings. *Scandinavian journal of psychology*, 42(3):225–238, 2001.

- [193] Brenton G. Cooper and Sheri J.Y. Mizumori. Temporary inactivation of the retrosplenial cortex causes a transient reorganization of spatial coding in the hippocampus. *The Journal of Neuroscience*, 21(11):3986–4001, 2001.
- [194] Seralynne D. Vann and John P. Aggleton. Extensive cytotoxic lesions of the rat retrosplenial cortex reveal consistent deficits on tasks that tax allocentric spatial memory. *Behavioral Neuroscience*, 116(1):85, 2002.
- [195] Hande Çelikkanat, Güner Orhan, Nicolas Pugeault, Frank Guerin, Erol Şahin, and Sinan Kalkan. Learning context on a humanoid robot using Incremental Latent Dirichlet Allocation. *IEEE Transactions on Autonomous Mental Development (accepted)*, 2015.
- [196] Radu Bogdan Rusu and Steve Cousins. 3D is here: Point Cloud Library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 9-13 2011.
- [197] Güner Orhan, Sertaç Olgunsoylu, Erol Şahin, and Sinan Kalkan. Co-learning nouns and adjectives. In *2013 IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pages 1–6, Aug 2013.
- [198] A. Kai Qin and Ponnuthurai N. Suganthan. Robust growing neural gas algorithm with application in cluster analysis. *Neural Networks*, 17(8-9):1135–1148, 2004.
- [199] Todd Veldhuizen. Ubigraph: Free dynamic graph visualization software, 2007.
- [200] Ross Kindermann, James Laurie Snell, et al. *Markov random fields and their applications*, volume 1. American Mathematical Society Providence, RI, 1980.
- [201] Almero Gouws. *A Python implementation of graphical models*. PhD thesis, Stellenbosch: University of Stellenbosch, 2010.
- [202] Akira Murata, Luciano Fadiga, Leonardo Fogassi, Vittorio Gallese, Vassilis Raos, and Giacomo Rizzolatti. Object representation in the ventral premotor cortex (area f5) of the monkey. *Journal of Neurophysiology*, 78(4):2226–2230, 1997.
- [203] Giacomo Rizzolatti and Luciano Fadiga. Grasping objects and grasping action meanings: the dual role of monkey rostroventral premotor cortex (area f5). *Sensory Guidance of Movement*, 218:81–103, 1998.
- [204] Luciano Fadiga, Leonardo Fogassi, Vittorio Gallese, and Giacomo Rizzolatti. Visuomotor neurons: Ambiguity of the discharge or ‘motor’ perception? *International journal of psychophysiology*, 35(2):165–177, 2000.



- [205] Igor Kononenko. Estimating attributes: analysis and extensions of relief. In *Machine Learning: ECML-94*, pages 171–182. Springer, 1994.
- [206] Ahmet Can Bulut. A multinomial prototype-based learning algorithm. Master’s thesis, Middle East Technical University, 2014.
- [207] Hande Çelikkanat, Güner Orhan, Nicolas Pugeault, Frank Guerin, Erol Şahin, and Sinan Kalkan. Learning and using context on a humanoid robot using latent dirichlet allocation. In *IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pages 201–207, 2014.
- [208] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.
- [209] Yee Whye Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Advances in neural information processing systems*, pages 1353–1360, 2006.
- [210] Kevin R. Canini, Lei Shi, and Thomas L. Griffiths. Online inference of topics with latent dirichlet allocation. In *International Conference on Artificial Intelligence and Statistics*, pages 65–72, 2009.
- [211] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57. ACM, 1999.
- [212] Dongshan Xing and Mark Girolami. Employing Latent Dirichlet Allocation for fraud detection in telecommunications. *Pattern Recognition Letters*, 28(13):1727–1734, 2007.
- [213] Mark Girolami and Ata Kabán. Sequential activity profiling: Latent Dirichlet Allocation of Markov chains. *Data Mining and Knowledge Discovery*, 10(3):175–196, 2005.
- [214] Bing Liu, Lin Liu, Anna Tsykin, Gregory J Goodall, Jeffrey E Green, Min Zhu, Chang Hee Kim, and Jiuyong Li. Identifying functional mirna–mrna regulatory modules with correspondence latent dirichlet allocation. *Bioinformatics*, 26(24):3105–3111, 2010.
- [215] Tomoaki Nakamura, Takayuki Nagai, and Naoto Iwahashi. Grounding of word meanings in multimodal concepts using lda. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3943–3948, 2009.
- [216] Tomoaki Nakamura, Takaya Araki, Takayuki Nagai, and Naoto Iwahashi. Grounding of word meanings in Latent Dirichlet Allocation-based multimodal concepts. *Advanced Robotics*, 25:2189–2206, 2011.

- [217] Takaya Araki, Tomoaki Nakamura, Takayuki Nagai, Shogo Nagasaka, Tadahiro Taniguchi, and Naoto Iwahashi. Online learning of concepts and words using multimodal LDA and Hierarchical Pitman-Yor Language Model. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1623–1630, 2012.
- [218] Takashi Bando, Kazuhito Takenaka, Shogo Nagasaka, and Tadahiro Taniguchi. Automatic drive annotation via multimodal latent topic model. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2744–2749, Nov 2013.
- [219] Charles E. Antoniak. Mixtures of Dirichlet Processes with applications to Bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174, 1974.
- [220] Thomas Griffiths and Zoubin Ghahramani. Infinite latent feature models and the indian buffet process. In *Advances in Neural Information Processing Systems (NIPS)*, 2005.
- [221] Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.
- [222] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476), 2006.
- [223] Chong Wang, John W Paisley, and David M Blei. Online variational inference for the Hierarchical Dirichlet Process. In *International Conference on Artificial Intelligence and Statistics*, pages 752–760, 2011.
- [224] Nancy L. Stein and Tom Trabasso. What’s in a story: An approach to comprehension and instruction. R. Glaser (Ed.), *Advances in instructional psychology*, pages 213–267, 1981.
- [225] Nelson Cowan. *Working memory capacity*. Psychology Press, 2004.
- [226] John D.E. Gabrieli, Russell A. Poldrack, and John E. Desmond. The role of left prefrontal cortex in language and memory. *Proceedings of the national Academy of Sciences*, 95(3):906–913, 1998.
- [227] Gregg C. Oden. Concept, knowledge, and thought. *Annual Review of Psychology*, 38(1):203–227, 1987.
- [228] Jaegwon Kim. Concepts of supervenience. *Philosophy and Phenomenological Research*, pages 153–176, 1984.

- [229] Alexander Klippel and Daniel R. Montello. Linguistic and nonlinguistic turn direction concepts. In *Spatial information theory*, pages 354–372. Springer, 2007.
- [230] Sabine Timpf, Gary S. Volta, David W. Pollock, and Max J. Egenhofer. A conceptual model of wayfinding using multiple levels of abstraction. In *Theories and methods of spatio-temporal reasoning in geographic space*, pages 348–367. Springer, 1992.
- [231] James A. Hampton. Conceptual combination. *Knowledge, concepts, and categories*, pages 133–159, 1997.
- [232] Albert H. Hastorf and Hadley Cantril. They saw a game; a case study. *The Journal of Abnormal and Social Psychology*, 49(1):129, 1954.
- [233] Sebastian Thrun and Tom M. Mitchell. *Lifelong robot learning*. Springer, 1995.
- [234] Rodney A. Brooks, Cynthia Breazeal, Robert Irie, Charles C. Kemp, Matthew Marjanovic, Brian Scassellati, and Matthew M. Williamson. Alternative essences of intelligence. *AAAI/IAAI*, 1998:961–968, 1998.
- [235] Minoru Asada, Karl F. MacDorman, Hiroshi Ishiguro, and Yasuo Kuniyoshi. Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous Systems*, 37(2):185–193, 2001.
- [236] Luc Berthouze and Tom Ziemke. Epigenetic robotics — modelling cognitive development in robotic systems. *Connection Science*, 15(4):147–150, 2003.
- [237] Jordan Zlatev and Christian Balkenius. Introduction: Why “Epigenetic Robotics”? In *Proceedings of the First International Workshop on Epigenetic Robotics: Modelling Cognitive Development in Robotic Systems*, pages 1–4. Citeseer, 2001.
- [238] Max Lungarella, Giorgio Metta, Rolf Pfeifer, and Giulio Sandini. Developmental robotics: a survey. *Connection Science*, 15(4):151–190, 2003.
- [239] Sebastian Thrun. Is learning the n-th thing any easier than learning the first? *Advances in Neural Information Processing Systems*, pages 640–646, 1996.
- [240] Richard S. Sutton. Verification, the key to AI. <https://webdocs.cs.ualberta.ca/~sutton/IncIdeas/KeytoAI.html>, 2001. [Online; accessed 09.10.2015].
- [241] Z. Chen, R. S. Siegler, and M. W. Daehler. Across the great divide: Bridging the gap between understanding of toddlers’ and older children’s thinking. *Monographs of the Society for Research in Child Development*, 65(2), 2000.

- [242] Gregory L. Murphy and Mary E. Lassaline. Hierarchical structure in concepts and the basic level of categorization. *Knowledge, concepts, and categories*, pages 93–131, 1997.

# CURRICULUM VITAE

## PERSONAL INFORMATION

**Surname, Name:** Çelikkanat, Hande

**Date and Place of Birth:** 08.10.1984, Ankara

## EDUCATION

<b>Degree</b>	<b>Institution</b>	<b>Year</b>
Ph.D.	Computer Engineering, Middle East Technical University	2015
M.Sc.	Computer Engineering, Middle East Technical University	2008
B.Sc.	Computer Engineering, Middle East Technical University	2006
High School	Ankara Atatürk Anatolian High School	2002

## PROFESSIONAL EXPERIENCE

**Jun. 2006 to present** - Researcher, KOVAN Research Laboratory, Department of Computer Engineering, Middle East Technical University, Turkey

**Mar. 2014 - Sep. 2015** - Researcher, The Scientific and Technological Research Council of Turkey (TÜBİTAK) Project no. 111E287, Development of Hierarchical Concepts in Humanoid Robots, Turkey

**Sep. 2011 to Sep. 2012** - Visiting Researcher, Centre for Robotics and Neural Systems, Plymouth University, UK

**Jan. 2011 to Mar. 2012** - Intern Researcher, Advanced Telecommunications Research Institute International (ATR), Japan

**Sep. 2006 to Sep. 2011** - Teaching Assistant, Department of Computer Engineering, Middle East Technical University, Turkey

## PUBLICATIONS

### Journal Publications

- **Hande Çelikkanat**, Güner Orhan, Nicolas Pugeault, Frank Guerin, Erol Şahin, and Sinan Kalkan, Learning Context on a Humanoid Robot using Incremental Latent Dirichlet Allocation, *accepted for publication by IEEE Transactions on Autonomous Mental Development*, 2015.
- **Hande Çelikkanat**, Güner Orhan, and Sinan Kalkan, A Probabilistic Web of Concepts on a Humanoid Robot, *IEEE Transactions on Autonomous Mental Development*, vol: 7, no: 2, pp. 92-106, 2015.
- Emre Uğur, Yukie Nagai, **Hande Çelikkanat**, and Erhan Öztop, Parental scaffolding as a bootstrapping mechanism for learning grasp affordances and imitation skills, *Robotica*, vol: 33, no: 5, pp. 1163-1180, 2015.
- **Hande Çelikkanat** and Erol Şahin, Steering Self-organized Robot Flocks through Externally Guided Individuals, *Neural Computing and Applications*, vol: 19, no: 6, pp. 849-865, 2010.
- Ali Emre Turgut, **Hande Çelikkanat**, Fatih Gökçe, and Erol Şahin, Self-Organized Flocking in Mobile Robot Swarms, *Swarm Intelligence*, vol: 2, no:2-4, pp. 97-120, 2008.

### Conference Publications

- **Hande Çelikkanat**, Erol Şahin, and Sinan Kalkan, Integrating Spatial Concepts into a Probabilistic Concept Web, *International Conference on Advanced Robotics*, 2015.
- **Hande Çelikkanat**, Güner Orhan, Nicolas Pugeault, Frank Guerin, Erol Şahin, and Sinan Kalkan, Learning and Using Context on a Humanoid Robot Us-

ing Latent Dirichlet Allocation, *IEEE Joint Conference on Development and Learning and on Epigenetic Robotics*, pp.201-207, 2014.

- **Hande Çelikkanat**, Erol Şahin, and Sinan Kalkan, Recurrent Slow Feature Analysis for Developing Object Permanence in Robots, *Proceedings of the Workshop on Neuroscience and Robotics at IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1-6, 2013.
- Emre Uğur, **Hande Çelikkanat**, Erol Şahin, Yukie Nagai, and Erhan Öztop, Learning to grasp with parental scaffolding, *IEEE-RAS International Conference on Humanoid Robots*, pp. 480-486, 2011.
- **Hande Çelikkanat**, Ali Emre Turgut, and Erol Şahin, Guiding a Robot Flock via Informed Robots, *Proceedings of the 9th International Symposium on Distributed Autonomous Robotic Systems*, pp. 215-225, 2009.
- Ali Emre Turgut, Christian Huepe, **Hande Çelikkanat**, Fatih Gökçe, and Erol Şahin, Modelling Phase Transition in Self-Organized Mobile Robot Flocks, *Proceedings of the 6th International Conference on Ant Colony Optimization and Swarm Intelligence*, vol: 5217, pp. 108-119, 2008.
- Ali Emre Turgut, **Hande Çelikkanat**, Fatih Gökçe, and Erol Şahin, Self-Organized Flocking with a Mobile Robot Swarm, *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems*, pp. 39-46, 2008.

### Conference Publications in Turkish

- **Hande Çelikkanat**, Güner Orhan, Erol Şahin and Sinan Kalkan. İnsansı robotlar için olasılıksal bir kavram ağı (*ing.* A probabilistic web of concepts for humanoid robots), *Proceedings of the 2nd Türkiye Robotbilim Konferansı*, 2015.
- **Hande Çelikkanat**, Güner Orhan, Nicolas Pugeault, Frank Guerin, Erol Şahin and Sinan Kalkan. İnsansı robotlarda bağlamın öğrenilmesi (*ing.* Learning of context in humanoid robots), *Proceedings of the 1st Türkiye Otonom Robotlar Konferansı*, 2014.

- **Hande Çelikkanat** and Sinan Kalkan, Yavaşlık İlkesini Kullanarak Öznitelik Seçimi: Alakalı Öznitelik Analizi (*ing.* Using Slowness Principle for Feature Selection: Relevant Feature Analysis), *Proceedings of IEEE 22. Sinyal İşleme ve İletişim Uygulamaları Kurultayı*, 2014.
- Ali Emre Turgut, Fatih Gökçe, **Hande Çelikkanat**, Levent Bayındır, and Erol Şahin, Kobot: Sürü Robot Çalışmaları için Tasarlanmış Gezgin Robot Platformu (*ing.* Kobot: A Mobile Robot Platform Developed for Swarm Robotics Research), *Proceedings of Türkiye Otomatik Kontrol Ulusal Toplantısı*, pp. 259-264, 2007.
- **Hande Çelikkanat**, Ali Emre Turgut, Erol Şahin, and Buğra Koku, Oğul Robot Sistemleri için Basit Bir Görüntüleme Sistemi Tasarımı (*ing.* A Simple Design for the Visual System of a Robot Swarm), *Proceedings of Türkiye Otomatik Kontrol Ulusal Toplantısı*, pp. 259-264, 2006.