

Spatial Experience in Humans and Machines

by

Çağrı Hakan Zaman

B.Arch., Istanbul Technical University (2009)

S.M., Istanbul Technical University (2011) S.M., Massachusetts

Institute of Technology (2014)

Submitted to the Department of Architecture

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Architecture: Design and Computation

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2020

© Massachusetts Institute of Technology 2020. All rights reserved.

Author

Department of Architecture

January 10, 2020

Certified by

Terry Knight

Professor of Design and Computation

Thesis Supervisor

Certified by

Randall Davis

Professor of Computer Science and Engineering

Thesis Supervisor

Certified by

Patrick H. Winston

Professor of Electrical Engineering and Computer Science

Thesis Supervisor

Accepted by

Les Norford

Chairman, Department Committee on Graduate Theses

Dissertation Committee

Terry Knight

Professor of Design and Computation
Massachusetts Institute of Technology
Thesis Supervisor

Randall Davis

Professor of Computer Science and Engineering
Massachusetts Institute of Technology
Thesis Supervisor

Patrick H. Winston

Professor of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Thesis Supervisor

George Stiny

Professor of Design and Computation
Massachusetts Institute of Technology
Thesis Reader

Spatial Experience in Humans and Machines

by

Çağrı Hakan Zaman

Submitted to the Department of Architecture
on January 10, 2020, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Architecture: Design and Computation

Abstract

Spatial experience is the process by which we locate ourselves within our environment, and understand and interact with it. Understanding spatial experience has been a major endeavor within the social sciences, the arts, and architecture throughout history, giving rise to recent theories of embodied and enacted cognition. Understanding spatial experience has also been a pursuit of computer science. However, despite substantial advances in artificial intelligence and computer vision, there has yet to be a computational model of human spatial experience. What are the computations involved in human spatial experience? Can we develop machines that can describe and represent spatial experience?

In this dissertation, I take a step towards developing a computational account of human spatial experience and outline the steps for developing machine spatial experience. Building on the core idea that we humans construct stories to understand the environment and communicate with each other, I argue that spatial experience is a type of story we tell ourselves, driven by what we perceive and how we act within the environment. Through two initial case studies, I investigate the relationships between stories and spatial experience and introduce the anchoring framework—a computational model of constructing stories using emergent spatial, temporal, and visual relationships in perception. I evaluate this framework by performing a visual exploration study and analyzing how people verbally describe environments. Finally, I implement the anchoring framework for creating spatial experiences by machines. I introduce three examples, which demonstrate that machines can solve visuo-spatial problems by constructing stories from visual perception using the anchoring framework. This dissertation contributes to the fields of design, media studies, and artificial intelligence by advancing our understanding of human spatial experience from a story perspective; providing a set of tools and methods for creating and analyzing spatial experiences; and introducing systems that can understand the physical environment and solve spatial problems by constructing stories.

Thesis Supervisor: Terry Knight
Title: Professor of Design and Computation

Thesis Supervisor: Randall Davis
Title: Professor of Computer Science and Engineering

Thesis Supervisor: Patrick H. Winston
Title: Professor of Electrical Engineering and Computer Science

I dedicate this work to the memory of my advisor Patrick Henry Winston, who recently passed away. I want to thank Prof. Randall Davis, who signed this thesis on behalf of him.

Acknowledgments

Foremost, I want to express my deepest gratitude to my late advisor Patrick Henry Winston. He was the reason I took a path in artificial intelligence and decided to dedicate my career to understanding the human mind, which, in his words, is “the greatest innovation of all times.” Patrick was a great communicator, a brilliant teacher, and a pioneer in the field of artificial intelligence. His life work on story understanding inspired my research on human spatial experience, to which his contributions were invaluable both in terms of vision, style, and technical rigor. I feel extremely lucky to have known Patrick, whose teachings will undoubtedly continue to guide me both professionally and personally.

My advisor Terry Knight has been by far the most supportive figure throughout my time at MIT. Without her guidance, support, and dedication, this research would not have been successful. Besides her mentorship for this thesis, Prof. Knight led the way for establishing the Virtual Experience Design Lab, which enabled me to take on the research challenges described in this work and expanded my research interests even beyond. I am grateful for her championship.

I am grateful to Randall Davis, who has taken over as my advisor after we lost Patrick. While ensuring that I complete my research in a timely manner, he substantially improved my work in every possible sense. He pointed out the right direction and provided invaluable feedback when I needed the most.

I want to thank my thesis reader, George Stiny, for insightful discussions over the years, during thesis meetings, over the hallways, and in his proseminar class, which has been an absolute joy to attend.

This thesis was made possible by a group of brilliant collaborators and study participants. Deniz Tortum and Nil Tuzcu are my co-creators for the project “September 1955”, and Nil Tuzcu is also my co-instructor for the class “Computational Ethnography and Spatial Narratives” class in the BAC, both of which are described in chapter 3. The visual exploration study described in chapter 4 was done in collaboration with Ainsley Sutherland, Danielle Olson, Ben Zuan, and

Leilani Gilpin, Ege Ozgirin, Bilge Zeren Aksu, and Zhutian Yang contributed to the artificial intelligence projects described in chapter 5. I want to thank all of my collaborators, students, and my study participants whose names I cannot disclose.

I am grateful to my friends and colleagues at CSAIL’s Genesis Group. I would especially like to thank Adam Kraft, who patiently edited my drafts and provided constant feedback. Most of my ideas drew from our weekly group meetings with my fellow travelers, including Dylan Holmes, Leilani Gilpin, Zhutian Yang, Yida Xin, Jennifer Madiedo, Jamie Macbeth, Henry Lieberman, and Gerald Sussman. Thank you all. To Mustafa Anil Kocak, I am thankful for his friendship and providing constructive feedback on my work.

I am indebted to my friends and colleagues at DCG, whose friendship will always stay with me. Among those, Katia Zolotovskiy has provided me with my daily dose of optimism; Moa Carlsson handed down her “How to Write your Dissertation in 15 minutes” book and Dina El-Zanfaly her clay printer. Athina Papadopoulou never fell short of inspiring and challenging me since my first day at MIT. I am thankful to Alexandros Charidis for his feedback and comments on my work. Also thanks Asli Arpak, Onur Yuce Gun, Theodora Vardouli, Woong Ki Sung, Rachelle Villalon, Eytan Michael Mann, Diego Pinochet, Nikolaos Vlavianos, Carlos Sandoval Olascoaga, and Paloma Gonzalez Rojas.

I have been privileged to have the opportunity to establish Virtual Experience Design Lab with Prof. Terry Knight, thanks to the leadership of the MIT School of Architecture, Dean Hashim Sarkis, Assistant Dean of Finance Ken Goldsmith, and Former Director of Communications Tom Gearty. I am greatly grateful for their generosity of support and guidance. In addition, MIT DesignX provided invaluable resources for setting up my virtual reality experiments. Special thanks to Denis Frenchman, Gilad Rosenzweig, and Svafa Gronfeldt.

To Marilyn Levine and Patricia Brenecke, thanks for their invaluable writing feedback and helping me improve my communication skills.

I am deeply grateful to my family, my parents Esin and Mehmet Zaman, my sisters Aysegul and Elif, my brother in law Firat, my nieces Lila and Gul, and my nephew

Can. I have always felt their love and support despite the long distances between us. Also, thanks to my extended family Ayfer Tuzcu, Gul, Ali, and Olivia Coskun, who provided invaluable emotional support whenever I needed.

Above all, I want to thank my wife, Nil Tuzcu. She has been my friend, colleague, and a source of courage and inspiration from the beginning. Without her patience, support, and love, I would have never taken on this challenge.

This dissertation was generously supported by the MIT Design Lab, DARPA, and NSF. Special thanks to Federico Casalegno for his support and mentorship.

Contents

1	Introduction	25
1.1	Vision	25
1.2	Motivation and Background	28
1.2.1	Human Spatial Experience is an Inner Story	30
1.2.2	Seeing and Thinking Require Space	32
1.2.3	Seeing and Doing are Composed into Stories	33
1.2.4	Anchoring is the Key Mechanism for Composing Stories in Space	35
1.3	Immersive Technologies Enable the Observation of Spatial Experiences	37
1.4	Overview	38
2	Human Spatial Experience	43
2.1	Perception of Space	44
2.1.1	Philosophical Underpinnings: Nativist and Empricist Agendas	44
2.1.2	Psychophysics and the Study of Human Eye	46
2.1.3	From Sensationalist to Ecological Psychology: “Ask not what’s inside your head, but what your head’s inside of”	49
2.1.4	Contemporary Cognitive and Neuroscience	51
2.2	An Argument for the Spatial Mind	53
2.2.1	Architects are Skilled Practitioners of Space	54
2.2.2	Dwelling versus Building	54
2.2.3	Spatial Ability and Spatial Knowledge	60
2.3	Computational Approaches to Understanding Visual Perception and Spatial Experience	62

2.3.1	Geometrical Models	63
2.3.2	Bottom-up Models and Deep Neural Networks	65
2.3.3	Top Down Models	67
2.3.4	Formal Approaches in Design	68
2.4	Discussion	71
3	Creating and Representing Spatial Experiences	73
3.1	September 1955: Immersing into the history	74
3.1.1	Project Setup	75
3.1.2	Streetscape and The Studio	77
3.1.3	Story Elements	78
3.1.4	Narratives	81
3.1.5	Qualitative Analysis of the Viewer’s Experience	82
3.1.6	Discussion	84
3.2	Computational Ethnography and Spatial Narratives of Urban Space .	85
3.2.1	Exercise 1: Draw Your Day	86
3.2.2	Exercise 2: Redefining the Urban Elements	89
3.2.3	Exercise 3: Physical Models of Observation	90
3.2.4	Discussion	93
3.3	Contributions	95
4	Spatial Experience as Inner Story: The Anchoring Framework	97
4.1	Inner Stories and the Anchoring Problem	98
4.1.1	Anchors Connect Symbols in Language to Visual and Active Information	100
4.1.2	The Anchoring Framework	104
4.1.3	Summary	109
4.2	Visual Exploration Study: The See, Act and Tell Methodology	109
4.2.1	Methodology	111
4.2.2	Study Results	117
4.3	Anchors and Domains	126

4.3.1	Spatial Anchors	130
4.3.2	Visual Anchors	141
4.3.3	Temporal Anchors	143
4.4	Summary	143
4.5	Contributions	144
5	Implementing the Anchoring Framework for Developing Spatial Experience in Machines	147
5.1	Learning Domain-Specific Representations	148
5.1.1	Background	149
5.1.2	Methodology	151
5.1.3	Spatial Cell Representations	152
5.1.4	Model	154
5.1.5	Training	155
5.1.6	Results	156
5.1.7	Using Self Organizing Maps as Place Descriptors	159
5.1.8	Summary	163
5.2	Connecting Visual and Symbolic Systems	163
5.2.1	Genesis Story Understanding System	164
5.2.2	Simulated robot replaces a cell phone battery	168
5.2.3	Genesis Describes the Environment	181
5.2.4	Contributions	188
6	Conclusions	191
6.1	Contributions	191
6.2	Open Questions and Next Steps	194
A	Appendix: See, Act, and Tell Methodology	195
A.1	An overview of methodologies for observing spatial experiences	195
A.1.1	Event Recording	196
A.1.2	Think Aloud Protocol	197

A.2	The use of Virtual Reality as an experimental tool	197
A.2.1	Limitations	199
A.3	Consent Form Sample	201
A.4	Verbal Description Samples	202

List of Figures

1-1	Illustration of a robot that replaces a phone battery	28
1-2	DNN's underperform in symbolic reasoning tasks	29
1-3	Symbolic domains in spatial language	36
1-4	Representing spatial experiences via drawings	39
1-5	Chapter 4: The See, Act and Tell methodology	40
2-1	Neural representations of space in the animal brain	52
2-2	Gestalt Fields in design process.	59
2-3	Learning a maze	61
2-4	Geometrical models of the perception of space	64
2-5	Society of Nearness by Minsky	66
2-6	Dynamism in visual perception is often used by artists to create illusions.	69
3-1	September 1955 VR installation at the Istanbul Independent Film Festival	76
3-2	September 1955 virtual spaces layout	77
3-3	Objects of interest	79
3-4	Animated characters in "September 1955	80
3-5	A viewer watching September 1955 in Keller Gallery	82
3-6	Exercise 1: Draw your experiences today up to the time you arrived in the classroom.	86
3-7	Representing a day on a map	88
3-8	Redefining urban elements	91
3-9	Representing the public activity with physical models	92

3-10	Representing noise with physical models	94
4-1	Learning depth representation from RGB	100
4-2	Graded material properties	101
4-3	Focal colors	103
4-4	An illustration of visual domains and symbols	105
4-5	Generating domain-specific representations for distances	106
4-6	Generating new domain functions	108
4-7	A still image from one of the experiment spaces	112
4-8	Sample transcripts with corresponding RGB images that are obtained from the study.	116
4-9	An overlay of the locations in Environment 3	118
4-10	Plan view of attention mapping.	119
4-11	Top 5 most attended parts of the three environments.	120
4-12	Movements and observations in Environment 2	121
4-13	Graph of camera parameters	122
4-14	Establishing location by looking around	123
4-15	Number of navigation actions per subject in each environment.	125
4-16	Most frequent words in the dataset.	128
4-17	Comparison of spatial word frequency among datasets	128
4-18	Comparison of visual word frequency among datasets	129
4-19	Spatial anchor domains	131
4-20	An illustration of domain-specific representation for demonstrative anchors.	133
4-21	Domain-specific representations for demonstrative anchors from the study.	134
4-22	Names of rooms mapped on the locations in the Environment where participants verbalized them	135
4-23	An illustration of domain-specific representations for proximity anchors.	136
4-24	Domain-specific representation for vertical relations	138

4-25	Domain-specific representation for attachment relations	138
4-26	Grouping anchors.	140
4-27	An illustration of material domains	142
4-28	Temporal anchors.	142
5-1	Neural representations of space in the animal brain	149
5-2	The training and test environments for the navigation task.	151
5-3	Biologically inspired representations of environmental location	152
5-4	Convolutional Neural Network (CNN) for predicting environmental boundaries	154
5-5	Loss values	155
5-6	Boundary predictions for 8-Cell and 12-Cell models	157
5-7	The training loss for the predicted activation values in two different training environments.	158
5-8	Predicted activations for 3 different grid features in a square environment	158
5-9	Self Organizing Map (SOM)	159
5-10	Spider graph visualization	160
5-11	Randomly initialized SOM with 20x20 cells	161
5-12	Final SOM trained with domain-specific representation of distances and orientations	161
5-13	SOM distinguishes distinct places in an environment.	162
5-14	Building blocks of Genesis	164
5-15	An elaboration graph	167
5-16	Robotic arm simulator	169
5-17	The Genesis Problem Solver communicates with a robotic arm simulator	172
5-18	Anchor Graph	175
5-19	Visual processes that generate anchors.	177
5-20	Different stages of battery replacement task	180
5-21	The test environments for the scene description task	182

5-22	Elaboration Graph generated from observations in the residential environment	184
5-23	Elaboration Graph generated from observations in the office environment	185
5-24	Answering questions using ConceptNet	187
A-1	Environment 1 Movement Overlay	213
A-2	Environment 2 Movement Overlay	214

List of Tables

4.1	Examples of anchors and domain-specific representations.	110
4.2	A sample of camera parameters.	114
4.3	Timestamped words generated by Gentle Aligner.	115
4.4	Closed-class words in English.	130

Glossary

Domain: A specific type of environmental information, such as distance, and a group of symbols that refer to this information, such as *close* and *far*.

Domain-specific representation: An image-like representation that contains a specific type of visual information. For example, a depth map is a domain-specific representation of the distance domain.

Domain Function: A function that generates a domain-specific representation from an input image.

Anchor: A function that generates symbolic relationships from visual information. An anchor operates on a domain-specific representation.

Inner Language: A cognitive mechanism that allows constructing complex symbolic descriptions of classes, properties, relations, actions, and events.

Inner Story: A collection of complex, highly nested symbolic descriptions of properties, relations, actions, and events, usefully connected with constraints such as causal, enablement, and time constraints.

Chapter 1

Introduction

1.1 Vision

How do we understand the environment around us? As we move through space, our physical perspective changes, creating a relationship between the changes we perceive and the movements we make. In so doing, we acquire information about where we are, what objects and materials are around us, and what can we do with them. Typically, this information is readily available to our awareness (for example, we quickly register the overall size of a room). However, some information becomes available to us only after dedicating our attention to a particular thing or property (for example, the material qualities of a table) while other pieces of information emerge in relation to a context (for example, a group of chairs around a table).

Understanding our environment is an active process that takes time and requires active attention. However, despite the substantial advancements in computer vision and discoveries in cognitive neuroscience, computational mechanisms that can model this type of spatial understanding have not yet been developed. I believe how we come to understand our environment is a crucial aspect of human intelligence, making computational models of this skill a cornerstone in developing artificial intelligence.

My objective in this dissertation is to take a step towards developing a computational account of human spatial experience —the process by which humans understand their immediate environment. I claim that human spatial experience is a

type of inner story that we tell ourselves, driven by what we perceive and how we act within our environment. My goal is to determine how our perceptions and actions within our environment develop into these inner stories, which then guide us to understand and interact with our environments. As a step towards this goal, I introduce a computational framework to integrate visual perception and symbolic descriptions, which remains one of the main challenges in artificial intelligence research.

My framing of human spatial experience as a form of inner story builds on Winston’s *Strong Story Hypothesis* (Winston, 2011). The Strong Story Hypothesis states that the mechanisms that enable humans to understand, tell, and recombine stories separates human intelligence from that of other primates. Winston argues that humans have a unique symbolic ability, *an inner language*¹, with which they can construct complex symbolic descriptions of classes, properties, relations, actions, and events. According to him, our inner language also enables us to connect complex symbolic descriptions with various constraints such as causal, means-ends, enablement, and time constraints (Winston and Holmes, 2019):

An inner story: A collection of complex, highly nested symbolic descriptions of properties, relations, actions, and events, usefully connected with constraints such as causal, enablement, and time constraints.

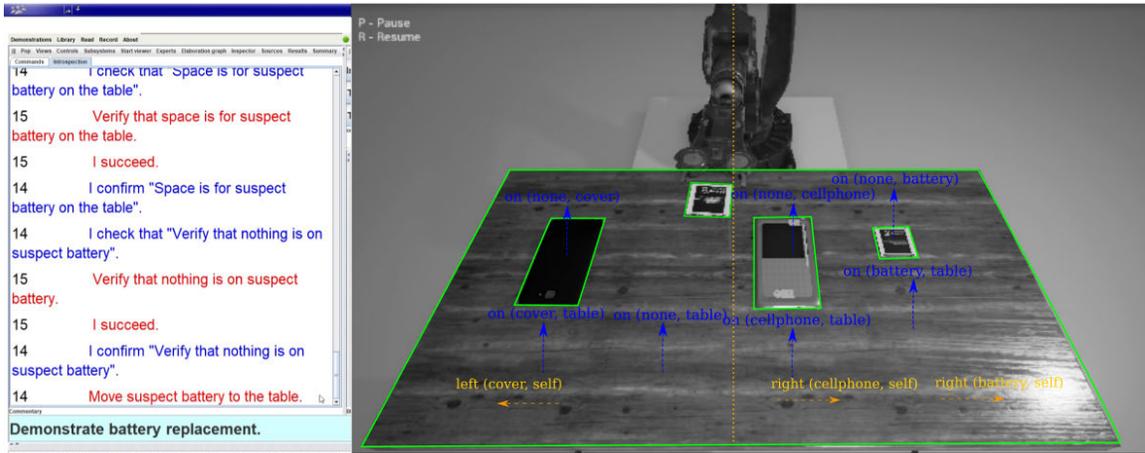
Language is primarily an instrument of thought, so are stories. I conjecture that the spatial aspects of our stories, which convey spatial and temporal relations among objects, events, and environments, are drawn from our spatial experiences. Just as we are able to verbally express these relations in our *external* stories, there must be cognitive mechanisms that allow us to extract these relations from our perceptions and actions and construct *inner* stories. These intuitions motivate what I call *the anchoring hypothesis*, which states that spatial experiences are inner stories that

¹This notion of inner language was introduced in (Berwick and Chomsky, 2016), who suggest that what people usually consider as language, that which we use to communicate with each other, is in fact an externalization of inner language.

chain together perceptions and actions by making evident the spatial and temporal relations among them through the use of symbolic expressions.

Nearly all animals have some form of spatial representation in their brains that enables them to navigate their environments, find shelter and food, and run away from predators. However, apparently, only humans can encode their spatial experiences in language, forming conceptual structures that represent spatial and temporal relations. I suggest that spatial and temporal relations are inherently present in perceptions and actions, but through our story-understanding faculty and our ability to assign symbols to perceptual and active information, we are able to make use of those relationships. For example, the symbol ON symbolically captures a visual relationship between two objects. I define an *anchor* as a function that constructs symbolic relationships from visual information. I present the motivation and the background of the anchoring hypothesis in the following section.

The anchoring hypothesis requires establishing (1) that spatial experiences can usefully be framed as inner stories, and (2) that composing stories from perceptions and actions can be done by assigning symbols to the spatial and temporal relations among these perceptions and actions. In order to test my hypothesis, I conducted case studies and experiments in which I investigated the relationships between stories, perceptions, and actions. Among those were (1) a study of how designers represent their spatial experiences using a variety of media, including drawings and physical models; (2) a virtual reality documentary that tells a story during which people spatially explore an environment; and (3) a virtual reality experiment in which subjects explore and verbally describe their environments as they do so. These studies provided initial evidence for my hypothesis and illustrated the importance of inner stories in processing spatial experiences. I gathered further evidence by implementing artificial intelligence models which can connect visual and symbolic systems by composing inner stories. For example, I demonstrated that a robot can replace a cell-phone battery by following the steps provided in English in coordination with a vision system that recognizes cell-phone parts and identifies the spatial relationships among them. (Figure 1-1). This experiment was particularly



— object recognition — vertical relation — horizontal relation

Figure 1-1: I develop a story-based visual problem solver, which guides a robot to replace a phone battery.

interesting because it illustrated how anchoring enables connecting symbolic and visual systems in order to solve problems.

1.2 Motivation and Background

Integration of symbolic and perceptual systems remains one of the most important topics in artificial intelligence research. Today’s computer vision systems, largely enabled by recent advancements in deep neural networks (DNNs), are able to extract pragmatically useful representations from images. Those representations are effective for multiple tasks, from object recognition to 3D scene reconstruction, and have become the mainstream approach in the field. However, deep neural networks continue to underperform in terms of symbolic reasoning. Superficially, DNNs can be explicitly trained to solve tasks that appear to require some form of symbolic reasoning (see Visual Question Answering (Kafle and Kanan, 2016), for example). However, even in such cases, DNNs are only learning the statistical dependencies between image and language (symbolic) representations and blindly producing outputs (Figure 1-2). Like a parrot ignorant of the meaning of the utterances it repeats, a DNN answering visual questions may easily amaze the audience yet



Q: What are they doing? **A:** Playing baseball
Q: What are they playing? **A:** Soccer



Q: Is the weather rainy in the picture? **A:** Yes
Q: Is it rainy in the picture? **A:** No

Figure 1-2: DNNs underperform in symbolic reasoning tasks. This figure depicts a Visual Question Answering (VQA) DNN model producing different answers to the same questions when they are posed in slightly different ways. (Kafle and Kanan, 2016)

remain unmistakably superficial. Therefore, we should be cautious about current DNN-based approaches to bridging visual and symbolic reasoning, especially in the context of language. We must clarify functional and computational imperatives behind perceptual versus symbolic systems, understand how and why they interact, and search for computational models that can demonstrate those interactions.

I propose a two-step approach to the problem of connecting visual and symbolic systems. First, we should explore the variety of ways in which people interact with their environments and characterize the symbolic and perceptual phenomena that unfold during those interactions. Second, we should develop algorithms and computational systems that are able to implement and test this characterization.

However, human spatial experience is an extremely broad concept, as it is a fundamental aspect of every human activity imaginable. Instead, we can focus on specific scenarios of spatial experiences by defining a scope within which we can discover, analyze and model particular observed behaviors. One way of limiting the scope is to focus on one sense only — the sense of hearing, or smell, or sight, for example. Although human spatial experience is multimodal, accounting for all of

the senses would hinder our ability to understand how each sense may contribute individually to the overall sensory experience. We can also limit the environmental context and geography, which is sometimes referred to as the *field*, where we can expect to observe only certain subsets of behaviors. For example, people’s spatial experiences on a playing field are very different from what they would be when working alone in an office. In a similar way, we can define specific activities for subjects to perform, so that comparison of their behavior will be easier to do.

I limited the scope of this work to studying visual explorations of environments and how these explorations were described by the subjects. The key competences to understand were (1) how people move around and explore an environment relying on their sense of vision and (2) how they describe what they see along the way. Throughout this thesis, this approach is demonstrated in a manner that allows for modeling different aspects of spatial experiences and furthering our understanding of symbolic-perceptual integration in the human mind. Although we cannot objectively measure the cognitive state of a person, verbal descriptions made by subjects can be used as an approximation of their cognitive states. Limiting the scope of my study to the sense of sight assists in narrowing down the inquiry so that we may better understand vision-specific features of spatial experiences.

1.2.1 Human Spatial Experience is an Inner Story

Why are inner stories are useful for encoding what we perceive and how we act within our environment? To elucidate the characteristics of human spatial experience, I identify three aspects of human intelligence that rely on our spatial experiences:

- **Communication:** Besides reacting to the immediate environment, one remarkable trait of human perception is the ability to communicate spatial experiences with others retrospectively. We have the ability to mentally represent things we have seen and actions we took in a way that we can recall and verbalize, and similarly we can imagine a spatial experience that is conveyed to us through language. For example, you tell your friend how to get

to a nearby restaurant by imagining yourself taking the path to the restaurant and describing it along the way.

- **Learning:** The previous example of describing the path to a nearby restaurant requires us to know where it is to begin with. We gather this type of knowledge by attending to various features of our environment such as a stop sign at an intersection, the presence of a brick wall, or the general size of a hall. We also learn environments based on our movements and activities; a library is a library because there are books there and people are reading them silently.
- **Proxy Action:** Once we have established an understanding of an environment, we can mentally examine the environment as if we were there and make modifications in our understanding. For example, imagine an empty white room with one of the walls painted blue. There is a picture on the blue wall. Opposite the picture is a door. Now imagine yourself looking into the room from the door. What do you see? This is an easy task, because we have the ability to perform spatial operations in our minds to represent our environments, reconstruct, and modify them (Kosslyn, 1980).

I believe that these three aspects of human intelligence can be enabled by our inner story apparatus. Through this mechanism, we are able to connect our actions and perceptions in an environment with our cognitive processes, which then facilitate communication, learning and proxy action. In my thesis, I argue that composing inner stories requires exposing spatial and temporal relations among our perceptions and actions. Our language already has a range of symbols that allows us to construct symbolic representations with those relations, such as "a cup ON the table." Symbols such as ON have previously been considered as mental-schemas (Talmy, 1983) or image-schemas (Johnson, 2017) in the semantics literature. In this work, I consider these symbols as references to particular visual, spatial or temporal relations that emerge in our experience. In order to form this referential relationship, anchoring relies on visual computing to identify these relations, and symbolic computing to assign corresponding symbols to them.

1.2.2 Seeing and Thinking Require Space

In this dissertation, I subscribe to the idea that our symbolic and perceptual skills are closely intertwined (Arnheim, 2004). According to this view, we humans do not simply use our eyes to collect and process information about our environment. Instead, we think with our eyes, ears, or hands in direct interaction with our environment. According to Arnheim, our ideas are tangled with imagery, and our perceptions inform our thoughts by identifying objects and material qualities (ibid). Creative practices, such as architecture, have been exploiting this fundamental unity of thinking and perceiving throughout history.

Theories of situated and embodied cognition are extensions of this view, positing that thinking happens not only through perception, but that it is enabled in particular by spatial perception. In this approach, thinking and perceiving do not merely operate as two different processes but work together through space. There is evidence in cognitive research and neuroscience that shows how space acts as a shared medium between thinking and perceiving (Tacca, 2011). Episodic memories, memories of autobiographic events, are formed in the hippocampal regions of the brain where spatial representations are also generated (Schiller et al., 2015). Subjects who experience verbal interference fail to combine geometric and non-geometric information in space (Shusterman and Spelke, 2005). Additionally, during spatial memory tasks, subjects who suffer from hemineglect, a neurological condition caused by brain damage, cannot remember the landmarks on either the left or right side of their recalled point of view (Unsworth, 2007). These and many other experiments demonstrate that symbolic and perceptual intelligence depend on the spatial abilities of human mind.

Perceptual and symbolic abilities cannot be considered in isolation from the environmental context within which they operate. Laboratories are particularly impoverished environments for the study of perception and cognition in contrast to everyday environments. This idea further motivated my research project to understand spatial experience as it unfolds in everyday environments. Similar

approaches were previously adopted by scholars who studied various aspects of cognition. In the context of memory, the psychologist Ulric Neisser showed that subjects' memories recounted in an everyday environment such as a living room are substantially different than those recounted in laboratory settings. He argued that the study of memory and other cognitive processes required experiments to take place in their natural contexts (Neisser, 2000). Following Neisser's theory, the cognitive scientist Edwin Hutchins developed a method he called *Cognitive Ethnography*. This method has been used to examine cognitive processes as they unfold in real-world settings (Hutchins, 2003). This method, also described as *Cognition in the Wild* (Hutchins, 1995), provides ecological validity to experimental studies and allows determination of relevant problems. The most recent and contextually relevant application of this method is the *Human Speechome Project*, initiated by Deb Roy, the director of the Cognitive Machines Group at the MIT Media Lab (Roy et al., 2006). Motivated to observe language development in infants, Roy and his team recorded nearly all of the activities of a newborn from birth to three years. Using audio-visual data, they discovered that there is a strong correlation between the learned corpus and the spatial context in which the infant is exposed to the language

1.2.3 Seeing and Doing are Composed into Stories

Our spatial experiences enable us to know where we are, how to navigate and discover new objects and places, and then relate what we have discovered to what we already know. Motivated by Winston's Strong Story Hypothesis, I postulate that spatial experience is an inner story that enables us to understand and interact with our environment. What are the mechanisms for composing inner stories when we explore our environments? I suggested that the critical component of such a mechanism is the ability to integrate perceptual and symbolic systems. In this dissertation, I investigate connections between vision and inner stories, which are inherently symbolic.

Symbolic computing works on predetermined symbols such as words in language. In contrast, visual computing operates on emergent information in our visual

observations rather than on predefined inputs. Visual computing works on sensible qualities such as colors, shapes, and materials, all of which are discovered in the environment during spatial experiences. The more we look, the more we can discover different shapes, materials, or additional information in the visual field of our perception. This provides the freedom to describe visual experiences in an indefinite number of ways and, crucially, not to be confined by a predetermined structure. Shape grammars offer a formal way to perform visual computing, and are used to analyze and generate designs with rules (Stiny, 2011). The inherent relationship between linguistic grammars and shape grammars makes shape grammars particularly relevant for this project.

Winston also points out that a story understanding system that relies solely on symbol manipulation is not complete:

Without connection to perception, story understanding reduces to disconnected symbol manipulation by a system that may appear to be quite intelligent, but depends too exclusively on linguistically supplied knowledge. Without connection to story understanding, an otherwise capable perception system can initiate reflex action, but lacks the ability to chain events together, to move backward and forward in such chains, to explain, and to predict (Winston, 2011).

A story understanding system can perform symbolic computations when the concepts are defined, for example, as commonsense knowledge or previously observed in the story. The connection of story understanding to perception is provided by visual computing, which allows computing with shapes, materials, and other sensible qualities (Knight, 1993). Algebraic calculations that allow for visual computation with spatial elements such as points, lines or surfaces have been introduced under the shape grammar formalism (Stiny, 2006). More recently, Stiny and Knight introduced making grammars (Knight and Stiny, 2015), which extend the shape grammar formalism from computing with shapes to computing with real, physical materials, and describe complex sensory qualities and spatial relations.

One can visually discover different types of knots on a string or different paint materials on a painting (Knight and Stiny, 2015).

From a computational point of view, our spatial experiences are inner stories in the following form: we perceive shapes, colors, and materials, making evident their spatial relationships, and chaining those spatially related observations into stories. Our actions, movements and attention drive this composition process through the anchoring mechanism.

1.2.4 Anchoring is the Key Mechanism for Composing Stories in Space

Research suggests that short-term memory has a duration of between 15 seconds and 30 seconds (Atkinson and Shiffrin, 1971). Any information that we receive earlier than that must be rehearsed or encoded in the long-term memory for later retrieval (ibid). This basic limitation alone demonstrates how crucial it is for us to be able to represent our visual and active experiences so that we can keep track of what is going on around us. If we had infinite short-term memory, we might have evolutionarily relied only on perceptual information, as it would never fade away from our immediate reach. However, after we experience something, we have to encode and compress the perceptual information in order to make it accessible.² Our inner stories and symbolic sequencing abilities enable us to perform these encodings.

Thus far, I have laid the groundwork for representing spatial experiences as inner stories, and have elaborated on basic computational mechanisms that would enable combining perceptual, active, and symbolic information. I argue that an essential mechanism is the ability to assign spatial and temporal symbols to what we observe. I refer to this assignment process as *anchoring* because this process is designed to group together visual observations with respect to a shared visual property, a spatial location, or the temporal order in which the observations are made. For example, in the sentence "The cup is on the table," the symbol *ON* anchors the cup and

²Research suggest that such compression and encoding take place in hippocampus. For example see (Squire and Alvarez, 1995)

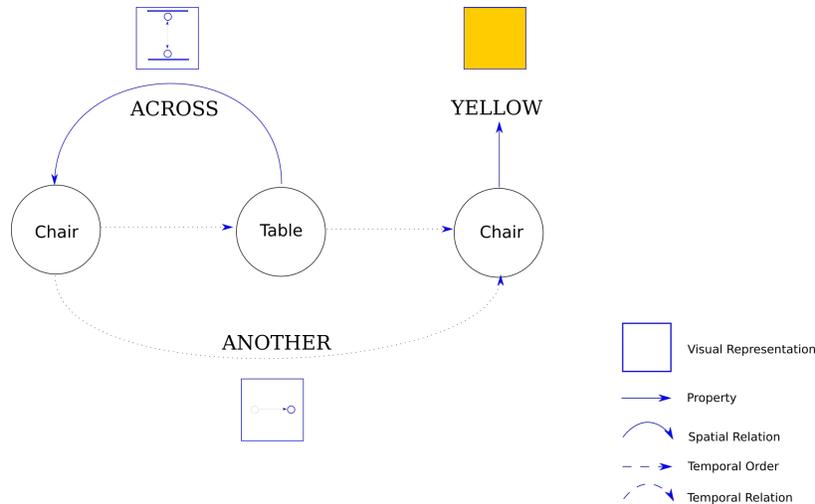


Figure 1-3: There are various symbolic domains in spatial language that rely on different visual representations.

the table with respect to their observed locations, which are exposed by perceiving the particular vertical relationship between two objects. Later, one can refer to this relationship by asking "Where is the cup?" obtaining the corresponding anchor as a result: *ON*.

In spatial semantics literature, spatial words are considered mental schemas that impose structure on space (Talmy, 1983). The main difference between the idea of mental schema and my approach is that anchors construct symbolic relationships from visual information. The word *ON* does not convey significant meaning by itself until there is a visual relation that we are interested in describing. A schematic representation of how anchors expose visual, spatial, and temporal relationships is illustrated in Figure 1-3.

Another similar and motivating idea is the *Feature Integration Theory* (Treisman and Gelade, 1980). According to this theory, despite appearing to be concrete and singular, our perception is the result of a fusion of multiple perceptual processes, each concerned with a different aspect of the world. Visual features are discovered by the brain in a pre-attentive state and integrated into a single representation that can be consciously attended to. The Philosopher Mohan Matten provides a similar account: "When a perceiver pays attention to a particular location, the features present at

that location (or thereabouts) get ‘bound’ together or ‘integrated,’ and she then sees a single image with a distinct characteristics” (Matthen, 2005, pp. 68). For example, consider looking at a letter of the alphabet printed on a page —“A” for example. We cannot see the “A” if we don’t also see the color it is printed in. We know that shape and color are separate, but to see the “A” here, we need both. In fact, all these features are identified by different specialized parts of the visual system, and are guided by coordinated actions and spatial contingency. A subset of features coincides in the same locality, so we perceive them as being the same (Matthen, 2005).

Inspired by feature integration theory, I suggest that we compose inner stories through anchoring by using visual features that are discovered in pre-attentive states, or in sub-symbolic layers of vision. Visual computing provides an easy way to pick out those features and then change our attention from a material quality to surface properties or to spatial locations of objects. However, implementations for computer vision are not as straightforward. With the advent of DNNs, the field of artificial intelligence has achieved one method of accessing those sub-symbolic representations. Accordingly, I use DNNs for extracting those representations, such as representations for the boundaries in an environment and for identifying objects and their relative locations in space.

1.3 Immersive Technologies Enable the Observation of Spatial Experiences

I suggested that we should define a scope within which we can observe particular behaviors of people during their spatial experiences. Accordingly, I limited the scope of the study to the visual exploration of environments and the verbal descriptions made by people during those explorations. Even within this scope, observing a person’s spatial experience and modeling his or her behavior are challenging tasks. There are many parameters and decisions to make, including the type of the environment to be explored, the nature of the data to be collected, and the methods

used to record the experiment. I decided to design these experiments within virtual reality (VR), because VR allows for quick testing of different situations without drastically changing the nature of the experiments. Previously, VR environments were used in research settings, for example, for studying the formation of episodic memories (Dimsdale-Zucker et al., 2018), and human action and navigation (van Veen et al., 1998). Similarly, in my research, I conducted visual exploration experiments in VR environments, collecting a variety of data from the participants. This data included the locations and orientations of the virtual camera, which enabled me to record both location and the gaze of the participant; a series of visual data such as RGB and depth, which were conveniently generated after experiments based on the camera parameters; and audio recordings that contain the verbal descriptions made by users.

In addition to its usefulness for spatial experiments, virtual reality is also a medium for creating spatial stories that can be an immensely creative tool in the designer’s hands. In this thesis, I also illustrate how designers and storytellers were able to adopt this medium to create novel spatial stories —stories that invited people to explore and actively engage with spatial content.

1.4 Overview

In chapter 2, I present the theories and basic assumptions about human space perception and spatial experience from an architectural and anthropological point of view. This information supports the thought that a central aspect of spatial experience is that it consists of multiple egocentric and allocentric representations that work in coordination. The human visual system must provide the information necessary for creating different representations based on the task at hand.

In chapter 3, I introduce two case studies on creating and representing spatial experiences. Through these case studies, I investigated the variety of ways in which people interact with their environments, compose and understand stories, and express their spatial experiences using words, images, drawings, or materials. The



Figure 1-4: Various methods that represent the aspects of spatial experiences, which include drawings, physical models and immersive simulations. In this example, the student introduces various drawing elements to describe the spatial experiences he has had in East Cambridge.

first case study is *September 1955*, a virtual reality documentary of the Istanbul Pogrom. Through this immersive experience, I explored how people construct a variety of inner stories based on their active participation within a virtual environment. Pivotaly, this work showed the flexibility of the perceptual processes that individuals use to understand novel environments. The second case study is the *Computational Ethnography and Spatial Narratives* class that I taught at the Boston Architectural College in Spring 2016. With my students, I investigated how we represent our observations and spatial experiences using different media, such as drawings, diagrams, or physical models (Figure 1-4). The work covers a wide range of representational techniques, modes of exploration, and creation of unique spatial experiences. Insights gained in this section motivated a more structured approach to observing and analyzing spatial experiences using virtual reality.

In chapter 4, I present a particular methodology, *See, Act and Tell*, which I developed for observing spatial experiences during visual explorations within virtual reality environments (Figure 1-5). With this methodology, I created a dataset of the locations and movements of participants in an environment, their visual experience



Figure 1-5: I present a novel approach to study human spatial experiences, *See, Act and Tell*. I investigate how story understanding enables spatial experience.

and attention, and their verbal descriptions. Having this type of rich, spatial dataset enabled me to model the processes by which people compose spatial descriptions from active explorations. The results of this study also highlighted the importance of the domain specificity of spatial language. A group of words referring to a certain aspect of spatial experiences, such as visual features, spatial locations and relationships, or temporal relations between observations, defines the domain. Each domain appears to rely on a particular aspect of visual observation and an accompanying visual representation. For example, the word *ACROSS* relies on the perceived boundaries of an environment in which one can contrast it to the words *THROUGH* or *ALONG*, but not to words that rely on other representations. For example, the word *YELLOW* relates not to boundaries but to another, color-related representation. Furthermore, some domains rely on temporal relations between observations, such as the word *ANOTHER* or *SIMILAR* (Figure 1-3). I compared the domains that emerged during a participant's visual exploration of space to previous categorizations in spatial semantics literature (Talmy, 1983; Langacker, 1987; Rosch, 1973). I concluded in this chapter that much of our flexibility in describing and reasoning about our environment comes from our ability to filter,

select, and compose, both visually and symbolically, different domain specific representations.

In chapter 5, I present the steps I took towards developing an artificial intelligence system that can learn, understand and communicate its spatial experiences by composing inner stories and relating their observations in terms of those stories. I present three demonstrations. The first is a demonstration in which a robot learns to extract boundary representations from vision and identify distinct parts of its environment. In a second demonstration, a robot replaces a cell-phone battery by identifying the spatial relationships between its parts. In the third demonstration, I introduce a story-composing system that can generate interpretable verbal descriptions from visual explorations, augment those descriptions with common-sense knowledge, and answer natural language questions regarding its observations.

In chapter 6, I present the contributions of my dissertation to the fields of design, media studies, and artificial intelligence. I discuss the further applications and possible future studies on the anchoring framework.

Chapter 2

Human Spatial Experience

In the first chapter, I introduced the main themes of this dissertation. One central idea is that human thought and perception are bound to space and time. Naturally, every experience is spatial, in the sense that it is unimaginable to speak of a "flat experience" stripped of spatial character. However, the word "experience" has a broader meaning, such as in "emotional experience," "user experience" or "life-changing experience." In order to prevent confusion, I will refer to concrete human experience in space in the scope of this work as "spatial experience." This definition confines what I mean by experience to one that occurs in a particular time and place, and which involves interactions between humans and their surrounding environments. My research deals with this type of experience.

In this chapter, I will present key ideas and theories about our perception of space and our spatial experience —two interrelated yet different aspects of the human mind. In the first section, I present a background and contemporary research regarding our perception of space. In the second section, I move into a discussion of spatial experience. Finally, in the third section, I present computational approaches to modeling the perception of space and spatial experience.

2.1 Perception of Space

More than two millennia have passed since Aristotle authored his philosophical treatise on space (Hussey, 1983). Since then, the study of space has been an integral part of scientific and humanitarian inquiries. Today, it is clear that space is more than a mere container of objects and an absolute quality of the environment: It is a relational quality of human perception that takes part in every cognitive and performative skill that a person exhibits—from memory to language, to problem solving, to creative practice. Therefore, the perception of space constitutes a relevant problem for a broad range of scholarship beyond the disciplinary concerns of psychology, cognitive science, and neuroscience. Moreover, space has been extensively discussed in the disciplines of sociology, anthropology, architecture and urbanism, geography, and more.

2.1.1 Philosophical Underpinnings: Nativist and Empiricist Agendas

Contemporary understandings of the perception of space have evolved since the establishment of 19th century experimental psychology and psychophysics, particularly in Germany and Britain. The proliferation of experimental studies that focused primarily on vision gave rise to new theories that regarded space as a psychological product. In 19th century, two opposing approaches were dominant. On the one hand, *empiricists* such as Wilhelm Wundt and Hermann von Helmholtz, who followed the ideas of George Berkeley, argued that space was a product of experience and that there was no spatial quality in the "real world." On the other hand, building on Kant's philosophy, Johannes Peter Müller, Edwald Hering, and other *nativists* believed that space was an innate quality of the mind (Hatfield, 1990; O'Keefe and Nadel, 1978). Fruitful discussions between these two camps—or sometimes violent arguments, as in the case of Helmholtz and Hering—are summarized in William James's *The Principles of Psychology* (James, 1890). As a response to those empiricist and nativist accounts, James proposed his

sensationalist theory, according to which the idea of space is a product of complex mental operations among senses that are already primitively spatial.

The philosophical problem of the perception of space that gained scholars' attention was rooted in the observation that all sensory surfaces of the human body, such as the retina or skin, are two-dimensional. How then did the human mind, without any three-dimensional input, produce the immediate feeling of three-dimensional space? 18th century philosophers saw that there was an unnecessary dichotomy between the concept of absolute space and the concept of mentally produced relational space; thus they abandoned the concept of absolute space that existed independently of a perceiving subject (Turner, 2014). Among the prominent scholars of the era, George Berkeley and Immanuel Kant proposed two epistemologies of psychological space. A follower of Locke's Associationist school in Britain, Berkeley believed that the concept of space derived from *Experience*, primarily through the interactions between the visual and tactile senses (Berkeley, 1922). According to Berkeley, one could visually understand distance, size, and shape only by relating them to the immediate tactile feelings generated by movement. He said that visual impressions only suggested the types of tactile feelings when one moves along places and that "neither distance nor things in places at a distance are themselves, or their ideas, truly perceived by sight" (ibid). In this approach, the mind was a blank slate, a *tabula rasa*, which did not impose any innate structure to space.

By contrast, Kant proposed that space was not an empirically acquired concept but was innate (Kant, 1933(1787)). He asserted that all knowledge came from experience (*a posteriori*) in accordance with the innate principles of the mind (*a priori*). Space was the main *a priori*, as it was not derived from experience and it permeated all knowledge. In fact, for any sense to be represented as "outside and alongside," the knowledge of space should have already existed in the mind:

Space is not something objective and real . . . arising by fixed law from the nature of the mind like an outline for the mutual co-ordination of all

external sensations whatsoever. (Inaugural Dissertation, 2004 [1770]).

The distinction between the two approaches of Berkeley and Kant, described above, is of primary importance in their theoretical foundations and their conception of space perception. In Kant's epistemology, the primitives of thought are unanalyzable and not decomposable into simpler units or processes. However, Berkeley, as an empiricist, believed that mental processes are natural associations and composed of elemental components (Hatfield, 1990). Accordingly, in Kant's account of perception, spatiality is purely a function of the mind that provides a unitary framework of perception, whereas Berkeley's conception assumes that the senses have inherent structures that can form the concept of space through the associations among them in a relational framework (Hatfield, 1990). Neither Kant nor Berkeley provided a detailed explanation of their theories or supported their approach with physiological evidence. These tasks were left to later generations of physiologists and psychologists.

2.1.2 Psychophysics and the Study of Human Eye

By the start of the 19th century, scientists had identified a series of visual cues, including the accommodation and convergence of the eyes, the apparent size of objects, and atmospheric perspective, all of which influenced one's perception of space (Hatfield, 1990). With the motivation of discovering physiological processes that integrated these cues into the visual perception of space, a nativist agenda was set by Johannes Müller, who believed that all the answers to the problem of space perception was in the specificity of neural pathways (Müller in Boring, 1942). According to Müller, sensory qualities were produced by specific nerve channels on which the source of energy had no determining effect. For example, when a visual nerve was triggered by touch (such as by poking the eyeball), it produced a visual impression, not a tactile one. He called this principle *the law of specific nerve energies*. Space, according to him, was nothing but a topographical arrangement of the optic nerves that define a visual field and, thus, was a pre-wired quality.

Although it differs in detail, Müller's work characterizes other nativist approaches of the era, most of which examine the physiology of the eye to find inherent structures that would allow the emergence of space.

The common neglect of nativist approaches is that they assume the spatial organizations of sense organs are completely parallel to the subjective impressions of the space they produce. For example, Hermann Lotze's concept of *local signs* attempts to show that the spatial relationships between two objects in space are preserved on their retinal reflections, or their local signs (Lotze, 1892). Therefore, any spatial character can be read off the local signs of objects. This idea takes a variety of forms in other nativist theories, such as that of Hering, who proposed that space was a property of every sensory signal and that each nerve carried its spatial signature along with visual information. Although today it is experimentally proven that visual images on the retina carry their spatial structures to deeper layers of visual processing (where the foveal region is magnified), this theory falls short in accounting for the three-dimensional character of space (James, 1890; O'Keefe and Nadel, 1978). However, as we shall see later (such as in William James' account), local signs might provide a reference frame against which spatial qualities are produced.

In addition to the assumed spatial parallelism between physiology and psychology, nativists since Kant have regarded Euclidean geometry as the true and objective structure of space: not only was this geometry manifested in the perceived space, but its axioms were accessible by judgment. In other words, Euclidean geometry was already present in the mind's eye, and that was clearly a proof of the innate nature of space (Hatfield, 1990; Turner, 2014). These assumptions were strongly criticized by empiricists, most notably by Helmholtz. He showed first and foremost that visual phenomenon did not obey the parallel postulate and that it was better explained with Riemann's geometry. He concluded that all geometrical axioms were learned, and the perception of space did not presuppose any geometry (Kahl, 1878). His own view,

which he called *unconscious inference*, assumed vision and touch merged deep within cognitive processes to produce the perception of space. This view is regarded as the precursor to the modern information processing approach and unsupervised learning models (Turner, 2014).

A tactile- and movement-based grounding for visual space was a common belief among the empiricists. For example, Spencer posited that all perceptions, including the visual and tactile, were decomposable into relative positions of subject and object, which were only possible through the movement of the subject (Spencer, 1895). He argued that the concept of space emerged from the uniform associations between visual and tactile senses and that the senses did not carry spatial information by themselves. What empiricists neglected to see as in this example, was that the elementary zero-dimensional structures were assumed to generate two- and three-dimensional ones. Space was not spatial in nature but, as James described it, instead a "mere symbol of succession." (James, 1890). Wilhelm Wundt, in his genetic theory, proposed a derivative of this approach that regarded local signs as elementary structures. Although there was no spatial quality in sensations, he suggested that the local signs of the sense organs would provide the basis for acquiring spatial knowledge. Local signs of touch would be associated with the local signs of vision through a process he called *psychic synthesis*. More importantly, he made the observation that the keenness of vision – the immediate visual impression, relying solely on the proximity of retinal elements – and the apprehension of directions and distances in the field of vision are not the same (Wundt and Judd, 1897). Yet, James, who stated that it was the "flimsiest" theory of its kind, also dismissed Wundt's theory. James responded, "Retinal sensations are spatial; and were they not, no amount of 'synthesis' with equally spaceless motor sensations could intelligibly make them so" (James, 1890, p. 907).

2.1.3 From Sensationalist to Ecological Psychology: “Ask not what’s inside your head, but what your head’s inside of”

At the turn of the 20th century, William James and Ernst Mach similarly identified the common mistakes of earlier approaches. In the *Contributions to the Analysis of the Sensations*, Mach argued that the separation between the real world and the perceived world is virtual, and accordingly, boundaries between ego, mind, body, and world are only provisional and practical. In fact, there is only a connection of sensations, and it should be the only subject of psychological investigation (Mach, 1897). This idea can be regarded as a precursor to Dewey and Bentley’s *theory of transaction* (Dewey and Bentley, 1960). Mach’s sensationalist account, later called *phenomenalism*, acknowledges the continuity of sensation (what James calls the “stream of consciousness”). His idea of space-sensation is grounded in the continuity of one’s motor behavior. He asserted, “The will to perform movements of the eyes [and later the head and the body], or the innervation of the act, is itself the space-sensation”(1897, p. 60). Moreover, Mach argued that physiological space is different from geometrical space. Geometrical space (i.e., recognition of a square and its transformations) is a product of the intellect and only slightly related to physiological space. However, physiological space has certain primitive features such as similarity (not geometrical, but perceptual), symmetry, and orientation.

Mach’s observation of physiological space was described by William James as the primitive sensations of space (James, 1890). James posited that the space that people perceive as real and singular is in fact produced with operations of assimilation, superposition, and summation of primitive spatial qualities —*voluminousness*— of the senses. A sense is by nature spatial: A higher-level computation is not necessary to differentiate vertical from horizontal, close from far, or small from big. Additionally, for James, there is not one unitary space but multiple relational spaces, each with comparable spatial properties based on their relative local signatures. The perception of space, therefore, is defined as a dynamic and continuous process of discrimination and association between multiple sense-spaces.

Nativist, empiricist, and sensationalist accounts of the perception of space in the 19th century were carried forward to the 20th century through a variety of studies. For example, Gestalt psychology emerged as an interpretation of nativism, pioneered by Kurt Koffka, Max Wertheimer, and Wolfgang Kohler. At the core of their theory, Gestaltists argued that objects and their environment were perceived together, according to the innate principles of mind (Wertheimer, 2012). Gestalt psychologists believed that the nature of perception demanded that each of its components (object and subject) be in dynamic interaction and that it was composed of both physical and mental properties. In the context of the perception of space, Kurt Koffka argued that "our space perception in all three dimensions is the result of organized brain activity and we can understand our space perception only in terms of organization (Koffka, 2013)." Well-known Gestalt principles exhibited what Koffka meant by organization.

James Gibson was one of the 20th century's most influential authors on the perception of space. Building on the ideas of William James and the Gestalt school, he synthesized a theory of perception that he called *ecological psychology*, (Gibson, 2014, 1950; Turner, 2014). Gibson changed the general understanding of perception by proposing that space was not reconstructed in the mind but rather was directly accessed in the environment. According to this theory, an observer actively seeks the invariants in the environment, and these invariants are not described by the patterns of the visual field but inherently exist in the visual world. He asked, "What are the qualities in the environment that matter to a person?" Accordingly, he coined the term affordance, which is the perceived quality of the environment that affords a certain action as the main reason we perceive space: to guide movements and to avoid obstacles. Similar active perception ideas were also present: Dewey and Bentley defined perception as a transaction, which is "action of and by the world in which the man belongs as an integral constituent" (Dewey and Bentley, 1960). Similarly, in *Visual Space Perception*, William Ittelson defined perception as an action (Ittelson, 1960). This view has been more recently supported by the

philosopher and cognitive scientist Alva Noe, who stated that perception is not something that happens to us but rather something we do (Noë, 2004).

2.1.4 Contemporary Cognitive and Neuroscience

A great deal of work has been produced in neuroscience and cognitive science, particularly under the information-processing agenda starting in the 1950s. In particular, the discoveries of the receptive field—a portion of sensory space that elicits neuronal response—and the simple and complex cells in the primary visual cortex, indicated a visual mechanism in brain that integrates low level orientation features with complex structures. Similarly John O’Keefe and Lynn Nadel discovered that there were cells in the hippocampus that spike only in particular areas in an environment, naming these cells *place cells* (O’Keefe and Nadel, 1978). David Marr was one of the most influential scholars in vision science, who, with the inspiration of this feedforward processing in brain, developed a computational theory of vision. Marr suggested that a three-dimensional world is gradually reconstructed in the brain from the elementary structures of visual stimuli such as edges, textures, and orientations (Marr, 1982). Yet, the inherent solution of such an approach remains unanswered, as documented in the previous paragraphs.

Hippocampus plays a central role in spatial experience

Our brains devote a tremendous effort to resolve space, integrate sensory information into practical information so that it can plan where to move, where to sit, and how to remember a place so it can navigate back when necessary. Basic characteristics of spatial perception have been discovered through studies on the rat hippocampus (O’Keefe and Nadel, 1978; Ranck Jr, 1973; Chen et al., 2012). Different firing patterns in place cell regions (see Figure 2-1) allow a rat to represent its location relative to the objects in the environment. While place cells respond to different visual cues in the environment, they also integrate proprioceptive information as to direction and velocity, as well as sensory information other than the visual, such as sound or odor.

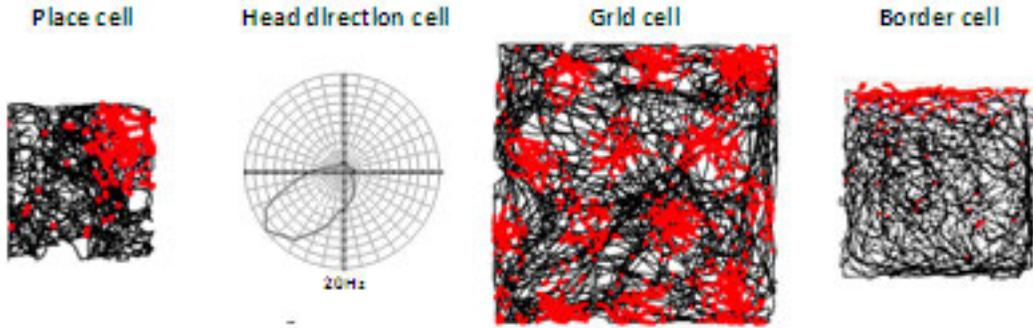


Figure 2-1: Spiking patterns of spatial cells of a rodent brain in a square environment, from (Marozzi and Jeffery, 2012). A *place cell* has a *place field* and is activated only in a particular part of the environment. A head direction cell is activated only when the rodent is oriented towards a particular direction in the environment. A *grid cell* has an activation pattern that has a particular distance interval which builds a "grid" when recorded in multiple locations. A *boundary cell* (border cell) is activated when the rodent is near a boundary towards a particular direction. In this example, the boundary cell is activated near the north wall of the room.

The spike pattern of a particular place cell is correlated with the change in size, orientation, and shape of the environment (O'Keefe et al., 1998).

In addition to place cells, there are other neural compositions in the brain such as *grid cells*, which have special grid-like firing fields in the environment; *head direction cells*, which are activated when the rodent is facing its selected direction; and *boundary cells*, which is activated when the rodent is near a border facing a particular direction. These cells contribute to the formation of reliable information about the environment (Figure 2-1). Additionally, the parahippocampal gyrus (PPA) region in the human brain is selective to environmental structures, which integrate low level visual features with higher level descriptions that allow detection and recognition of individual spaces.

As we can see from the significant work undertaken by scholars, there is more to the perception of space than the realization of the three-dimensional character of space. Today, there is a vast amount of work demonstrating that the perception of space is closely related to memory and navigation (Tolman, 1949; Neisser, 2000; O'Keefe and Nadel, 1978; Montello et al., 2014; Wilson and McNaughton, 1993),

cognition, (Kosslyn, 1980; Tversky, 2011; Newcombe and Huttenlocher, 2003), language (Spelke, 2002), and action (Sperry, 1952; Sheets-Johnstone, 2011; Millar, 2008; O'Regan and Noë, 2001). Moreover, discourses on situated cognition, ecological psychology, enactivism, and phenomenology provide a conception of the human mind that is not confined to the neural processes in the brain but instead consider particular social, cultural, and material environments. Such a conception motivates another question: "How does human perception (and cognition) come into being in space?"

2.2 An Argument for the Spatial Mind

Spatial experience is the subjective phenomenon of being immersed in a vast area within which one can locate oneself, qualitatively discriminate left from right and close from far, and identify places one can move along in, around, or between. As James Gibson notes in *The Ecological Approach to Visual Perception of Space*, the life of a person is driven by and attached to the experience of space (Gibson, 2014). Understanding the experience of space constitutes a major component of the social sciences, the arts and architecture. Recent advances in computer vision and immersive technologies make this type of understanding relevant for an even broader community as technology penetrates our spatial experiences more than ever. With assistive robots inhabiting our everyday environments, classes taking place in simulated environments, and remote working becoming a standard for workspaces, the study of human spatial experience will become a central topic for scholars and practitioners from all fields. In the past, creative practices, especially architecture, were privileged to examine and create spatial experiences because concrete experience took place only in physical space. It is therefore a good starting point to look into architecture and related fields to see what they offer to the study of spatial experience.

2.2.1 Architects are Skilled Practitioners of Space

Architects have always been interested in understanding spatial experiences, which they could then integrate into templates or rules for future uses. How does an architect translate spatial experiences into her designs? Donald Schön, who famously framed the design process as a *reflective practice*, observes that "the practitioner (architect) allows himself to experience surprise, puzzlement, or confusion in a situation he finds uncertain or unique" (Schoen, 1983, p.68). The design process is an experiment that recursively generates novel phenomenon for the architect to tackle over and over. The source of an architect's ability to create is not some abstract knowledge about architecture, but a recursive process where she finds opportunities to pay attention to her own interactions and experiences. Designers rely on their own experiences within the material world, induce changes with their actions, and over time build a repertoire of "examples, images, understandings, and actions" (ibid). This observation about spatial experience is extremely relevant. Schön's repertoire is not a knowledge base. It is a skill set an architect builds (habituates) over time; she distills this skill set from her experiences and applies it to novel conditions when she is called upon to design. More often than not, she finds herself in a unique condition, for which she may not yet have developed an understanding or a suitable action, and for which she may not have a familiar image. Or, she may find herself in a familiar condition to which she brings a new understanding that completely changes her perception. Constant discovery and re-evaluation, driven by a cycle of acting and perceiving, is the backbone not only of architecture but of any creative practice.

2.2.2 Dwelling versus Building

In the *Perception of the Environment*, the British Anthropologist Tim Ingold presents an account of livelihood, ranging from wayfinding to cultural and creative practices (Ingold, 2002). Two contrasting views of the human world are presented in the book: the dwelling perspective (the one Ingold defends) and the building perspective. The building perspective represents the major assumptions of cognitivism and objective

thought: humans deploy their cognitive skills, which exists independently of their environment, to perceive and reconstruct the world according to a cognitive schema. This perspective posits a Cartesian, disembodied intellect that performs in isolation against a frozen and neutralized environment. By contrast, according to the dwelling perspective, humans do not reconstruct an existing environment in the mind but create it in place. They do so because they are integral parts of their environments: they can only exist by inhabiting them or dwelling in them. Ingold points to James Gibson's ecological psychology and the phenomenologies of Martin Heidegger and Maurice Merleau-Ponty as the primary sources of this framework. Accordingly, I will examine each point of view by looking at how these perspectives were adopted by scholars, what they imply for the experience of space, and how they relate to the works of architects and artists. The discussion will be presented regarding two topics: people's experiences of particular localities (places) and the movement of people between places

Dwelling, according to Martin Heidegger, is not a behavior that people adopt in order to live in their environments. In "Building, Dwelling, Thinking" he observes that there can be no a priori condition to dwelling (Heidegger, 1971). Dwelling, or being-in-the-world, is a prerequisite of human existence; thus, the human mind presupposes a corporeal place. In this phenomenological approach, space is not "out there" but is embodied (Ingold, 2002). Anthropologists Setha M. Low and Denise Lawrence-Zuniga define embodied space as "the location of where human experience and consciousness take on material and spatial form" (Low and Lawrence-Zúñiga, 2003). Their anthropological stance to understand place regarding orientation, movement, and language draws from the philosophy of Merleau-Ponty. A strong opponent of objective thought, Merleau-Ponty dismisses the concepts of "object" and "subject" in favor of a singular intentional body (Merleau-Ponty, 2013, 1964). According to him, "the primary condition of all living perception is spatial existence." It is the body, not as a mechanistic object or a "bundle of functions" but as a phenomenal body in which vision and movement are intertwined, that enables

one to take a position in objective space:

The possession of a body implies the ability to change levels and to ‘understand’ space, just as the possession of a voice implies the ability to change key. The perceptual field corrects itself ... because I live in it, because I am borne wholly into the new spectacle and, so to speak, transfer my center of gravity into it (Merleau-Ponty, 2013, p. 251).

As presented by Merleau-Ponty, dwelling is a unity of acting and perceiving through a body. In *Ways of Worldmaking*, Nelson Goodman defends a similar embodied and pluralistic view of the world. He states, "Knowing cannot be exclusively or even primarily a matter of determining what is true... if worlds are as much made as found, so also knowing is as much remaking as reporting"(Goodman, 1978, p. 17).

In contrast to Goodman’s multiple worlds, the building perspective is the search for a single universal, either in the innate structures of the mind, according to cognitive scientists, or in the cultural schemata established in particular societies (Ingold, 2002). For example, Herbert Simon defines the human as an information-processing system, the goal of which is to adapt to its complex environment, an idea that is matched also by connectionists and behaviorists (Simon, 1969). On the anthropological side, the building perspective is best represented by Edward T. Hall’s idea of proxemics. In his entry remarks to *The Hidden Dimension*, he says, "All my books deal with the structure of experience as molded by culture, those deep, common, unstated experiences which members of a given culture share" (Hall, 1910). Accordingly, he introduces the three spatial distances that all people employ when they experience space: intimate, personal, and social. The building perspective, then, is characterized by goal-driven behavior and planned action in a homogeneous uniform space (which Simon admits is Euclidean geometrical space), which take the experience of space out of its natural place —from the subjectivity of the perceiving body.

By contrast, from the dwelling perspective, experienced space – or lived-space, as suggested by philosopher O.F. Bollnow – is not homogeneous but, rather, qualitatively discriminated (Bollnow, 1961). The ways in which humans perceive and move through space dynamically determine the center (which is the person's body), the major vertical and horizontal axes, and the directions of front, back, left, and right in the space that he or she experiences (Bollnow, 1961; Merleau-Ponty, 2013; Tuan, 1977; de Certeau, 2011). The anthropologist Yi-Fu Tuan argues that the upright structure of the human body establishes this unique human space (Tuan, 1977). Merleau-Ponty also distinguishes this space, which he defines as the *anthropological space*, from the geometrical space. Similarly, in *The Practice of Everyday Life*, Michel de Certeau observes that “[experienced] space exists when one takes into consideration vectors of direction, velocities, and time variables.” Tim Ingold in his essay "Temporality of Landscape" points out that temporality is a fundamental part of the experience of space (Ingold, 2002).

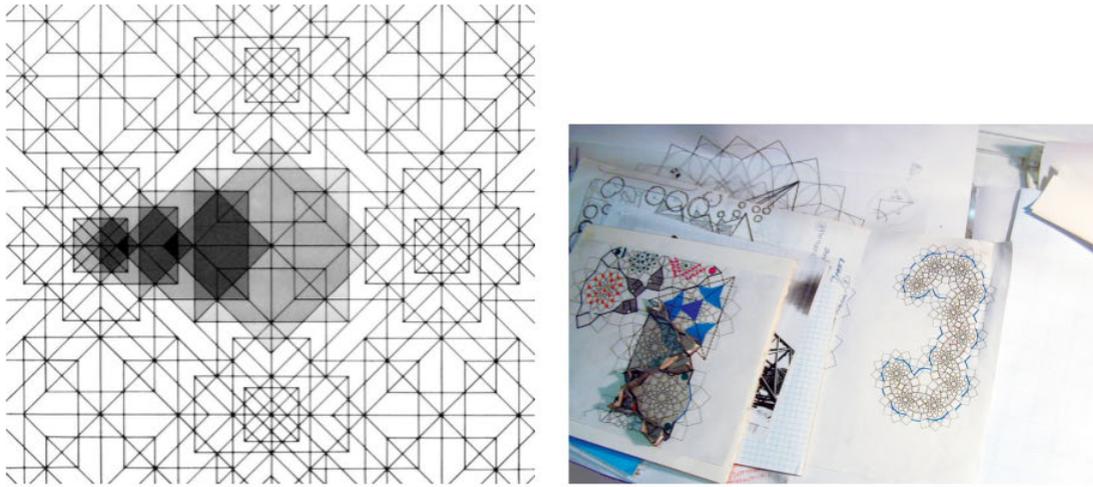
The Norwegian architectural theorist Christian Norberg-Schulz proposes a formal description of the experience of space (Norberg-Schulz, 1971, 1980). He argues that the experience of architecture cannot be described through universals, as it is endemic to a particular locality and a personal perspective. However, this phenomenological, dwelling approach is impugned when he introduces existential space as a unity of perceptual space (the subjective experience) and a universal space schemata embedded in the mind. He argues that egocentric perceptual space is assimilated into stable experiences through the subject's schemata; thus, the idea of existential space appears to be a hybrid of the dwelling and building perspectives. Norberg-Schulz's theory is strongly influenced by Piaget's developmental psychology, where he makes use of schemata to define spatial universals: centrality and place, direction and path, and area and domain. Nevertheless, these categories prove to be practical in characterizing architectural space in relation to perception and movement without making an assumption that "architectural space really exists" (1972). German architect Jurgen Joedicke offers a similar view, but instead

of a spatial schemata, he proposes that there are certain constants explained in cultural backgrounds (Joedicke, 1985). His idea of the experience of indoor space is a one-directional flow of information from space to person, where "subjective filters" somehow produce "experience."

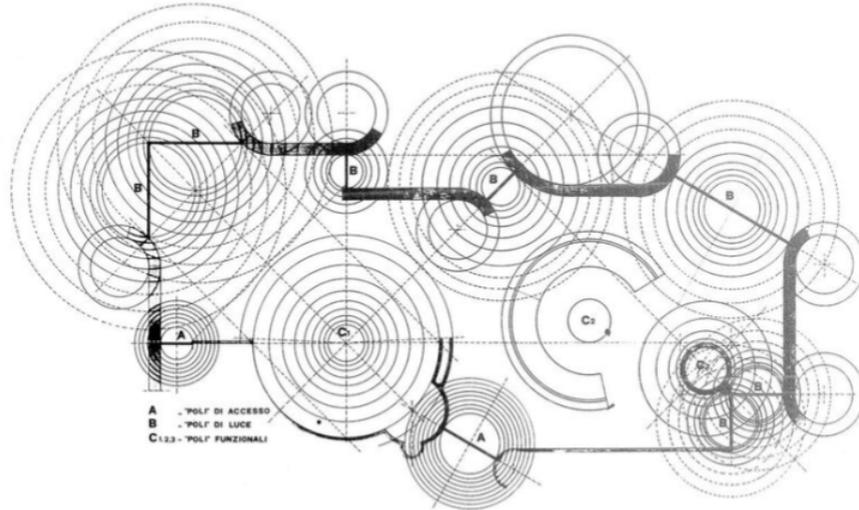
The Swiss architect Sven Hesselgren's approach to characterize space can be considered closer to a dwelling perspective when compared to the approaches of Norberg-Schulz and Jodicke. In *The Language of Architecture*, Hesselgren remarks on Hering's observations about a roomy feeling —what James later describes as *voluminousness*— arguing that such feeling is essential to discriminate the infinite outdoor and the bounded indoor, as the perception of space is strengthened when one moves his or her attention in space (1969). Moreover, such an indoor feeling of voluminousness directs the attention outwards in all directions to the "room" as "something that surrounds oneself" This character is lacking outdoors, in landscapes and townscapes. Later, Swiss critic Vogt-Goknil integrates this distinct character of indoor into her concept of *Umraum* (surrounding space) (Vogt-Goknil in Norberg-Schulz, 1972).

Architects such as Walter Netsch sought to quantify the character of indoor space in their design processes (Netsch, 1969). Netsch employed what he called a *field theory*, arguably inspired by Gestaltists, for the spatial organization of his designs (Figure 2-2(a)). Norberg-Schulz points out similar combinatorial and intersectional field theory attempts in the works of Dientzenhofer, Guarini, Portoghesi, and Gigliotti (Figure 2-2(b)). These hybrid perspectives on the human world are comparable most notably to the pure building perspectives of Christopher Alexander's *Pattern Language* (1977), Michael Benedikt's notion of *isovist* (Benedikt, 1979), or Bill Hillier's theory of *space syntax* (Hillier, 1999), all of which completely disintegrate the human element from space.

Regarding architectural scale, one last remark should be made on the inherent problem of representation. In *Architecture as Space*, the architectural historian



(a) Walter Netsch's implementation of his *field theory*. Netsch used a system of overlapping lattice to trace out boundaries or shade different areas (Netsch, 1969)



(b) A floor sketch plan by Gigliotti and Porthogesi based on *Field Theory*

Figure 2-2: Inspired by the Gestaltists, architects developed formal methods to quantify experiential aspects of their designs.

Bruno Zevi argues that architecture has "infinite dimensions" and that no means of representation can capture what architecture really is—even a video of a building is limited representation based on a finite set of observations (1974). He asserts:

There is a physical and dynamic element in grasping and evoking the fourth dimension through one's own movement through space. Whenever a complete experience of space is to be realized, we must be included, we must feel ourselves part and measure of the architectural organism (Zevi, 1974, p.59).

2.2.3 Spatial Ability and Spatial Knowledge

Using Ingold's framing of dwelling and building perspectives, we were able to compare different approaches to examining spatial experiences. One important distinction is that from the building perspective, the world is seen as universal and invariant, whereas from the dwelling perspective the main concern is human activities and interactions with their environments. Based on the task at hand, we humans adopt those perspectives interchangeably, and think about our environments both in terms of universals and subjective experiences. Our inner stories must make use of both types of understandings when we experience our environments. Knowing spatial universals, such as knowing that you are in an office, is clearly advantageous for us to understand our environment, because they provide handles for relating our inner-stories to previous experiences. On the other hand, through our embodied experiences we attend to different parts of environments, creating the content for our inner stories. Knowing and doing are essential characters of the building and dwelling perspectives, respectively. The anthropologist Yi-Fu Tuan offers a complementary distinction between *spatial knowledge* and *spatial ability*:

Spatial ability is essential to livelihood, but spatial knowledge at the level of symbolic articulation in words and images is not. Many animals have spatial skills far exceeding those of man; birds that make transcontinental migrations are an outstanding example. ... Spatial

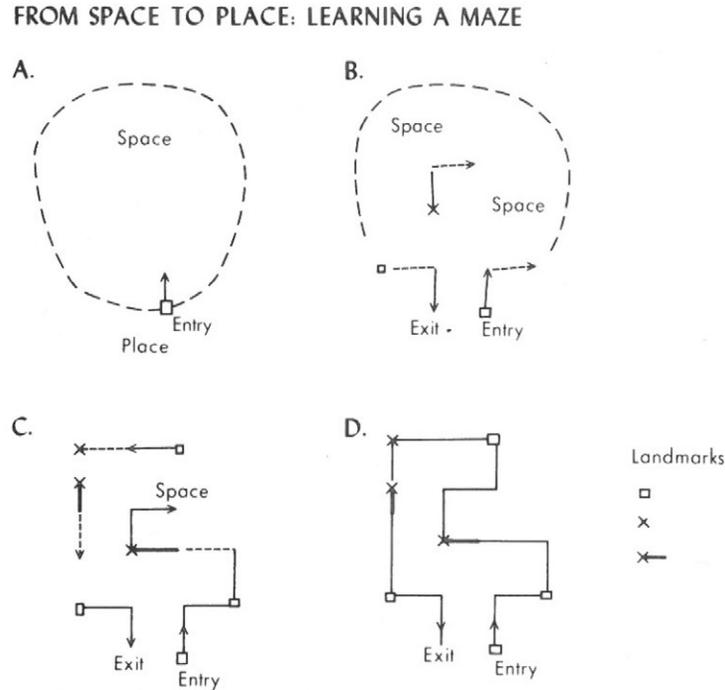


Figure 2-3: Learning a maze. (A) A subject enters a maze. She has no idea where she is. (B) She begins noticing spatial and visual landmarks during her exploration. (C) She gradually learns the route she takes with the help of landmarks. (D) She understands the maze completely, and she can find the exit without any mistakes.

ability precedes spatial knowledge. Mental worlds are refined out of sensory and kinesthetic experiences. Spatial knowledge enhances spatial ability.(Tuan, 1977)

We humans use our spatial abilities to move about and around our environments, without necessarily attending to our own actions and perceptions. Going back to Schön's argument, our spatial abilities are skilled practices whether when we play piano or walk down the street. Tuan suggests that whenever we reach beyond this automated behavior we use our spatial knowledge. One such piece of knowledge he suggests we use is the knowledge of landmarks that are learned as part of a journey. In Figure 2-3 he demonstrates how a person gradually learns a maze by integrating his movements from the entrance to the exit. This maze learning example is based on an earlier work by the psychologist Warner Brown(Brown, 1932). Initially, the subject does not have a sense of where the entrance and the exit are located in space.

With more trials, she begins learning distinct localities in her journey by identifying a perceptual quality such as a "rough spot" on her route or by making a structured decision, such as making a "double turn." These landmarks become handles for the subject, leading to a more and more confident understanding of the environment, until eventually she identifies the path between the entrance and the exit. In the end, the subject has a journey composed of discrete stages discovered through her explorations. This process, Tuan suggests, is how people convert an unknown space into a place. An interesting result of the maze learning experiment is that even those subjects who had learned to navigate the maze failed to understand the entire layout of the environment. When they are asked to draw the maze they explored, they often made mistakes regarding the lengths and angles of the paths, or omit multiple turns. This suggests that the subjects did not acquire a mental map but rather a practical knowledge that they can use when they are actually in the maze..

The maze learning study illustrates some of the most characteristic aspects of spatial experiences I introduced in the previous section: the aspect of *self-location*, that is, knowing and not knowing where we are; the aspect of *orientation*, which is based on qualitatively discriminating directions such as front and back, and left and right; the aspect of *action and movement*; the aspect of *material perception*, that is, the ability to locate objects and materials within an environment; the aspect of *temporality*; and finally the aspect of *discovery and learning*, the ability to distill our spatial abilities into spatial knowledge.

2.3 Computational Approaches to Understanding Visual Perception and Spatial Experience

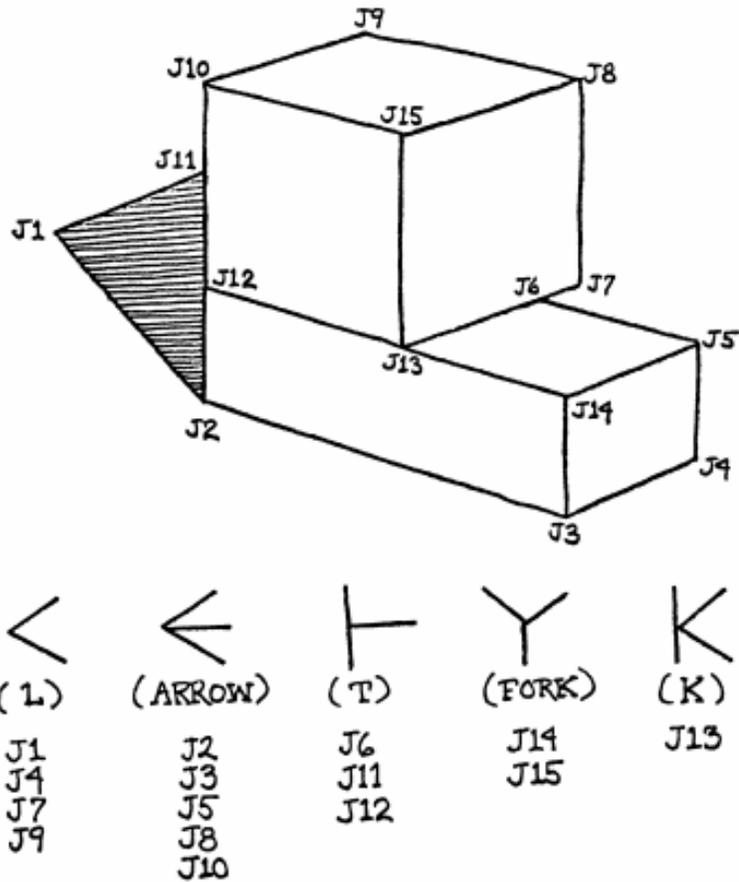
In the context of computation, a variety of models and tools have been proposed for the analysis of the perception and the experience of space. A particular understanding of perception as a form of information-processing has been employed in cognitive science and neuroscience, and in the development of artificial intelligence, which aims

at the mental reconstruction of space and the identification of the environment and objects within. The experience of space is mostly dismissed in this approach, except for a particular interest in place representation and navigation. By contrast, architects are mostly interested in the analysis of the qualities of design through which they evaluate space and material, and are concerned with using those qualities in design processes in order to provide a particular experience. There are parallels and conflicts between these two discourses, which I present in this section. First, I explore a variety of computational models that attack the problem of the perception of space, present underlying computational theories, and evaluate their relevance to design problems. Then, I investigate alternative approaches in the context of design and computation.

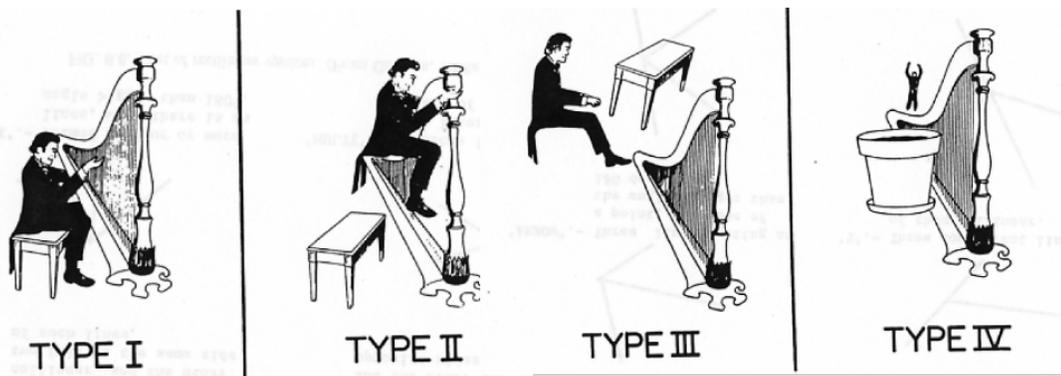
The computational models of the perception of space are presented under three categories. First, early geometrical models propose deterministic descriptions of three-dimensional objects, with an attempt to identify objects according to a certain geometrical constraint or criteria. Second, bottom-up models, or feedforward systems, aim at gradually building a higher-level description of the environment from basic visual features, and then apply learning models for identification. Finally, top-down models construct representations through a set of generative rules and compare these representations to the images. As a common motivation, all three of these models deal with the problem of identification and classification of a given environmental image or sequence of images. In the following section, I will discuss whether this motivation is appropriate for designers, and how it relates to design problems.

2.3.1 Geometrical Models

In geometrical models, space is examined in a purely structural sense in which perception is an analytic operation on the structure of the world. This approach is often associated with James Gibson's *ambient array*: he argues that the visual world has an inherent structure, which is exposed in the ambient array through perspective and vanishing points (Gibson, 2014). Geometrical models use this simple rule of perspective to determine the three-dimensional structure of objects by



(a) Geometrical models uses perspective to define structural rules. Each edge can follow one of the five rules described here. Waltz uses a constraint search to determine the three dimensional structure. David Waltz, 1972



(b) Irving Biederman's semantic rules. Type 1: a semantically correct scene. Type 2: a position violation Type 3: a support violation Type 4: a size violation. Biederman, 1981

Figure 2-4: Early attempts in AI research to solve the problem of perception focused on the development of rule-based, geometrical models.

identifying different types of edge junctions, as seen in Figure 2-4(a) (Guzmán, 1968; Winston, 1970; Waltz, 1970). David Waltz observes that each particular setting can be solved in terms of constraint propagation. Irving Biederman later augments the geometrical rules with a set of semantic ones. He suggests that objects themselves do not pose a three-dimensional structure and location, but only appear so when they are in a semantically correct scene (Biederman et al., 1982). He defines five relational violations in a scene: support, an object that is floating in air; interposition, a background that appears inside an object; probability, an object that is in an unusual scene; position, an object that is in an unusual location; and an object with an unusual size (Figure 2-4(a)). However elaborate and detailed, geometrical models failed to address real world scenes, as it was revealed over time that vision was not a simple task; thus, this approach was replaced by more complex bottom-up and top-down models.

2.3.2 Bottom-up Models and Deep Neural Networks

The bottom-up models have a range of theoretical and technical developments behind them. As I indicated in the first section of this chapter, empiricist theories of the 19th century, particularly that of Helmholtz, suggested that spatial qualities were unconsciously learned through seeing and moving, and that their particular associations were stored in memory (Kahl, 1878). Later, the discourse of artificial intelligence began with Alan Turing (Turing, 1950), and later John McCarthy, Marvin Minsky, and others introduced fundamental ideas for symbolic computation. Finally, discoveries in neuroscience generated an interest in neural networks. In particular, a better understanding of how the visual cortex and the hippocampus works motivated researchers to develop algorithms that imitate neural computing in the human brain (Hubel and Wiesel, 1959; O'Keefe and Nadel, 1978).

One of the earliest examples of the neural computational model of the perception of space is found in Marvin Minsky's *The Society of Mind* (Minsky, 1988). In the chapter "The Shape of Space," Minsky builds a theory of nearness: any point in space exists only in relation to another point; without such reference it is not possible to

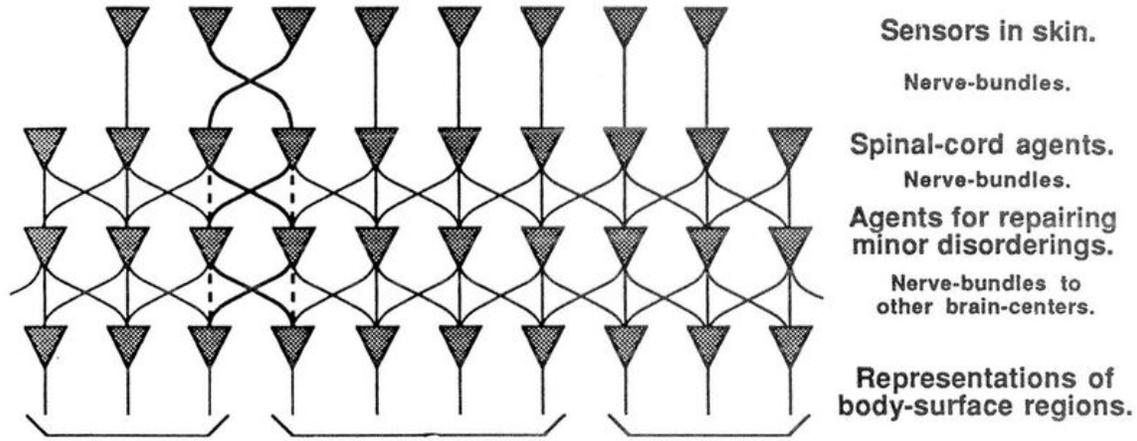


Figure 2-5: Society of Neighbors by Minsky Minsky (1988)

develop a concept of where for such a point. This is true not only for vision, but also for any other senses, including touch. Thus, space is a “vast society of nearness” (Figure 2-5). According to Minsky, a child starts by learning the space of the skin and extends that understanding to the outer world: “The nerve pathways that preserve the physical nearness relations of our skin-sensors, can make it easy for inner agencies to discover corresponding nearness about the outer world of space” (Minsky, 1988, p. 112).

In Minsky’s model, a dense network of neurons adjust themselves to allow the emergence of a spatial organization. By contrast, the neuroscientist David Marr proposes that the brain uses retinal representation that is already organized to reconstruct a three dimensional space through a computational process. In his seminal book, *Vision*, he defines a three-step process used by a machine to carry out a visual information-processing task: first construct a primal sketch to discover blobs, edge segments, boundaries, and other low level features; then make a 2.5D sketch to determine surface orientations and distances from the viewer; and finally construct a 3D representation through which to carry out higher level spatial operations (2010). Marr enumerates a list of procedures that are involved in the spatial reconstruction process, including binocular image integration, texture analysis, shading analysis, extracting shape and surface contours, all with reference to biological neural computations.

Marr’s computational theory played a key role in the establishment of computational neuroscience and inspired many subsequent studies (Poggio in Marr, 2010). Today, convolutional neural nets (CNN), which use Marr’s method of analysis, are widely employed in image processing, particularly for recognition tasks. CNN imitates basic procedures of the human visual field within which individual neurons with overlapping receptive fields—different spatial regions in image space—are densely connected. They produce rich, unique descriptors out of large data sets and can identify individual images with very low error rates (Fukushima, 1980). In their recent work, a group of researchers led by Antonio Torralba and Aude Oliva use this model for visual space perception in different contexts, such as the discrimination of scenes with regard to categories of indoor, outdoor, natural, and urban; the localization and recognition of objects in particular spaces; the discovery of visual identities of cities; and the evaluation of memorability of architectural and urban spaces (Oliva and Torralba, 2006; Zhou et al., 2014b,a; Isola et al., 2013). They also make use of the gist of a scene through what they call gist descriptors as well as the spatial envelope—a semantic description of the perceptual characteristics of a scene, such as openness, density, and verticality (Oliva and Torralba, 2006).

2.3.3 Top Down Models

Top-down models facilitate a nativist idea by assuming that perception is driven by top-down cognitive processes rather than being constructed out of sensory stimuli. For example, Rodney Brooks proposes a contextual, model-based method for object recognition, ACRONYM, in which three dimensional models are produced to match two dimensional images for recognition and measurement purposes (Brooks, 1981). In the context of indoor space reconstruction, Feng Han and Song-Chun Zhu use a similar method to match 2D images with three dimensional planar surfaces that represent walls, ceiling, and floor (Han and Zhu, 2008). Hierarchical Bayesian models (HBM) are also proposed as generative and predictive methods for accomplishing perceptual tasks, where a priori concept structures, or hypotheses, probabilistically

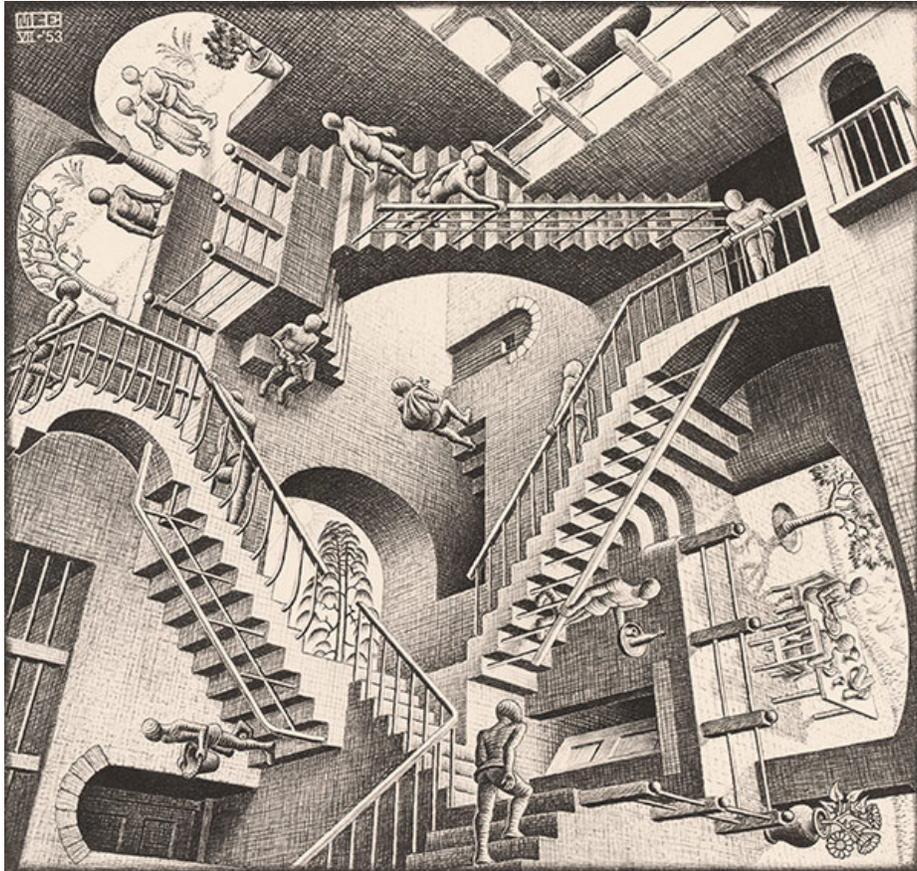
match an object to a category or make predictions about spatial locations (Tenenbaum et al., 2011). For example, in the context of visual scene recognition, this model has been employed to limit the expected locations of objects in images to particular zones in which they are most likely to occur (Belongie et al., 2007).

2.3.4 Formal Approaches in Design

The computational models presented above are mostly concerned with (1) solving the "ill-defined problem" of vision, which is a problem of reverse engineering of three-dimensional worlds from two dimensional images, and (2) classifying and identifying objects and places. However, from a designer's perspective, the qualities of space are limitless, and structure and identity can only be a subset of what is perceived and experienced in space. Moreover, neither structure nor identity are fixed; their interpretations are subjective. James calls this sagacity, the ability to discover parts in wholes, and remarks: "The properties which are important vary from man to man and from hour to hour" (James, 1890, p. 961). Artists and designers often make use of this dynamism of perception, which can be seen, for example, in Escher's *Relativity* or in Steiner's *Look Alikes* (Figure 2-6). How many grounds are there in *Relativity*? Are there stove tops, or spiders, or something else in *Look Alikes*?

On the other hand, the aforementioned computer vision models provide powerful and reliable methods for measurement and quantification of visual data. Particularly in digital modes of design, accurate three-dimensional reconstruction of a space provides a rich representation of the environment that can be digitally explored from a variety of viewpoints (Galor et al., 2009). This method also is capable of making metric assessments that are crucial for architectural production phases. Quantification of the perceptual characteristics in architectural space has become of particular interest to designers. Michael Benedikt points out the neglect of perception studies regarding space Benedikt (1979). He quotes William Ittelson: "the overwhelming bulk of perception research has been carried out in the context of object perception rather than environment perception" (Ittelson, 1978).

Benedikt proposes that each environment can be characterized by its *isovist*,



(a) *Relativity* by Escher



(b) *Look Alikes* by Steiner

Figure 2-6: Dynamism in visual perception is often used by artists to create illusions.

which is the sum of all vantage points in that environment. According to him, an isovist may essentially help describe the behavior of people in a space as well as explain the experiences of people regarding the physical qualities in the environment. The idea of an isovist was later integrated into the theory of space syntax by Bill Hillier (Hillier, 1999). Space syntax provides a series of tools and measurements with which to quantify spatial properties, mostly regarding accessibility and complexity. This method is often employed in urban settings to make qualitative measurements of buildings and street networks. In the context of computation for design, the perception and experience of space is considered not only to be an aspect of the final design but also an important part of the design process. Dönald Schön points out the perceptual and dynamic character of designing (Schön, 1988):

In this paper, I shall treat designing not primarily as a form of "problem solving," "information processing," or "sketch," but as a kind of making. On this view, design knowledge and reasoning are expressed in designers' transactions with materials, artifacts made, **conditions under which they are made**, and manner of making. (emphasis added)

For Schön, a designer's knowledge is beyond her or his abstract thinking, and is shaped within the design world she or he inhabits. This indicates the necessity of plurality in the designer's experience of space, which should allow a designer to see things or touch things in space in multiple ways.

Shape grammars offer a formal method of visual computing in which shapes preserve their ambiguity without being converted to symbols (Stiny and Gips, 1971). As rule-based systems, shape grammars make it possible to generate and analyze designs with shapes, which are embedded by designers interactively. Embedding is the sagacity of the designer, each time a new thing can be seen and integrated to design rules (Stiny, 2006). Shape grammars have been applied to design problems as descriptive tools (Stiny and Mitchell, 1980; Knight, 1989), as generative tools (Duarte, 2005) and as educational tools (Knight, 1999; Özkar, 2011). This approach

does not presuppose a fixed environment within which to search for good designs (Simon, 1969), but instead makes it possible to dynamically generate the design. In their recent work, Terry Knight and George Stiny describe designing with shape grammars as “doing (drawing) and seeing with basic spatial elements that make shapes” (Knight and Stiny, 2015). In this work, the formalism of shape grammars is carried forward to materials and making activities, such as knotting with strings.

In the context of the experience of space, Ferreira, et. al. propose an implementation of shape grammars to describe the relationship between body, motion, and space (Ferreira et al., 2011). They explore the possibility of including motion and temporality in formal descriptions, which are often neglected in works that approach this problem. They propose a movement grammar for the human body, with the aim of describing architectural space through motion. Their approach, which they call “corporeal view,” has affinities with the theory of practice (Bourdieu, 1977; Lave, 1988). In *Cognition in Practice*, Jean Lave argues that to understand cognition, one should take whole person in action and in the context of activity. Mind is not “in the head” but in the practice (Lave, 1988).

2.4 Discussion

In this chapter, I presented a review of the literature on spatial perception and experience, and computational approaches in this field. I discussed how scholars dealt with puzzling questions of how and why we perceive space, from ancient theories to more contemporary understandings of experienced space. One important idea is that each species has a unique experience of the world determined by its particular goals and abilities. Humans, as well, perceive and understand their environment based on their goals such as finding shelter and navigating between places.

Our subjective spatial experiences in the environment has also been a subject of interest for architects and designers throughout history. I presented how designers adopted and tested theories of perception in their work, and invented ways of

representing the subjective experience of architecture. Among the computational approaches in design, shape grammars and making grammars provide a formal method of computing with shapes and materials, taking into account the movements and interactions of the designer. In the context of spatial experience, making grammars provide important insights on how to connect symbolic descriptions with perceptual and active information.

Another important insight in this chapter is that thinking and perceiving are not separate but are parts of the same process. Often defined as embodied and situated cognition, this idea suggests that our symbolic abilities and perceptual abilities are connected to each other. Rudolf Arnheim makes this observation when he studies the idea of visual thinking —the type of thinking designers and artists engage when they create. I showed that Arnheim’s idea is also supported by the studies in cognitive and neuroscience. In an interview, the neuroscientist Matt Wilson suggests that we create an internal narrative from our experience (O’Connor, 2019, p. 172).

Both [navigation and memory] depend on a critical function, linking things in time. It is how you put the pieces together, how you create an internal narrative of your experience. It is not a record, or videotape of experience. It involves evaluating, selecting, and sorting things. Rats create an experience of moving around in space. We create the stories of our lives.

I agree with Wilson that we create the stories of our lives through our spatial experiences. What are those stories and how do we construct them? In the following chapter, I will search for the answers to this question through two case studies that investigate the relationships between inner stories and human spatial experience.

Chapter 3

Creating and Representing Spatial Experiences

In the previous chapter I introduced key ideas and studies on human spatial experience. I showed that, beginning in the nineteenth century, scholars began to distinguish the differences between *perception of space* and *spatial experience*. Perception of space is a fundamental aspect of all animals, which enables them to translate sensory signals into a three-dimensional representation. Perception of space enables an animal to navigate the world and find shelter and food. In contrast, spatial experience is unique to each species, and is defined by the particular goals, abilities and constraints of a specific animal. Spatial experience is the particular way in which an animal senses the world, then filters and composes the available information through an ongoing interaction with its environment. Similarly, human spatial experience is the unique way we humans perceive our world and make available particular types of environmental information through which we can take action: surfaces on which we can walk or enclosures we consider safe to enter. Our spatial experiences produce an understanding of the environment in a way we can mentally examine and communicate to others. Within the scope of this thesis, I consider human spatial experience as an inner story that we tell ourselves, one which enables us to understand and interact with our environment. In which ways do we filter the information in our environment and construct inner stories?

How does an inner story enable spatial experience?

In this chapter, you will learn about my investigations into human spatial experience undertaken in two case studies. These projects will further clarify the distinct features of spatial experiences and show the inherent relationships between inner stories and spatial experience. The first case study, “September 1955” is a historical virtual reality documentary. Through this project, I examined the ways in which people understand a story experienced in a fictional virtual space. The second case study is of a class entitled “Computational Ethnography and Spatial Narratives”, which I co-taught with Nil Tuzcu at the Boston Architectural College, in Spring 2016. In this class, we explored visual and material representations of spatial experiences in the particular context of urban exploration.

3.1 September 1955: Immersing into the history

“September 1955” is a 7-minute virtual-reality documentary of the Istanbul Pogrom, an organized, government-initiated attack on the ethnic minorities of Istanbul on September 6-7, 1955. In this interactive installation, viewers move through a soon-to-be-attacked photography studio recreated from 1955. They experience the events from various points of view, including that of onlookers outside in the street and that of the owner of the photography studio—one of the victims—inside the building. By encouraging the viewer to move around the virtual space and participate in the mundane activities within the studio, this piece aims to make the viewer a part of the story and generates individualized and embodied understanding of a historical event.

The Istanbul Pogrom, during which thousands of homes, businesses, and churches that belonged to Greek and Armenian minorities were raided by a mob, marked a turning point in the social history of Turkey. Frequently referenced as one of the most shameful events in the country’s history, the pogrom exemplifies how polarization within a society and constant provocation can result in the dissolution of otherwise peacefully co-existing communities. In this way, it also teaches important lessons for today’s world.

Motivated by the idea that the inner story apparatus is the main driver of spatial experience, I and my co-creators Deniz Tortum and Nil Tuzcu created a virtual-reality documentary of the Istanbul Pogrom. We explored how we could use space as a tool to create a new narrative about an event that had devastating effects in the past. We wanted to explore how empathy could be created through embodied interaction and immersive perception of the events that occurred during the pogrom. The viewers wore virtual reality headsets that guided them through several spaces in which the realities of the time were recreated. The installation was designed to allow participants to move through these spaces, examine objects from the period, listen to the chatter of the people, and witness the events as they unfolded —both from the points of view of the onlookers outside and the victim inside the portrait studio. Because the viewers found themselves within a "real" space, they became more engaged in the events as they unfolded and had the opportunity to form an emotional connection to them. In the following sections, I examine this piece from various perspectives to show how the story unfolds in space, whether we take on the role of an onlooker or that of a victim. Figure 3-1 shows a person viewing September 1955 in the Istanbul Independent Films Festival in February, 2017.

3.1.1 Project Setup

“September 1955” takes place in a photography studio in the Beyoglu district of Istanbul. Because of its majority of Greek inhabitants and business owners, Beyoglu was one of the epicenters of the Istanbul Pogrom. The decision to build the story around a portrait studio was made after we discovered the archives of photographers Osep Minasoglu and Maryam Sahinyan, both of whom were members of ethnic minorities (Greek and Armenian, respectively) who witnessed the events. Minasoglu moved to Paris soon after the events while Sahinyan stayed in Beyoglu to continue her work. Sahinyan’s work resulted in a large collection of photographs that focused on the socio-cultural transformation of the country, particularly with regard to the socially segregated communities of rural migrants, the trans-gender community, and religious minorities. Inspired by the lives and the works of these photographers, we



Figure 3-1: September 1955 VR installation at the Istanbul Independent Film Festival, 2017. (Photo Credit: Deniz Tortum)

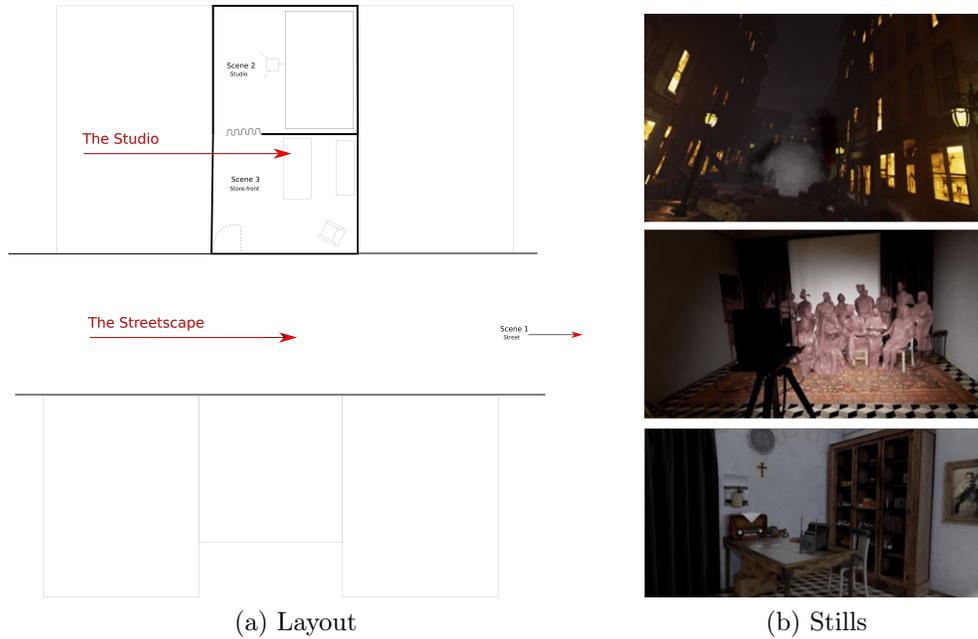


Figure 3-2: : The story takes place in two main scenes, on the street and in a portrait studio. (a): Layout of the virtual spaces in September 1955. (b): Stills from the VR documentary, from top to bottom: streetscape, the studio, and the waiting area at the front of the store.

decided to digitally reconstruct (and partially re-imagine) a portrait studio of the period, which was raided during the events.

The story unfolds in a digital space that corresponds to a physical space of 5x5 meters. In this virtual documentary, which is experienced with virtual reality (VR) glasses with 6 degrees of freedom, the viewer physically walks within the digital space and is able, for example, to look under furniture or closely investigate objects . The viewer thus feels completely immersed in the digital space.

We followed a series of technical steps in order to create the experience. These steps included creating digital reconstructions, sound and light editing, creating and adding character animations, and creating spatial transitions.

3.1.2 Streetscape and The Studio

The digital space consists of two main scenes: the streetscape, in which we witness the aftermath of the events, and the studio, in which we take a look at the everyday life a

photographer in the 1950s. To collect information on how a photography studio would look at that time, we consulted the photographic archive of Sahinyan. Architectural details such as floor tiles, papier-mache architectural ornaments, doors, and everyday objects such as chairs, tables and mirrors, were digitally reconstructed based on those photographs. The studio space consisted of two rooms separated by a black curtain. In the front, there is a waiting area, and in the back, there is a photography studio. The front of the store faced the aforementioned street-scape, which is designed to resemble Istiklal Street, the commercial avenue at the heart of Beyoglu, or one of its adjacent streets. The layout and stills of the studio and the streetscape are illustrated in Figure 3-2.

3.1.3 Story Elements

We introduced various story elements that would gradually pull viewers into the context without disturbing their individual spatial experience. These elements were: (1) *objects of interest*, such as an old radio, bottles of chemicals involved in photography, a newspaper of the day, and a wall covered by Sahinyan's photography; (2) a *spatial soundscape*, which included a song playing on the radio, a ringing phone, and the indistinct chatter of people; and (3) *animated characters*, both as shadowy figures at a distance who represented the mob, and as clients within the portrait studio. These elements drove the viewers' engagement with the virtual space and helped create an immersive spatial experience, which became increasingly grim toward the end of the documentary experience.

One of the sub-goals of the project was to present archival materials in a spatial medium, which we believed would improve viewers' engagement with the historical subject. In our fiction, objects of interest served a dual purpose. First, they captured the viewers' attention, inviting them to take a closer look at the scene. People often moved around in the virtual space in order to get closer to these objects, gathering new perspectives and building their own understanding of the environment. For example, a viewer bent under the table to read what was written on a stack of boxes underneath, while another looked up to the ceiling to study the oriental patterns of



Figure 3-3: Some of the *objects of interest* in September 1955 included an old radio, some lacework, and the calendar of the day.

the papier-mache. Second, the objects of interest provided a contextually relevant representation of archival materials, such as the newspaper of the day found on a coffee table, or a lace-work doily covering the old radio —typical aesthetic choices of the time (Figure 3-4). Because these objects were part of the *everydayness* of the studio, they helped create the sensation of actually being in the 1955 time-frame.

Another equally important story element in September 1955 was the *spatial soundscape*. Visual immersion provided by the VR headset was complemented by an audio component, which was created by spatially located sound sources within the digital space. Spatially located sound sources allowed the viewers to perceive the sounds as if coming from various distances and directions, increasing the sense of reality. For example, a song playing on the radio in the other room sounded as if it actually came from behind the wall. People could locate the radio by following the song to its source. The additional element provided by the soundscape further immersed viewers in the story, "breaking our grasp in the real world," as some of them commented.

The immersive aspect of the story within a fictional space was empowered by the



Figure 3-4: Animated characters enabled viewers to connect to the lived experiences of those who were involved in the life of the studio. 3-D character animations morphed into photographs of real people from Maryam Sahinyan’s archive, reminding the viewer that these ghostly animations depicted real people. The perpetrators gathered around the shop also depicted real people.

visual and aural story elements described in the previous paragraphs. In order to tell the stories of those who participated in the everyday activities within the studio, and bridging those lived experiences to the horrific events of the pogrom, we introduced another story element, *animated characters*. These animated characters, created by a special scanning technique, provided an opportunity to the viewer to glimpse the everyday life of the studio’s customers, most of whom were Greek or Armenians. The viewer watched as shadowy animated figures prepared for a portrait shot, and listened to their mundane chatter. A scene in which two women appeared to be preparing for a sitting transitioned into one in which the flash of a camera filled the space, suddenly revealing a real photo of two sisters from Sahinyan’s archive. In that split second, participants understood that the documentary was depicting the experiences of real people, people who had been customers at a studio in 1955.

The viewers had approximately seven minutes to experience the story unfold and to engage with the elements we introduced. Through their individual spatial experiences, their decisions to pay attention to a certain detail resulted in their unique understanding of the experience. The only time the viewer had no agency in driving the narrative was the moment when the story transitioned from the streetscape to inside the portrait studio. Whereas in the beginning of the experience

the viewers find themselves as onlookers outside the studio, the transition places them within the studio earlier in the day, where they find themselves with the victims of the upcoming events.

3.1.4 Narratives

Narrative of the Spectator

The story of “September 1955” begins on the night of September 6th. The viewer finds herself in the middle of a street looking at a group of looters. She is at a safe distance which we refer to as *the narrative of the onlooker*, which depicts the scene as it was captured in the photographs at the time. By placing the viewer at a distance from the events, we isolate her from it, in much the same way as when we show someone a photograph: we, the onlookers, have no agency in changing the course of an event that has already taken place.

While the viewer examines this brief introduction, the scene transitions to an earlier time inside the studio, this time placing her in the middle of the events that are about to occur.

Narrative of the Victim

In the next scene, the viewer finds herself inside the studio, which is soon to be attacked. It takes some time for the viewer to realize she has traveled back in time, and the events she observed as an onlooker in the previous scene are about to happen. As the event unfolds, she finds herself surrounded, trapped inside a studio. This is *the narrative of the victim*. Here, the devastation manifests itself not as a spectacle to watch, but through the intensified sensory inputs of the crashing, screaming, mob gathering outside. These narrative and the previous one complete the piece.

The viewer’s ability to move around while participating in the installation gives a certain agency to him or her, which in turn enhances the effect of these narratives. Both the presence of agency —the ability to move around— increase the viewer’s participation within the story, and when this agency is curtailed by the mob that



Figure 3-5: A viewer watching September 1955 in Keller Gallery

gathers around, the viewer feels trapped. In her 2018 review of the piece, the historian Angela Andersen makes a similar remark about her agency and entrapment within the studio (Andersen, 2018):

I strained my eyes to read the headlines on a newspaper, and looked at personal objects in the room. I turned around to see who was smashing the windows of a burning apartment building, and made my way to a door that seemed as if it might provide an escape route from a horde intent on destruction.

3.1.5 Qualitative Analysis of the Viewer’s Experience

“September 1955” was exhibited in different venues and festivals, including at the MIT Keller Gallery in November, 2016 and !F Istanbul Independent Films Festival in February 2017. We had a mix of audiences —some were familiar with the subject of the Istanbul Pogrom and some were not. We collected various information regarding the viewer’s experience through informal interviews, screen recordings of their experiences, and a diary in which people recorded their thoughts after the

screening. An overwhelming majority of the viewers remarked on the emotional connection made with the subject of the event —ethnic discrimination and mob action— even those who were not familiar with the subject. Screen recordings revealed the diverse ways in which people explored the virtual space during the documentary, which supported our initial thought that the spatial medium would allow more subjective experiences. With regard to the overall hypothesis of this dissertation, that spatial experience is a type of inner story driven by action and perception, two important insights were made: that active participation is a natural constituent of being-in-space, and that selective attention gives rise to multiple narratives.

Active participation is a natural constituent of being-in-space

The story that was told in “September 1955” provided a subtle background for the individual spatial experiences of the viewers. Instead of providing a complete historical narrative of the pogrom, we put viewers into the environment where the events took place. Although each viewer experienced the piece differently, all were subject to the same sequence of elements, starting out on a street at night where they heard the sounds of broken glasses and sirens, then moving into the photography studio. There they heard the chatter of people, and moved around the studio to take a look at different objects. Some discovered that they could pass through objects, stand inside a table or look through a wall. Some stood still for the most part, without paying attention to the details of the space. Nevertheless, at the end of seven minutes, everyone was left with a story that they could remember and talk about.

The way in which the visual and aural story elements captured the sensory channels of the viewers made it difficult for them to break out of their immersion in the space of “September 1955.” The majority of viewers we interviewed after the screening remarked on the realism of the virtual space, which they reported truly made them feel they were experiencing the events themselves. Affected by the anxiety evoked at the very end of the documentary, some users decided to remove

their headsets immediately in order to escape. In their diary notes, some viewers used the word "empathy," and wrote that they "had a chance to form empathy with those", and "experienced their horror and pain". If we examine the viewers' experiences in terms of Ingold's dwelling perspective(see pp. 54), we can conclude that by virtue of being immersed in the virtual space the viewer finds herself as an integral constituent of it; he or she is "borne wholly into this new spectacle" (Merleau-Ponty, 2013). Spatial experiences, including those that unfold in virtual environments, naturally demand our active participation, even when we think we are not paying attention.

Selective attention gives rise to multiple narratives

In any spatial experience, we pay attention to certain details while knowingly or unknowingly ignoring others. Our perception has a bandwidth. Our attention—visual and otherwise—is drawn to what we consider the most relevant or interesting given the task in hand. In "September 1955," selective attention plays an important role because there are many details that are not possible to capture in a single screening of seven minutes. The 1950s photographs that decorate the walls of the studio, taken by well-known Armenian artists, treat themes ranging from immigration to gender to personal identity. Not all of these can be seen at the same time, and some viewers fail to pay attention to them altogether. Some may focus on the other details instead, such as the tools and photography equipment in the studio, or on the postcards pinned to the wall or the newspaper folded on the table. Each person decides what he or she is going to spend time on while the outside events unfold. We thus end up with infinitely many different versions of the same story, each recreated by a different user.

3.1.6 Discussion

In this project, I explored the characteristics of spatial experiences of people exploring a fictional world of 1950's Istanbul. The use of virtual reality for creating stories is

a relatively new area. With this project I explored various representation techniques and introduced a series of story elements for creating stories in digital space. These elements encouraged viewers to actively explore an environment and increased the impact of the underlying story. The project also presented an opportunity to explore the underlying thesis of this dissertation: the relationships between spatial experiences and inner stories. After doing a qualitative analysis of the viewers' experience, I provide two answers to this question: (1) being in space naturally demands our participation with the environment —our cognitive state cannot be isolated from the environment we are in; and (2) selective attention leads to individual experiences. The mechanisms that select and filter information in the environment determine the type of experience we have, and each individual has a slightly different understanding of the environment. The field of spatial story telling offers many opportunities to understand how cognitive, perceptual and active information builds these individual understandings.

In the next case study, I will focus on constructing visual and material stories from observations.

3.2 Computational Ethnography and Spatial Narratives of Urban Space

“Computational Ethnography and Spatial Narratives of Urban Space” was a studio I co-instructed with Nil Tuzcu at the Boston Architectural College in Spring 2016. In this studio, we focused on methods for making observations in the urban environment and representing those observations through drawings, images, and models. We were interested in building on previous studies on urban perception and mental maps, as in the works of Kevin Lynch and William Whyte. We structured the class around two main themes: representations of personal narratives and representations of environmental features. Our major goal as instructors/experimenters was to see how much of the students' personal stories



Figure 3-6: Exercise 1: Draw your experiences today up to the time you arrived in the classroom.

included features of the environments and vice versa, our working hypothesis being that there must many overlaps between the two. With this goal in our mind, we created three exercises:

3.2.1 Exercise 1: Draw Your Day

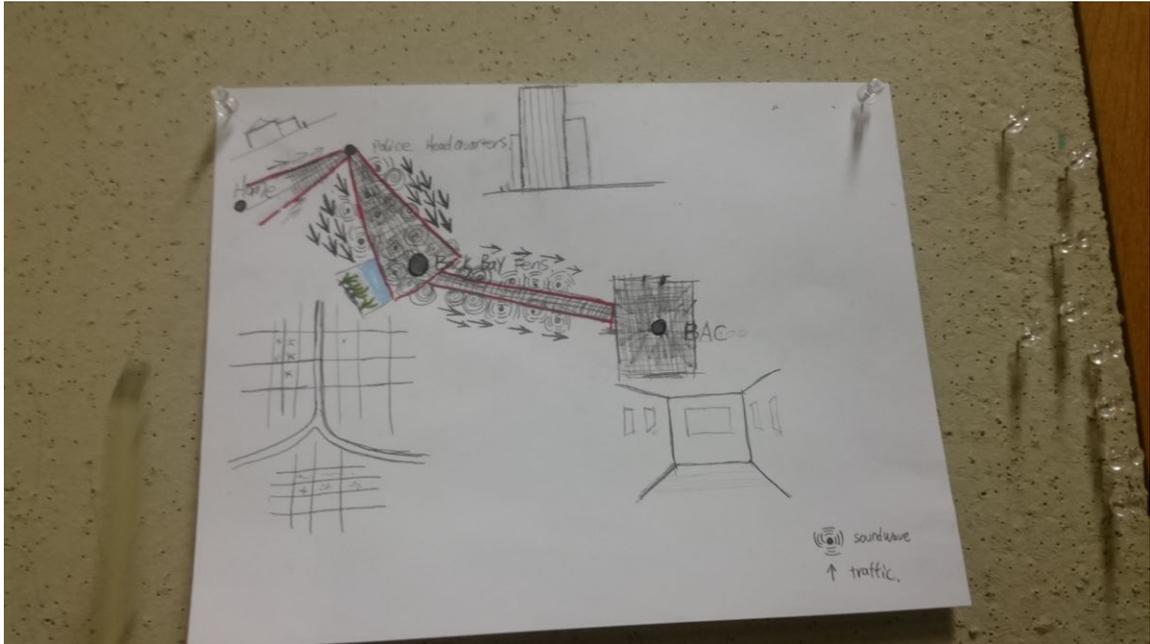
In this first exercise, students were asked to produce drawings and diagrams to tell the story of their day. They were instructed to trace their activities from the morning to the time they arrived to the classroom, and construct a visual story to represent their experience.

The resulting drawings included many interesting features, illustrating what the students considered important parts of their experiences. This exercise also demonstrated the challenges posed by reducing a spatio-temporal experience into a fixed drawing. Unlike the straightforward process of drawing a map, drawing an experience includes not only places and their spatial organization, but also the drawer's actions, thoughts, and feelings. Figure 3-6 illustrates one of the resulting drawings.

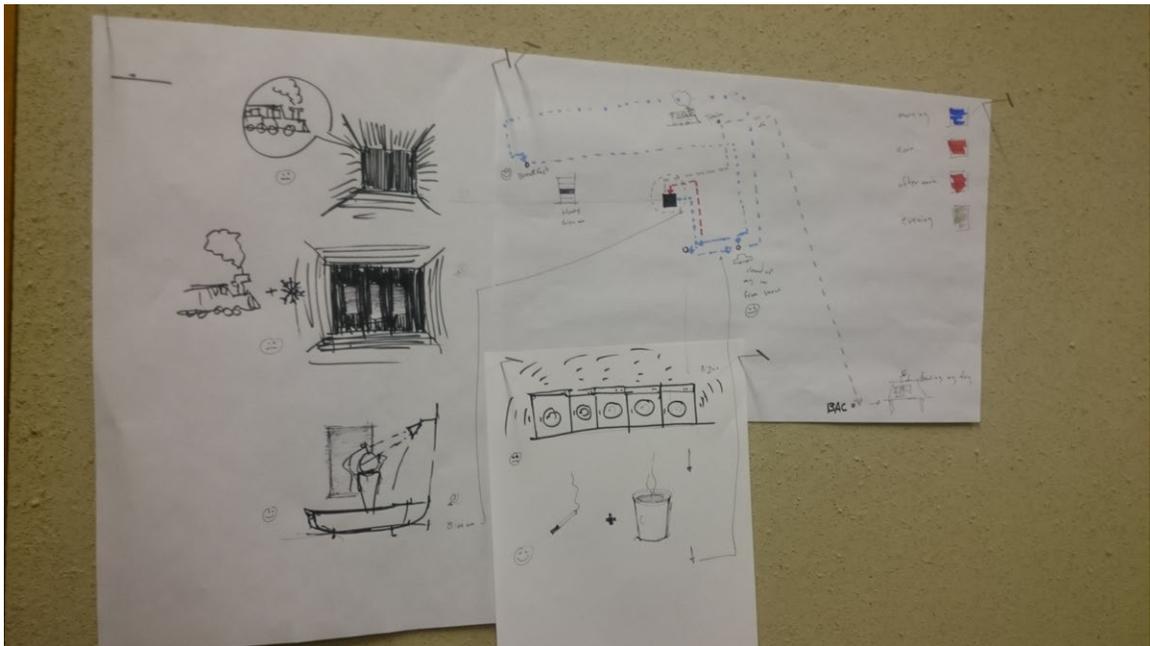
Each student also provided a verbal narrative to accompany his or her drawing.

The student who produced the example in Figure 3-6, explained that the horizontal axis represented time, but the duration of events was not linearly mapped to the timeline. Instead, as the student explained, the size of objects represented in the image corresponded to how much impact each had on the student's memory. In other words, while the order of drawing elements from left to right was chronological, the events themselves were represented in size based on their importance. A large computer screen was drawn from the perspective of the student, and invited us to look at the most important experience of her day: sitting in front of a computer screen and drawing. Interestingly, the computer screen was centered on the frame, making it easier for the observer to center it while examining the rest of the day. Every other visual element in her drawing was visually anchored to the drawing of the computer screen. She elaborated on the fact that other events happened during her time in the office, such as attending a meeting or talking to a colleague, but they all seemed to be smaller parts of her experience in front of her computer. Surrounding the central visual element in her drawing—the computer screen—were a series of steering wheels spread across the drawing. Like the computer monitor, the size of the steering wheel reflected her time commuting. She explained that the places she passed through did not matter as much as her experience within her car, so she decided to represent her journey from home to the office and from office to the classroom with the steering wheel. Time itself was represented as an observation: if the morning alarm rang or lunch time arrived, she noticed it.

It is often a challenge to encourage design students to think about the environment in terms of their own experiences instead of using conventional methods of drawing maps and orthogonal representations. In this particular exercise, some of the students decided to represent their day on a map, pointing out where they were at a particular time and how they navigated to another place. (Figure 3-7). The contrast between the dwelling and building perspectives were again very clear in these two different approaches. Exercise 2, below, aimed at encouraging students to focus on their own experiences within the city, and to construct narratives based on their observations.



(a)



(b)

Figure 3-7: Some students described their day on a map in addition to various drawing elements that describe different events that took place during the day.

3.2.2 Exercise 2: Redefining the Urban Elements

Students were introduced to Kevin Lynch's *The Image of the City* (Lynch, 1960), and William H. Whyte's *The Social Life of Small Urban Space* (Whyte, 1980). In their books, each urbanist defined city elements using tools such as mapping, direct observation and documentation. Together with the students, we analyzed and discussed these tools. Later, students were asked to develop their own tools of observation and create visual representations of their observations. Students were limited to selecting one urban area.

Results

Each student came up with a different observation technique and method of visual representation. For example, one student decided to adopt photography as a tool of observation, and then overlaid drawings to photographs as a visualization method. In the first step, he followed Walter Benjamin's definition of *Flaneur*; he took multiple walks around the neighborhood and took many photos of the areas/elements that caught his interest. In the second step, he made sketches on these photos to highlight the urban elements he discovered. The student used various drawing styles to highlight different elements. The drawing/highlighting changed in terms of line thickness (bold vs light), line style (smooth vs rigid) or drawing style (free hand drawing vs more regularized drawing) This change in drawing style was his unique way of representing what he observed in the city. The following are the visual elements the student used in the process:

Visual Representation Elements:

- Hatching: Identify various surfaces, such as a surface of a building or a green patch between the street lanes
- Dotted Lines: Identify a material aspect that he deemed important, such as piece of stone or concrete.
- Framing: Identify boundaries and surfaces.

- Arrows: Identify directions and paths
- Spirals and Stars: Identify sound and lighting, respectively.

Using these elements, he represented five different urban elements he discovered: paths, landmarks, materials, sound and light. Figure 3-8 shows some of the drawings the student produced.

3.2.3 Exercise 3: Physical Models of Observation

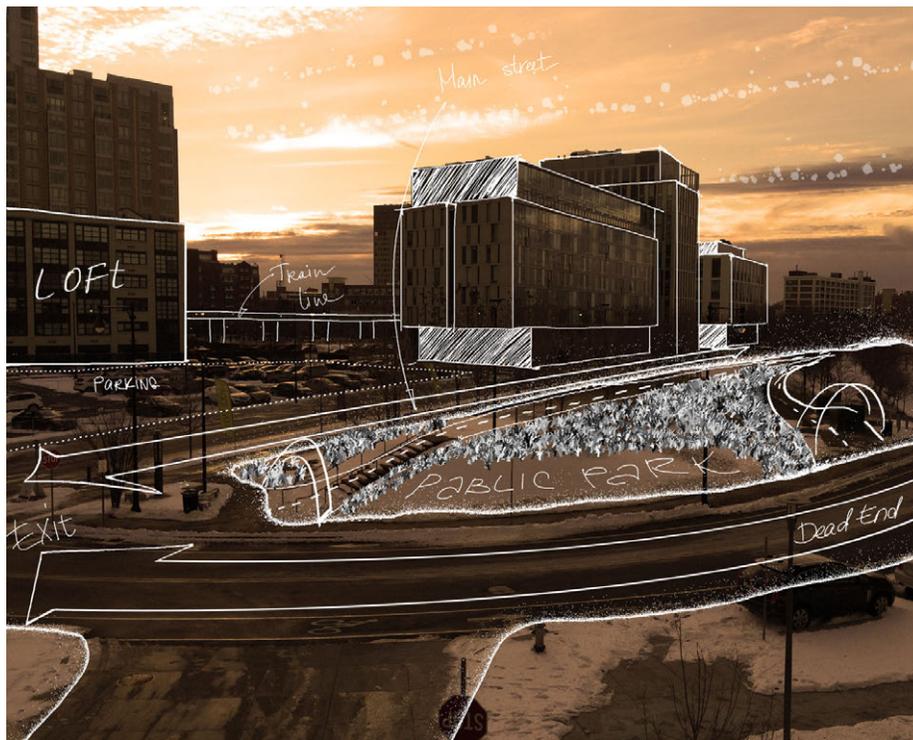
In the final exercise, students were asked to make physical models based on the observations made in the second exercise. They integrated their visual representations and the urban elements that they discovered into a material representation. The purpose of this exercise was to understand the differences between the inner stories conveyed in physical representations and those conveyed in visual representations. The students were instructed to focus on one of the urban elements they had discovered and represent the environment based on that element. In this exercise, each student used a different method in representing their observations with physical materials. They also created digital models from which they constructed their physical models. Below I present two examples of student work, showing their different models and the corresponding urban element they each identified.

Results

Example 1 - Representing the public activity The student represented the public activity within the environment she observed (Figure 3-9). She focused on her selected field in East Boston. She created two models that have different characteristics. Three different colors in the first model (a) represents three different activity types she observed: recreational activity, which occurs near the shore where people enjoy the view across the river; commercial activity, which occurs around various shops and stores; and navigational activity, which occurs around the road intersections and pedestrian crossings near the highway. She used curved stripes of

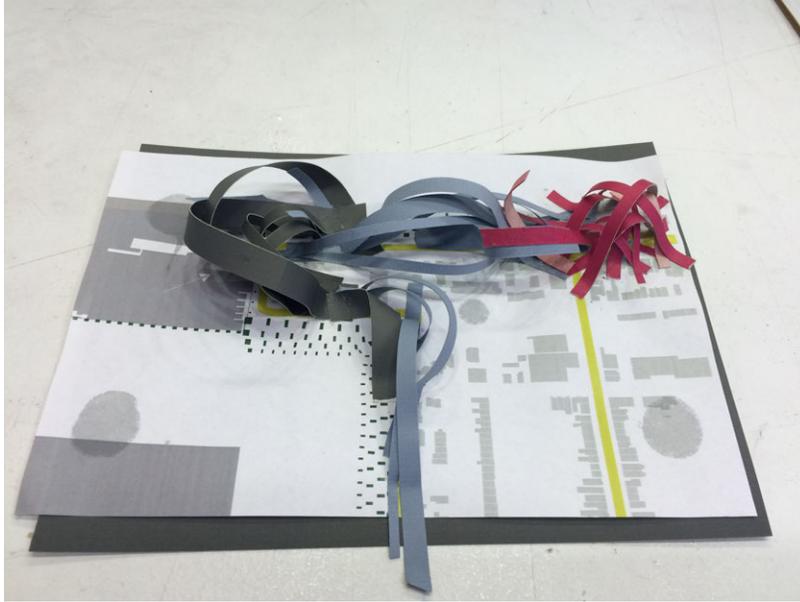


(a)

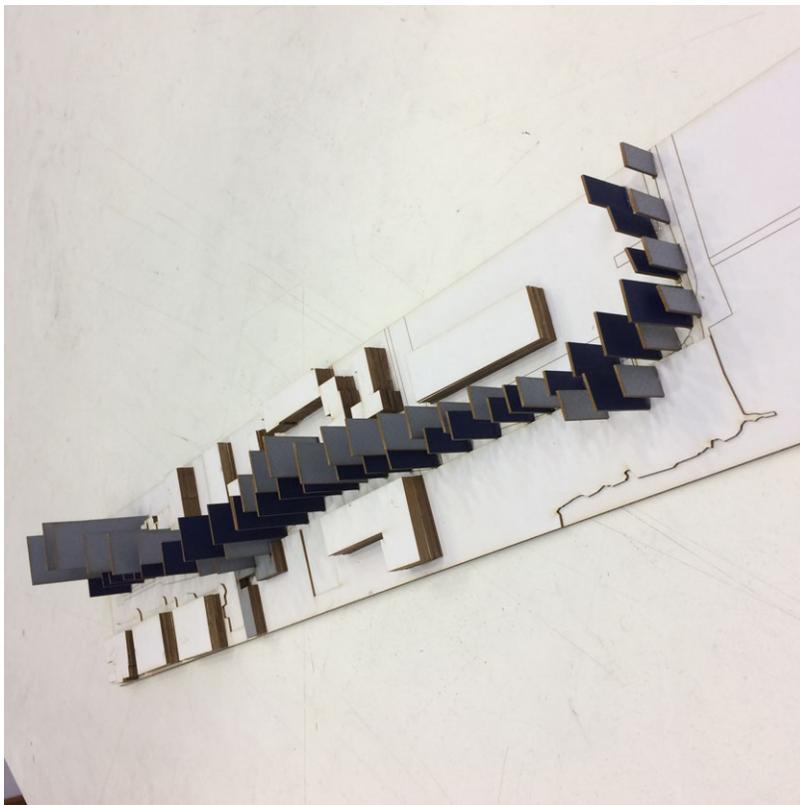


(b)

Figure 3-8: Redefining urban elements. Drawing is used as the primary tool analyze urban elements in the city. This example shows the use of lines, arrows and text for representing how the observer experienced this streetscape.



(a)



(b)

Figure 3-9: Physical model representing public activity in East Boston. The different colors in (a) represents three different types of public activity: recreational, commercial and navigational. In (b) the types of represented activities are reduced to two: navigational and non-navigational. The heights of the rectangular pieces represent the amount of activity in a particular location.

cardboard to represent those activities. In the second model, she reduced these activities to two: navigational and non-navigational. In this model, she quantified the amount of each activity and represented these amounts using rectangular elements with different heights (b).

Example 2 - Representing the level of noise in different times In this example, the student focused on the noise element he identified in his field. He also created an interactive representation of the sound sources that created the noise. The amount of noise was represented by transparent acrylic pieces that were cut into curved shapes, the height of which represented the level of noise at a particular location. The student engraved names of the noise sources on each piece. According to him, the main reason for the noise was the highways surrounding the field, which had different impacts at different times of day. In order to represent the varying levels of noise during the day, he installed strips of LED lights on the roads in his model: the colors of the lights represented the time of the day, and the intensity of light represented the amount of traffic. Figure 3-10 shows a few images of the model.

3.2.4 Discussion

In this case study, I investigated how students represent their experiences using a variety of media, and construct visual and material qualities. This study revealed the diverse ways we construct inner stories, from establishing temporal relationships between observations to isolating and emphasizing a distinct material quality. The design process guided by story-making encouraged students to dwell on their subjective experiences, and discover ways to analyze and represent their spatial experiences in visual and material media.

In each exercise we focused on a particular aspect of spatial experience. In the first exercise students constructed a visual narrative from their memories to describe their day. Their challenge was to decide on what to represent and how to combine various elements in their representations. We observed two distinct approaches:



(a)



(b)

Figure 3-10: In this example, the student made an interactive model that shows the level of noise at different times of a day.

taking personal perspectives and making maps. Personal perspectives enabled students to express their subjective experiences and emphasize the importance and impact of events from their own points of view. The map representations enabled students to construct a spatial representation, on which they could identify the different locations of events.

In the second exercise, students were asked to develop their own observation and representation methods and identify various urban elements they discovered. The results of this exercise pointed to the diversity of elements one can discover, and demonstrated how the various senses contribute to this discovery process. From visual and material qualities to lighting conditions and noise, many elements make up one's spatial experience. The types of representations the students picked for representing various urban elements are parallel to the perceptual and active information they gathered during their observations: arrows indicate walking directions, and surface hatches indicate horizontal and vertical boundaries. A further study can be developed to introduce constraints to drawing techniques to further understand the relationships between the spatial experiences and how they are represented.

The final exercise enabled students to increase the complexity of their representations using different materials and interactive elements. The main difference between the second and third exercise is that the physical models allowed new ways of adding spatial and temporal dimensions to the representations. The three dimensional volume created by the strips of cardboard in the Example 1 and the interactive lighting that represent the level of traffic in Example 2 are good examples of the added advantage that physical models provide.

In summary, this case study enabled me to explore the idea of constructing inner stories using a variety of media, and to observe how one converts his or her perceptual and active experience into a visual or material story.

3.3 Contributions

My contributions in this chapter are as follows:

Developed and exhibited a virtual reality documentary, "September 1955".

- Introduced various elements with which to create stories in virtual reality, including objects of interest, spatial soundscape, and animated characters.
- Provided a qualitative analysis of the viewer's experience based on exhibition diaries, screen recordings, and interviews. I argued that VR environments demand active participation of the viewer because these environments completely capture his or her visual and aural sensory channels. I also argued that limitations in attention caused each viewer to have a slightly different experience, which enabled multiple different inner stories of the same documentary.

Developed and co-instructed a class titled Computational Ethnography and Spatial Narratives.

- Through a series of exercises, I observed how students construct visual and material stories from their explorations of the city.
- I presented of example works and explained the approach taken by each student.
- I identified common spatial and temporal elements in the students' representations, including arrows, surfaces, color and light, which represented certain perceptual or active information they obtained from their spatial experiences.

Chapter 4

Spatial Experience as Inner Story: The Anchoring Framework

An ancient memorization technique called *the method of loci*, also known as *the memory palace*, illustrates two crucial facts about human memory: (1) remembering a series of unrelated pieces of information is very difficult unless it is organized, and (2) an effective way of doing this is to imagine a familiar environment, linking the information to be memorized to various locations in that environment, and taking a tour through it, visiting each location along the way. Memory athletes use this technique to remember thousands of pieces of unrelated information in order, such as arbitrary numbers. They consciously hijack a spatial mechanism in their brain that the rest of us use unknowingly when we interact with our environments. We are all natural memory athletes: imagine a route you frequently use and try to recall everything along that route. You will realize that you can effortlessly remember a great number of details. Neurological studies show that when memory athletes start recalling the information they have memorized, the regions in their brain that are related to spatial awareness light up.

The memory palace demonstrates that composing stories in space is clearly useful to memory. Is it also useful to our spatial experiences? I believe so. The ways in which we interact with our environments might be tightly connected to the way we describe and remember them. If that is the case, then there may be connections between the

language elements that we use to describe our environments and the perceptual and active information we register in our environments. This chapter describes my steps towards developing a computational account of these connections.

4.1 Inner Stories and the Anchoring Problem

One of the most interesting aspects of human spatial experience is that it can be framed as an inner story. By making inner stories, we connect perceptual features of our environments with cognitive processes. Forging these connections enables us to generate complex spatial representations that we can reason with. Research suggests that inner stories play a critical role in spatial experience. For example, a series of studies conducted by Shusterman and Spelke (Shusterman and Spelke, 2005) established that there is a causal relation between language learning and spatial reorientation ability. According to the authors, animals and humans hold multiple language-independent representations of space, such as representations for *proprioception* (sense of one's location in an environment) or for the senses of *left* and *right*. However, the authors suggest that only through language can the information from different spatial representations be combined. For example, infants 18-24 months old cannot solve a task that requires the use of the information *left of the red wall*, although they can solve a task that requires only the information *red wall* or *left of the wall*. The authors suggest that the missing link is a structure that can combine these two concepts. Only after the child learns the words "left" and "right" is he or she able to form those structures and perform the task. This study demonstrates that our inner stories enable us to combine language-independent visual and spatial representations, which otherwise remain unconnected.

How then are language-independent visual and spatial representations combined via inner stories (i.e. left of the wall + red wall = left of the red wall)? I frame this problem as a problem of **anchoring**: How are the symbols in language connected to—anchored in—visual, spatial and temporal aspects of environments?

I recently discovered that Jon Barwise and John Perry made similar use of the

term in their work on *situation semantics* (Barwise and Perry, 1981). In the context of situation semantics, the authors coined the term *anchor* as a function that links actual entities to a cognitive schema with indeterminate variables. For example, an anchor function links a particular entity, *John*, to a cognitive schema *A person is tired* and helps identify a particular situation *John is tired*¹. This approach was later adopted in cognitive robotics research in the context of symbol grounding (Coradeschi and Saffiotti, 2003). Silvia Coradeschi and Alessandro Saffiotti define the problem of anchoring as finding a mapping between a symbol and a corresponding visual signal. The main purpose of an anchor is to keep track of which symbol is connected to which signal during an agent’s interaction with the world. Forming these types of connections between symbols and images is useful, as in the case when a robot needs to search for a specific object given a list of properties, such as its shape and size, e.g. a bag that is red and small. Coradeschi and Saffiotti assign certain thresholds to determine the hue values that correspond to *red* and size values that correspond to *small*. Then, they use those selected thresholds to process the signals in the image to identify the particular bag (e.g. BAG2) they are looking for.

However, this notion of anchoring —finding and keeping track of identities of objects— offers only a narrow view of how we might make use of the available perceptual and active information in the environment. An ideal system with this type of anchoring mechanism should have a large dictionary of symbols in language, especially those that describe how things are related to each other in terms of where they are, how they look, or what they are used for. I argue that many of those symbols do not match any signals directly measurable in the images but nevertheless enable us to understand and describe complex visual scenes. I therefore suggest that we should not rely on predetermined thresholds to identify which type of signal is a match for which symbol. Instead we should learn visual representations that enable identifying different properties and relationships, which can then be used for generating symbolic descriptions.

¹This is a simplified exposition of the anchoring idea introduced by the authors. See (Barwise and Perry, 1981) for further details.

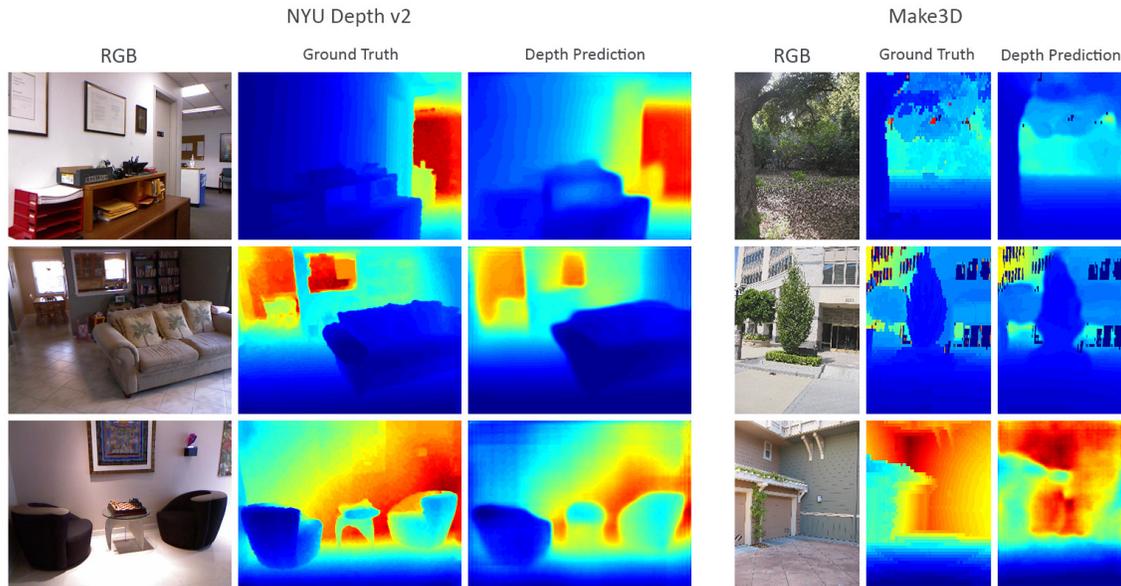


Figure 4-1: Learning depth representation from RGB enables identifying distances (Laina et al., 2016). A point in the resulting image can be qualified as CLOSE or FAR

4.1.1 Anchors Connect Symbols in Language to Visual and Active Information

I propose that solving the anchoring problem requires (1) determining the symbols relevant to the information in the environment, and (2) building representations that make this information evident. In some cases, these two tasks seem straightforward. For example, a symbol concerning distance, such as *close* or *far*, might be connected to the relative depth information in an image.² A representation that contains this information can be learned, for example, by a DNN that constructs a depth representation from an RGB image (such as in (Kraft, 2018; Cao et al., 2018; Laina et al., 2016)))(Figure 4-1). This representation makes evident what parts of an image can be considered close and what parts can be considered far. Depth is a feature with a clearly defined distance metric that enables making an easy comparison between two depth values and then identifying them

²*Close* and *far* are two examples of symbols that concern distances. They denote a relative distance of a point given a certain context. For example, something that appears close in an urban context can be far away in an indoor context.

SHADER CALIBRATION SCENE

METALLIC VALUE CHARTS

ALBEDO RGB

ALBEDO DEFINES THE **OVERALL COLOUR** OF AN OBJECT
VALUES USUALLY MATCH THE PERCEIVED COLOUR OF AN OBJECT

MEDIAN LUMINOSITY



NON-METAL sRGB RANGE **50-243**

METAL sRGB RANGE **186-255**

NON-METAL EXAMPLE VALUES



METAL EXAMPLE VALUES



METALLIC R

METALLIC DEFINES WHETHER A SURFACE APPEARS TO BE **METAL** OR **NON-METAL**
WHILST PURE SURFACES WILL BE EITHER **0.0** OR **1.0**, BEAR IN MIND FEW PURE, CLEAN, UNWEATHERED MATERIALS EXIST IN REAL LIFE
WHEN **TEXTURING** A METALLIC MAP, THIS VALUE WILL ALWAYS BE **GREYSCALE** AND IS STORED IN THE **R CHANNEL** OF AN RGB FILE

GREYSCALE



ALMOST ALL **NON-METALS** = **0.0**

ALMOST ALL **METALS** = **1.0**

SMOOTHNESS A

SMOOTHNESS DEFINES THE PERCEIVED **GLOSSINESS** OR **ROUGHNESS** OF A SURFACE
FOR TEXTURES, THIS IS STORED AS THE ALPHA CHANNEL OF THE **METALLIC MAP**

METALS



NON-METALS

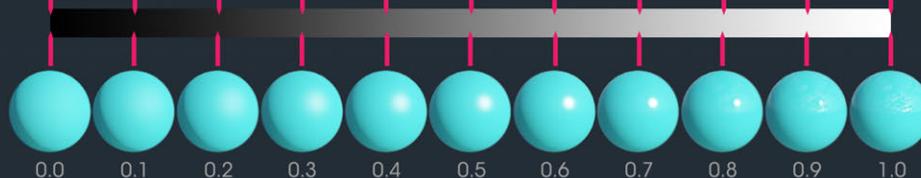


Figure 4-2: Graded material properties are used for rendering materials. Figure by Körner (2015)

with the symbols *close* and *far*. Note that the words "close" and "far" simply indicate two opposing distance qualities, and are often used in comparative and superlative forms. Things do not appear close or far in and of themselves.

However, the connection between a symbol and visual information is not always as straightforward. For example, if we are interested in material qualities —e.g. to describe something as *wood* or *metal* strictly from images— then we need a visual representation that makes evident the properties of *metal* and *wood*. In this case, it is more difficult to determine what type of representation we need that would discriminate the visual properties of *metal* from those of *wood*. Unlike "close" and "far", the visual properties of metal and wood do not have clear distance metrics that would allow a comparison. Moreover, it is difficult to determine how many other symbols might be in the same domain of materials, such as *glass*, which identifies completely different properties from those of *metal* and *wood*.

Alternatively, we can represent *wood* and *metal* as combinations of surface properties such as reflectivity, roughness, or color. In so doing, we can distinguish different material qualities in their respective representations; for example, we can describe a material as *shiny* or *matte*, *dark* or *bright*, *red* or *green*. Singling out visual dimensions with respect to these type of properties appears to be a good heuristic in the domain of materials. In fact, computer graphics implements this idea to compose and render materials (Figure 4-2). Consider, for example, the *luminosity* in the figure, which is represented as a linear scale between black and white. *Dark* and *bright* are two symbols that we can place on this scale at the two opposing ends. Any value in between can be considered either *dark* or *bright* based on its location on the scale. This property, when integrated into a final material, gives a certain luminosity to the material composite, while the luminosity property itself is no longer visible. Human vision might not be different from computer graphics in this regard. Our sensory-motor system creates many representations, all of which are fused together in our spatial experiences.

The idea that symbols can refer to locations in their corresponding representations is supported by studies on color perception. Research suggests that

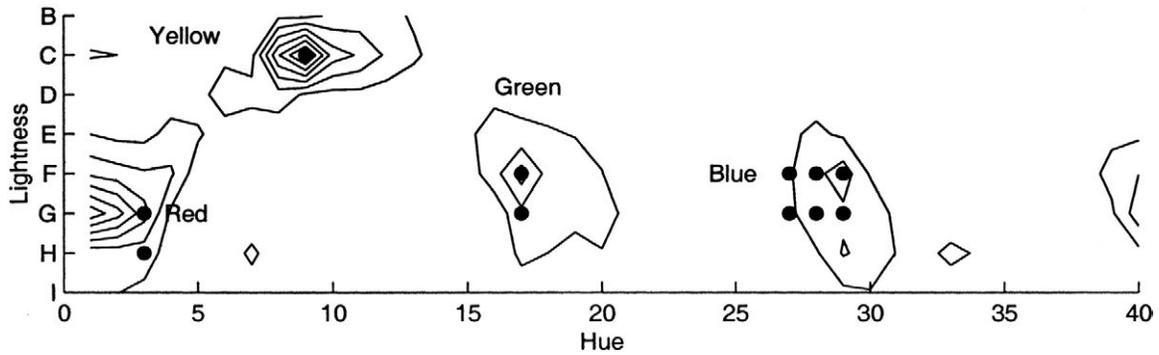


Figure 4-3: Contour plot of best representatives of color names in 110 languages, compared to English. This plot shows color names universally cluster around certain hue/lightness values, which are called *focal colors* ((Regier et al., 2005)).

regardless of culture and language, humans always prefer *focal colors* to which they assign a name and perform better in differentiating those colors (Regier et al., 2005). Figure 4-3 shows a contour plot of best representatives of color names in 110 languages in comparison to English. Focal colors are "pure" hues of colors such as red, green or yellow and are perceptually at maximal distance from each other. To put it simply, although one can see an enormous number of colors and perceive similarities and differences among them, there are a limited number of symbols that represent important locations (focal colors) in color space. One can still refer to a specific color outside one's color vocabulary, such as *greenish yellow* or *dark red*. This works insofar as the speaker has an understanding of how two color symbols are related to each other as in *greenish yellow*, or an understanding of luminance as an aspect of color as in *dark red*. We make use of the inherent relationships between what we perceive when we combine these symbols. For example the phrase *reddish green* makes no sense, because there is no perceptual similarity between these two colors.

How many such groups of symbols are there in language that differentiate one type of environmental factor from another based on a common visual representation they share? Although it would not be practical to consider all the ways in which every word can relate to a particular environmental factor, it is possible that our minds contain such mapping. In order to demonstrate the feasibility of this approach, I

focus on a limited vocabulary of visual descriptions that I obtained from a visual exploration study conducted in simulated environments.

4.1.2 The Anchoring Framework

So far, I have introduced a particular approach for connecting symbols in language to visual information. In this approach, the symbols refer to localities in domain-specific representations that make evident a particular type of information. A depth representation makes evident information concerning distances, and a color representation makes evident information concerning colors. I define an anchor as a function that generates a symbolic relationship given a domain-specific visual representation and a particular location(s) within that representation. Consider the sentence "There is a red cup on the table." Figure 4-4 illustrates some of the domain-specific representations and symbols that might be used for producing this description with anchors. In this example, there are three different domains: the object identity domain, which concerns identifying objects such as a cup or a table, the color domain, and the vertical relations domain. Each of these domains helps to generate one part of the description: *cup* and *table* are generated from the object domain, *red* is generated from the color domain, and *on* is generated from the vertical relations domain. Below, I introduce a series of computations that generate descriptions from images using anchors:

Assume there are two points p and q in an image that we are interested in describing.

- 1. Select a series of visual properties to describe:** There are many ways to describe a visual scene, and each description might concern a different subset of visual properties within the scene. In order to generate symbolic relationships with anchors, we must select the properties that we are interested in describing, such as distances, colors or spatial relations. The study in the next section will illustrate the most common properties people described while they explored an environment.

- 2. Generate domain specific representations:** Describing a feature in the image is a domain-specific question to be answered by the vision system. A domain

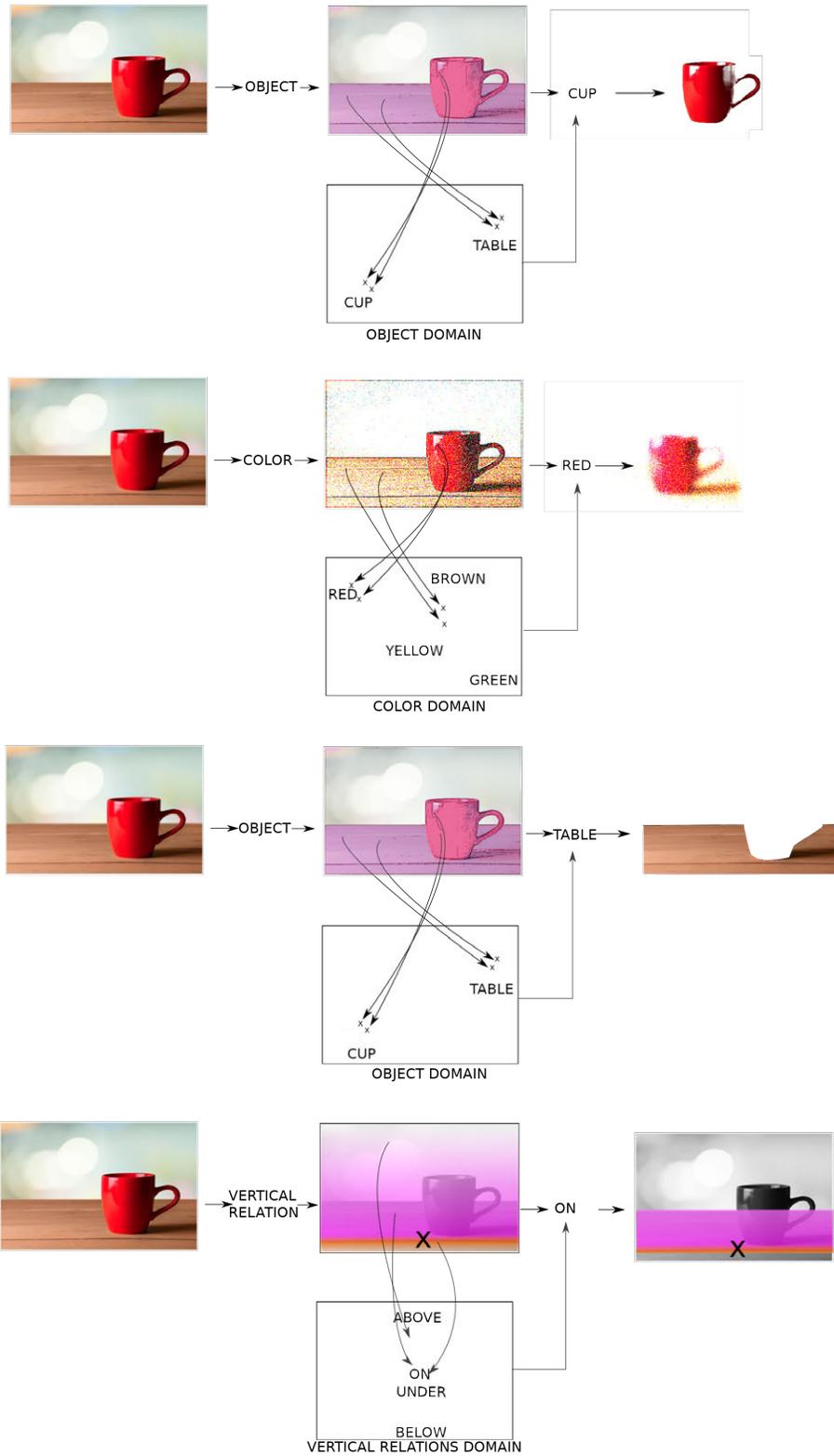


Figure 4-4: Some of the visual domains and symbols that generate the description "There is a red cup on the table". Note that domain specific representations and corresponding domains are only for illustrative purposes.

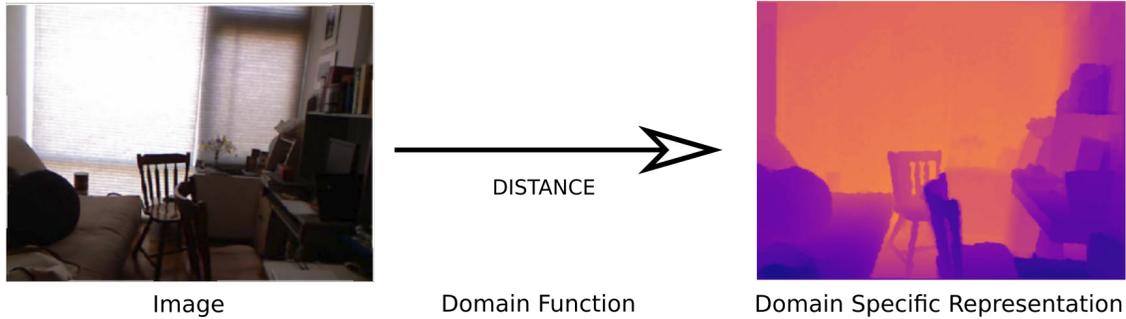


Figure 4-5: Generating domain-specific representations for distances

function takes an image and sometimes other arguments, and returns a domain-specific representation that makes it possible to answer certain questions unambiguously. A domain function is denoted as $D(x, [args]) \rightarrow x'$, where x is the input image, x' is a domain-specific representation, and $args$ are the additional parameters that might be required to generate the domain-specific representation (angle of the camera, for example).

For example, to answer a question about whether an object is above or below the observer, a domain function takes an image, along with additional information about the orientation of the observer's head when the image was captured. The domain function is responsible for producing from these arguments a map of Y coordinates relative to the observer, a map that resolves ambiguity and makes it possible to decide whether the surface that generated each pixel in the original image is above or below some 3D reference point.

The domain-specific representation x' is image-like in that it contains an entry for each of the pixel locations in the original image. At each of these pixel locations, x' contains some domain specific representation that pertains to a certain type of question answering —e.g. "How far is this point?" or "What color is this point?" For example, the output of the domain function for reasoning about distance could simply be a depth map estimated from the image. The output of the domain function for reasoning about object types could be a dense semantic labeling of the image, where each pixel maps to a set of object categories.

Anchor functions take as input a domain-specific image description, x' , and often

other anchor-specific arguments. For example, the anchor ABOVE takes as x' a map containing Y coordinates in the 3D world, and a reference Y coordinate Y_{ref} . ABOVE returns true for a given image coordinate iff $x'[image_coord] > Y_{ref}$, and constructs a symbolic relationship for the given coordinate.

3. Generate anchors: Each domain has a finite set of anchor functions that answer questions about the presence of a particular property. Anchor functions are Boolean-like functions in the sense that they return true (construct a symbolic relationship) if a corresponding feature is present at a point. For example, an anchor function RED returns true if the color value in the representation corresponds to the color red and constructs the description "x is red", but returns false otherwise. This function takes in a domain-specific representation (e.g. color) and an input point.

Consider the description "There is red cup on the table". Below are some example anchor functions that may return true or false for the given points p and q . Each domain specific representation is denoted with a subscript, such as x_{color} . Let p denote a vector that corresponds to a location on the cup, and q a location on the table:

$$[1] \text{ RED}(x_{color}, p) = true$$

$$[2] \text{ GREEN}(x_{color}, p) = false$$

$$[3] \text{ CUP}(x_{identity}, p) = true$$

$$[4] \text{ TABLE}(x_{identity}, q) = true$$

$$[5] \text{ UNDER}(x_{vertical}, p, q) = false$$

$$[6] \text{ ON}(x_{vertical}, p, q) = true$$

4. Generate Descriptions: Once we have obtained results from anchor functions, we can generate descriptions for the points p and q for each pair of anchors and obtain the following result:

For the point p : (RED, true) \rightarrow "p is red" (CUP, true) \rightarrow "p is cup"

For the point q : (TABLE, true) \rightarrow "q is table"

For the pair of points p and q : (ON, true) \rightarrow "p is on q"

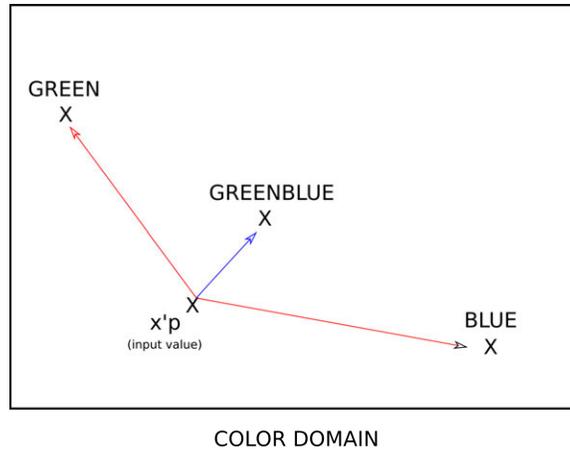


Figure 4-6: In some cases, using an average location of two domain symbols might generate the closest location to the input value. In this example, a color value is tested against BLUE and GREEN, both of which are at similar distances. A new anchor function, GREENBLUE is generated, which has a new preferred location at the average of the original two. GREENBLUE has a closer location to the input value. A color here can be described as "greenish blue."

5.Combine Descriptions: Descriptions of two anchors can be combined if both anchors use the same input parameter:

["p is red", "p is cup"] → "p is red cup"

["p is on q", "q is table"] → "p is on table"

["p is red cup", "p is on table"] → "red cup is on table"

Generating new anchor functions: In most cases, anchor functions from the same domain are not combined, because only one of them can return true for a given value. However, if the distances between an input value and the preferred locations of two different anchors are similar, that is, if the location lies between two anchors, they can be combined. For example, a color between green and blue can be represented as "greenish blue" by combining the two color anchors. For those cases, a new anchor function can be added to the domain, (i.e. GREENBLUE) which has a new preferred location in between the original two (Figure 4-6).

The five steps of anchoring I described above enable us to generate inner stories from visual observations, as well as to make inquiries about a visual scene using natural language. For example, the question "Is the cup on the table blue?" can be

represented in the form of anchors, then tested against the input image, returning false. This is due to the fact that symbols in the input query are associated with specific domains and domain specific representations.

4.1.3 Summary

In this section, I introduced the idea of anchoring, a novel mechanism that enables connecting visual information with symbolic representations, namely inner stories. The anchoring framework considers inner stories as fundamental mechanisms that establish connections between visual and spatial information, and enable reasoning with them. There is much supporting evidence in cognitive and neuroscience that this type of mechanism might be biologically implemented in the human brain. I presented a few of those studies in the beginning of this chapter.

In order to better understand the range of domain-specific representations and particular language elements that are connected to those representations, I conducted a visual exploration study, through which I observed the way people explored an environment and verbally described what they saw. Based on this study, I identified various types of anchors that can generate a significant portion of the verbal descriptions observed in the study. Table 4.1 illustrates a few of those anchors and corresponding domain-specific representations.

4.2 Visual Exploration Study: The See, Act and Tell Methodology

In this section, I outline a methodology I call *See, Act, and Tell*, which I developed for observing spatial experiences in virtual reality environments, and describe the study I conducted using this methodology. The *See, Act, and Tell* methodology aims to capture the relationships between visual perception, action, and inner stories. Because we cannot observe inner stories directly, I asked people to explore a virtual reality environment and verbally describe it during the experiment. I selected the task of free

Information	Domain-Specific Representation	Function	Example anchors in the domain
Vertical Locations	Heights of points in space	Create vertical relationships	ABOVE, BELOW, UP
Distance	Distances of points in space	Create distance of objects	HERE, THERE, CLOSE, FAR
Depth Order	Distances of points in space	Create depth order of points	BEHIND, FRONT
Horizontal Locations	Horizontal vectors originating from a point in space	Create horizontal relationships	LEFT, RIGHT
Surfaces	Surface orientations	Create attachment relationships	ON, AT
Surfaces	Distances and Surface Orientations	Create path relationships	ACROSS, ALONG
Surfaces	Distances and Surface Orientations	Create containment relationships	INSIDE, OUTSIDE

Table 4.1: Examples of anchors and domain-specific representations.

spatial exploration and verbal descriptions for multiple reasons: it is relatively less biased in comparison to other types of spatial tasks; it is easy to make comparisons between verbal descriptions; and asking people to describe an environment naturally encourages them to move around space, creating more opportunities to explore their spatial experiences.

I conducted this study in collaboration with Ainsley Sutherland and Dani Olson, over a period of two months in the summer of 2017. Initial findings of the study were released in (Zaman, 2018; Gilpin et al., 2018; Olson et al., 2019), which introduced several potential applications enabled by the methodology and the collected data. Gilpin et.al. showed that the particular data generated in the study would enable machine learning systems to evaluate the validity of visual scene descriptions (Gilpin et al., 2018). Within the scope of this dissertation, I will focus on the relationships between environmental information and verbal descriptions. I will show examples of how anchors enable modeling these relationships. This study will also help in

understanding the limitations of the current anchoring framework and provide insights for future developments.

The *See, Act, and Tell* methodology builds on previous methods for observing people while they engage in a task. I used VR for this task, which has inherent advantages and limitations. In order to keep the chapter focused on the anchoring idea, I provide a discussion on the methodology in Appendix A.

4.2.1 Methodology

In See,Act, and Tell methodology, I immersed participants in a simulated virtual reality environment and asked them to explore it, describing it as they explored. In each session, I recorded the movements of participants, their gazes, and their verbal descriptions. This methodology had several advantages over *the think aloud* protocol. Instead of asking participants to verbalize their thoughts, as in the *think aloud protocol*, I asked them to verbalize their perceptions — that is, I asked them to describe the environment and objects in terms of their locations sizes, textures, and colors, as if they were seeing these objects for the first time. People could easily relate to this task and found it intuitive to describe what was in front of them. More importantly, because I recorded the visual field and the movement of the participant, I had multiple ways to analyze the verbal descriptions based on where and when an observation was made.

Virtual Reality Environments

In order to perform the visual exploration study, I created three interior environments in virtual reality. Each environment consisted of a series of familiar objects that are found in everyday spaces, such as chairs, tables and books. Figure 4-7 shows a still from one of the environments. All environments were obtained from Evermotion (Evermotion, 2019) and placed in Unity Game Engine (Unity, 2019). The environment specific details were as follows:

1. Residential Environments: Out of three environments, two were residential



Figure 4-7: A still image from one of the environments in the experiment. The environments consisted of indoor spaces with everyday objects.

indoor settings. Both of these environments had similar, open layouts, with a kitchen and a living room contained in the same space. They both had a mezzanine where the bedroom was located. Observers were not allowed to go upstairs but could see some parts of the second levels.

2. Office: The third environment was an office space. There was an open office area with desks and computers in the center, and an adjacent meeting room. This environment also contained some unusual objects such as a motorcycle in front of what appeared to be a garage entrance.

Subjects were instructed to explore and describe these environments. They were given the prompt:

" Describe the environment for people who are not there and who may not be familiar with the everyday objects and their uses. You may assume you are describing the environment to an alien."

I created this particular prompt for several reasons: First, I wanted to ensure that my subjects would not mistakenly focus on only the interesting parts in the environment instead of describing their experience as it unfolded. Second, asking the participants to describe the environment to an alien caused them to pay attention to details. For example, in a kitchen, a participant would see a sink and a tap, and

describe each item in terms of how it looks, what material it is made of, and its apparent function. Without this particular prompt, a participant would probably omit those details, assuming that they were understood.

Methods and Procedures

Each observation study took between 10 to 15 minutes. I determined 15 minutes as the maximum length of the study, but the participants were instructed to stop the task any time they wanted. The task took place in a conference room in which I set up an empty space for the participant to move freely. The participant was not interrupted during the study unless there was a technical issue, such as the participant getting too close to an obstacle in the experiment area or the VR headset losing tracking. In those cases, the participant was interrupted and repositioned in the study area. By doing so, I made sure the headset was properly positioned, the participant was comfortable, and she understood the basic mechanism of navigation, explained in the following section. See section A.3 for a consent form sample provided to the participants.

Navigation within the virtual space

Participants had limited means of physical movement within the space due to the technical limitations and space considerations. They could lean over and get closer to the objects in the virtual space, but they could not move over large distances by walking. In order to help the participants to navigate within the virtual environment, I developed a simple interaction method using the Oculus Remote Controller. In this interaction method, participants pressed the remote button, which initiated a slow movement towards the direction of their gaze. Once the movement was initiated, the participant kept moving in the same direction as long as she or he kept pressing the button, even if she or he turned their gaze to another direction. This essentially allowed the participant to keep observing the environment while moving. I performed several pilot studies to test this navigation mechanism before implementing it for the final study.

Time	x	y	z	q1	q2	q3	q4
2.011325	-4.588	-1.273	-4.125	0.2216	-0.09706	0.009789	0.9702
2.021552	-4.589	-1.274	-4.124	0.2246	-0.0977	0.009816	0.9695
2.033029	-4.59	-1.275	-4.123	0.2273	-0.09885	0.01022	0.9688
2.043853	-4.59	-1.275	-4.122	0.229	-0.1001	0.01062	0.9682
2.054788	-4.591	-1.276	-4.121	0.2304	-0.1008	0.01064	0.9678
2.066194	-4.592	-1.276	-4.119	0.2316	-0.1024	0.01055	0.9673
2.077256	-4.593	-1.277	-4.118	0.2331	-0.1035	0.01063	0.9669
2.089143	-4.593	-1.277	-4.117	0.2346	-0.1054	0.01057	0.9663
2.099061	-4.594	-1.278	-4.116	0.2363	-0.107	0.01108	0.9657

Table 4.2: A sample of camera parameters. $[x,y,z]$ is the camera location in meters, $[q1,q2,q3,q4]$ is a quaternion representing the camera orientation.

Data Collection Procedure

During each session, I recorded several parameters of the observations made by the participant, including the camera locations and orientations, and made a voice recording of the participants' visual descriptions. The camera parameters were timestamped so that I could align them with the verbal descriptions. The location was represented in meters relative to the scene origin via a three-dimensional vector. The orientation parameter was a four-dimensional vector corresponding to a quaternion, which described the three dimensional orientation of the camera in the space. I sampled the camera parameters in every frame, with a varying frame rate between 70 to 90 frames-per-second(fps). Table 4.2 shows a few samples from the collected camera parameters.

The voice recording was encoded in WAV format and saved at the end of each session. Each session was assigned a unique session ID, which was used to identify camera parameters and voice recordings.

Data Processing and Generation

I performed a series of data processing and generation procedures for transcribing the voice recordings, aligning verbal descriptions with camera data, and generating various image data such as RGB, depth or surface normals. I did not store the visual information during the sessions because it would cause the VR system to have latency issues. I wrote a script that imported the camera parameters to a scene and generated

Word	Start	End
across	23.23	23.65
from	23.65	23.76
me	23.77	24.18
with	24.26	24.69
lots	25.4	25.73
of	25.73	25.88
books	25.88	26.14

Table 4.3: Timestamped words generated by Gentle Aligner.

visual information for each camera parameter.

I used an online service for transcribing the voice recordings. The transcription did not have any time stamp and did not eliminate interruption words such as "um" or "er." The following is a sample transcription:

To my right, I see a curtain that, can be used to, separate the hallway from the rest of the house. It's gray, can't really see through it. To my right I see, long, thin, tall windows. Uh, don't allow too much light in. In front of me I see stairs. Um, they look pretty spacious, kinda scary looking. Lots of space in between them. I'm kinda afraid to go up them because there's no rail. Above me I see that there's a second floor. I can see paintings upstairs along the wall. The paintings, are lined with a white border, different from the paintings downstairs. They look less abstract and more, like baskets and realistic.

In order to align the transcript with the camera data, I used a software called Gentle Aligner (Ochshorn and Hawkins). Gentle aligner took a sound file and a transcript and produced a timestamped alignment data. The resulting data had a timestamp for each phoneme and word, indicating the start and end times in seconds. I used the word data only for further processing. Table 4.3 shows a sample of aligned word data.

Using this information I generated a sentence level alignment data, such as the one below:

'sentence': "I'm looking above in the center of the room.", 'start':



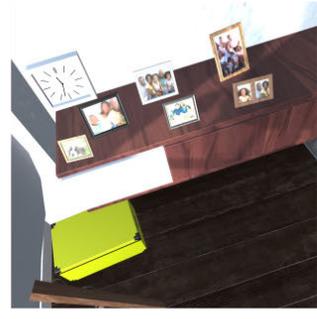
t=157s

...



t=161s

...



t=165s

{'sentence': "Then there's another lamp and there's a really small, not that small but there is a yellow box below this lamp.", 'start': '158.63', 'end': '165.67'}



t=631

...



t=633s

...



t=634s

{'sentence': 'So, getting here I see this huge plant back in this room.', 'start': '631.28', 'end': '634.84'}



t=306

...



t=308

...



t=312s

{'sentence': 'All right, I have what looks like a knock off of an Eames chair.', 'start': '305.96', 'end': '312.93'}

Figure 4-8: Sample transcripts with corresponding RGB images that are obtained from the study.

'693.45', 'end': '695.97'

Timestamped camera parameters and voice transcriptions enabled us to relate descriptions with what people saw and how they moved within the environment. Figure 4-8 shows some examples of aligned visual and verbal data.

4.2.2 Study Results

I conducted the study with a total of 27 participants (11 females / 16 males). The participants ranged in age between 21 and 32 years old. Before each session, the participants confirmed that they could comfortably see the virtual environment and were able to navigate within the environment using remote controller. The resulting dataset consisted of a total of 6.5 hours of recordings. One of the participants did not want to continue the study. I omitted the recordings from this participant. I originally intended to conduct a total of 30 studies but I could not achieve this goal within the time frame of the study. Out of three environments I had ten participants for each residential environments and six participants for the office environment. The limited data collected in the office environment introduced challenges for making comparisons between different types of environments, and this type of comparison was omitted from the present study.

First, I will present my observations regarding the way participants explored the environments. These observations include the amount of time spent by each participant observing different parts of the environment, the actions involved in their observations, and how those actions were involved in establishing relationships between observed objects and qualities. Following these observations, I will present an analysis of verbal descriptions using the anchoring framework.

Analysis of Explorations

Although each participant took a different route and spent time in different parts of the environments, there was a significant overlap among the routes and locations of different participants (Figure 4-9). There were parts in each environment that

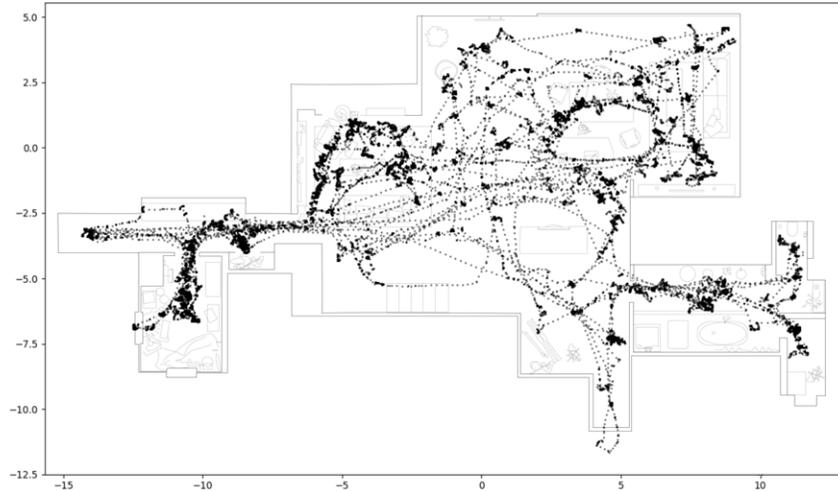
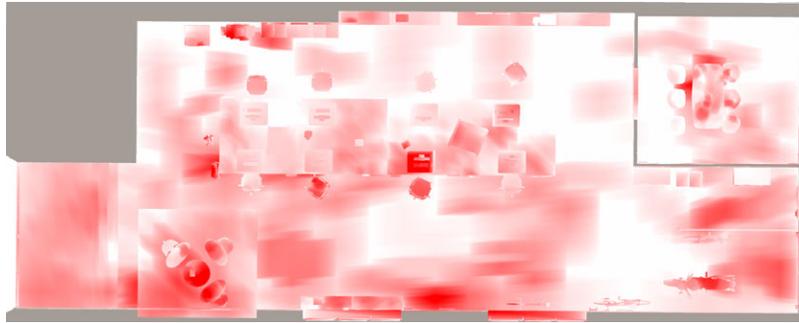


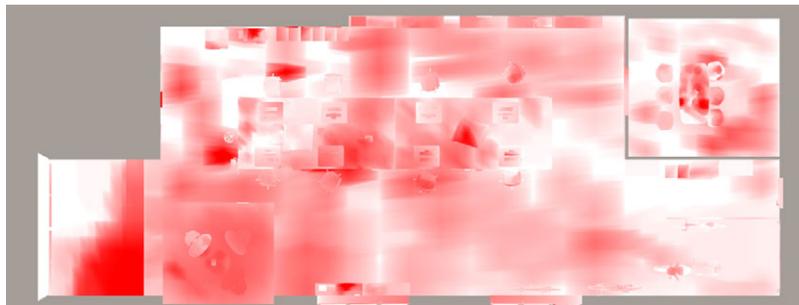
Figure 4-9: An overlay of the locations of all subjects in Environment 3. See Appendix A for other environments.

the participants neither navigated nor described. Their interest centered primarily around the objects such as furniture or distinct material qualities of the environments such as the green paint on a ceiling.

I produced heat maps to visualize the most common parts and aspects of the environment that people paid attention to (Figure 4-10). Because I did not have actual gaze data, I assumed that each participant was looking more or less at the center of her or his visual field. In these heat maps, a gradient value from white to red represents the time of attention at a particular location, where white represents 0 seconds, and red represents >10 seconds. I overlaid these heat maps across participants and generated an average attention map for each environment. I then used these heat maps to determine the average attention value for each frame of recording from the participant's point of view—the average time of attention for all visible objects in a frame. Figure 4-11 shows the top five most attended parts of all three environments. Both the heat map in Figure 4-10 and the frames in Figure 4-11 illustrate that the "hot spots"—where people spent more time attending—were generally spread across the environments. For example, for Environment 1 (first row), these spots correspond to



(a) Subject 1



(b) Subject 2

Figure 4-10: User attention mapped in each environment. The brightness of the color represents the gaze time.

a corner of the living room, the table in the dining room, the oven in the kitchen, the sink in the bathroom, and a maroon chair between the living room and the dining room. These spots might have functioned as landmarks within each environment, which people use to determine their relative location and orientation, although further evidence is required for supporting this idea.

The way people acted within the environment provided important insights about their spatial experiences. By moving around the environment and attending to different locations, the participants gathered information regarding where they were, how they were oriented in regard to the objects and the overall geometry of the environment, and how objects and materials related to each other. The verbal descriptions that followed or accompanied these movements illustrate how this information was integrated into descriptions. Figure 4-12 shows the movements of two participants within an environment and the locations of objects or qualities each described, indicated by red arrows. In the following section, I will examine the

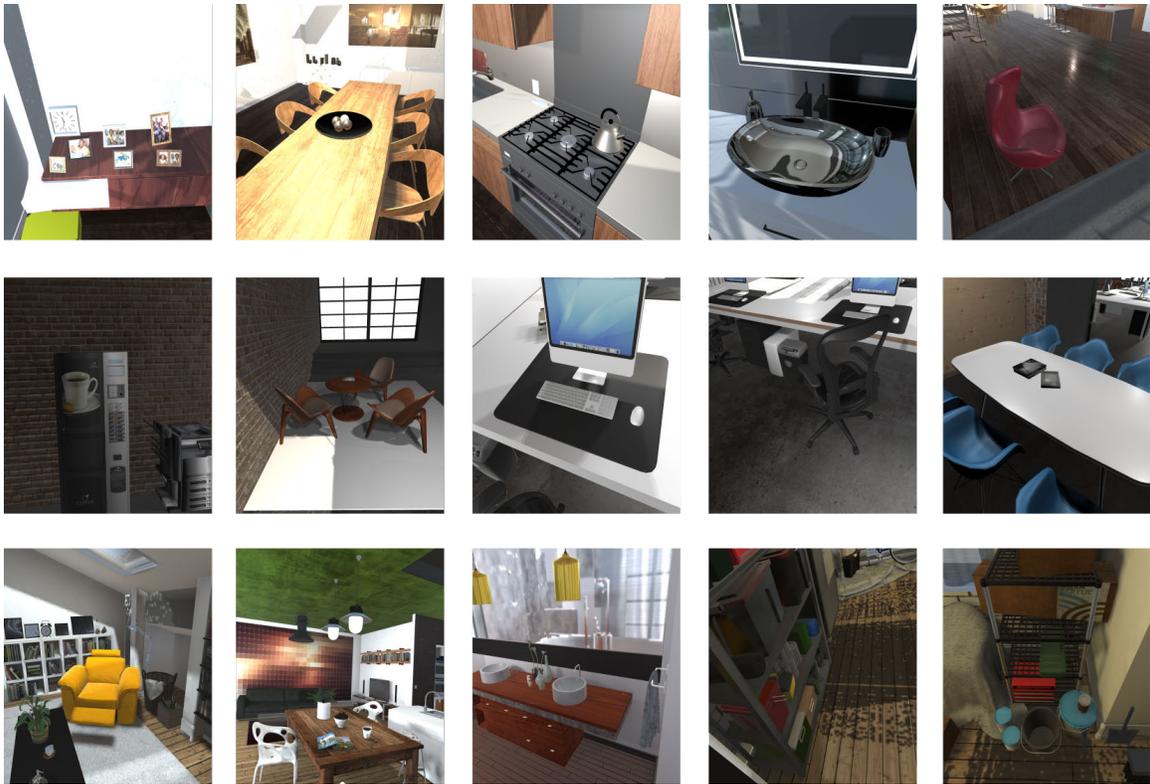
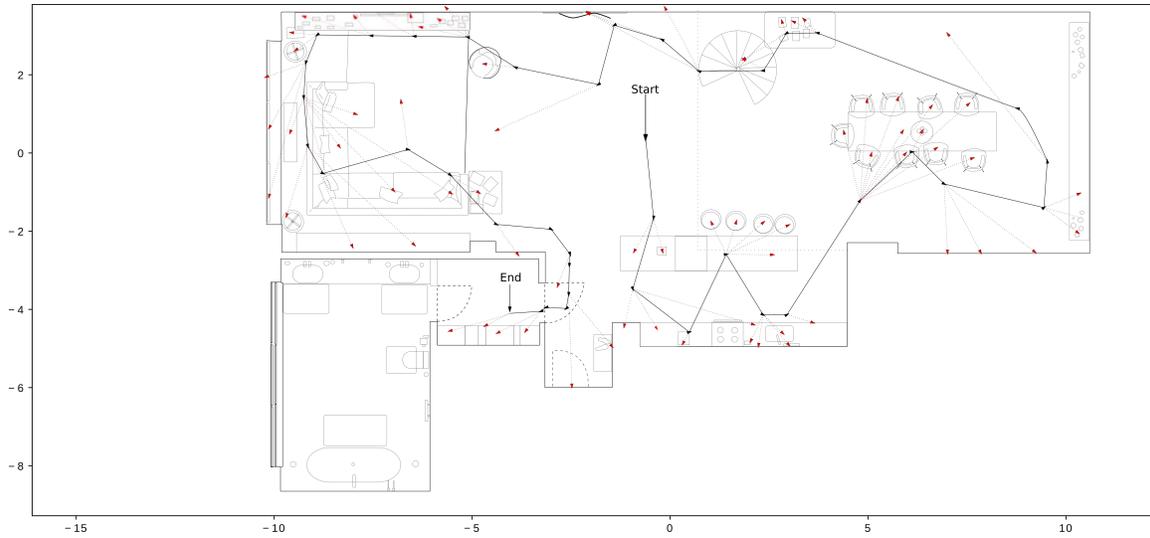


Figure 4-11: Top 5 most attended parts of the three environments.

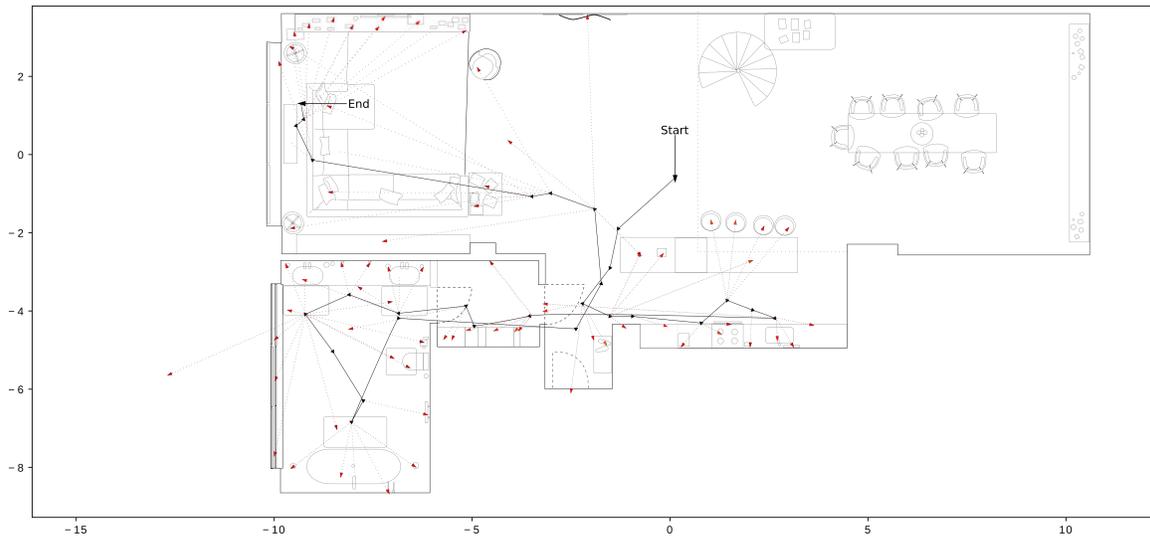
actions the participants took when they observed the environment.

Actions

The recorded camera parameters allowed me to identify the location and orientation changes during an exploration. The graphs in Figure 4-13 show the changes in camera parameters over time. The spikes in these plots indicate a drastic change in either location or orientation of the camera. When I examined the verbal descriptions and corresponding camera images in those moments, I identified three distinct actions that caused these changes: navigation, turning, and looking around (surveying). I discovered that each of these actions enabled a certain type of observation in the environment. For example, surveying enabled a participant to understand where she or he was, which was often followed by a remark on her or his location. Similarly, the navigation and turning indicated that the participant was about to focus on a different part of the environment.



(a) Participant 1



(b) Participant 2

Figure 4-12: Movements and observations in Environment 2. This figure shows the movements of two subjects exploring the same environment. Red arrows indicate that subjects made an observation about the object or material quality that the arrow is pointing to. Each participant made a 15-minute observation. They navigated similar distances (Participant 1 70 meters, Participant 2 60 meters) and observed similar number of objects or qualities, 66 and 69 respectively.

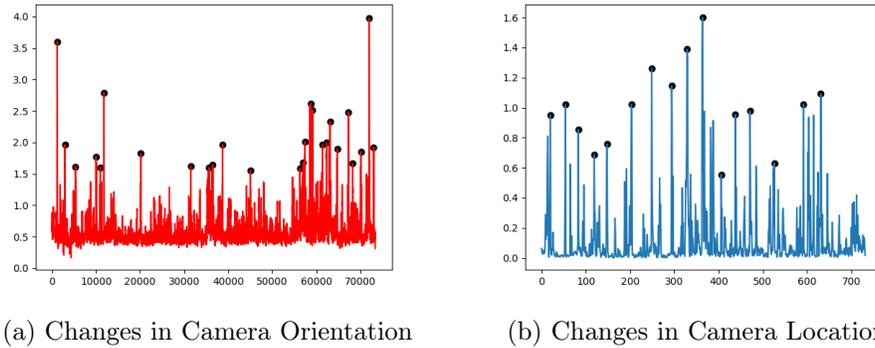


Figure 4-13: Changes in orientation and location during an exploration. (a) Orientation changes. This graph shows the difference in degrees between forward angles of the camera parameters obtained in successive frames. The black dots indicate the most drastic changes in orientation. (b) Location distance. This graph shows the distance in meters between two camera locations obtained in every 100th frames. The black dots indicate the most drastic change in locations.

Establishing Location by Looking Around (Surveying)

One of the most characteristic aspects of spatial experience is that it allows us to know where we are relative to the environment around us. There are many factors in determining one's location and relative orientation within the environment including the geometry of the space and landmarks represented by salient objects or material qualities. This information is spread across the environment, and one needs to survey the environment by looking around into different parts. In my study, I discovered that 21 of 27 participants (73%) began their explorations by surveying the environment. They turned around to capture a wide panorama, some of them making large turns of about 180 degrees (16 /27) and some of them making smaller turns. This initial survey usually ended with the participant making a remark about where she or he was or by describing the space she or he was in. Below are a few examples of the remarks made after an initial survey action:

[1] "So, I am in the middle of a space with a large communal work desk in the center"

[2] "I am in a conference looking room"

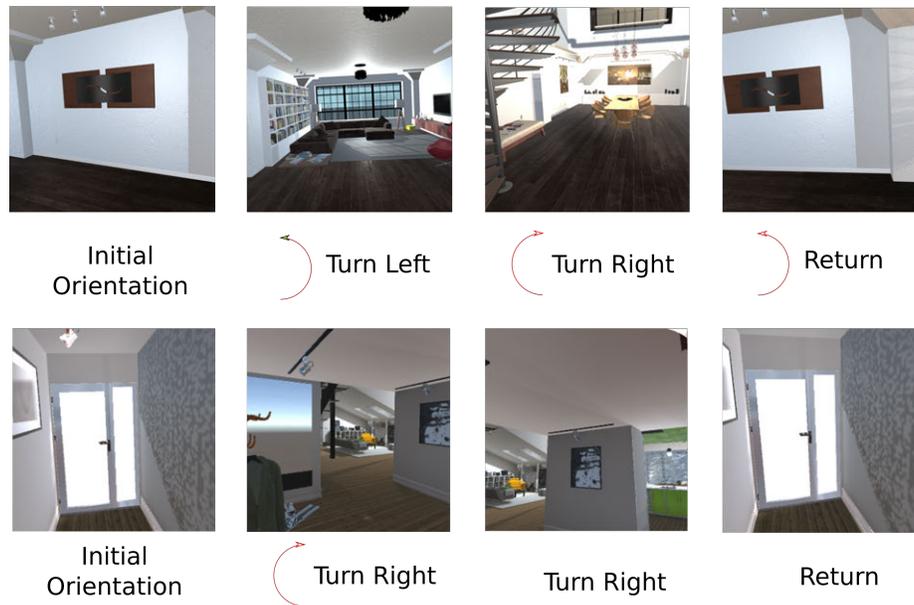


Figure 4-14: Look around action enables establishing a location for the observer. This action was observed multiple times for a participant's exploration of the virtual environment. 73% of the participants looked around in the beginning of the task.

[3] "This space is very wide open and nice, has a lot of lighting"

[4] "I am in a room with what looks like laminate flooring"

[5] "This is somebody's fancy apartment"

The surveying action was frequently performed when the subject needed to make location references between objects. For example, while describing a TV, a subject quickly looked behind to locate the couch, and made the observation: "I can sit on this couch to watch TV." Another noticed two identical lamps located in two opposite corners by successively looking back and forth between those corners. I also observed a similar behavior for establishing vertical relationships between objects. For example, a subject looked above to the ceiling to locate the spotlights and remarked that the light above him illuminated the painting in front. Figure 4-14 illustrates a few examples of *look-around* actions.

The looking around was a distinct action from turning around because in looking around, the subject returned the initial direction of his or her gaze, often not changing body position. This action was also more saccadic, usually taking only

a few moments for user to return to the original orientation. The lack of body movement and return to the original gaze direction indicates that the subjects might have used body orientation within the environment as a reference (e.g. not changing the direction they consider as the front) while establishing spatial relationships between their observations.

Changing Location by Navigation

Subjects navigated the environment by using the controller or by actually walking. Due to space limitations subjects were advised to use the controller over large distances. The navigation action, in the scope of my experiments, therefore were mostly linear movements between two locations, triggered by the controller. Subjects used navigation sparingly, and mostly stood still and described what they saw from their vantage point.

Navigation action was often preceded by a remark by the subject such as "moving to the kitchen", "let's go over there," indicating she or he was about to move. Most of these remarks included a reference to an object or location in space especially when it was relatively far away. Navigations over short distances were frequent, often going unremarked upon by the subject. The participants often navigated back to a place they had visited before, essential making a loop. Some of the participants explicitly made a remark about going back to a previous location, such as "Let's go back to the kitchen" or "So, I am back in the living room."

I counted how many times each subject navigated using the remote controller. The results show that there was no significant variation across the subjects or the environments. The average number of navigations were 38, 33, 37 times for Environment 1, 2 and 3 respectively. Figure 4-15 shows the number of navigations per subject across three environments. This result suggests that regardless of the size and the type of the environment, the subjects visited a similar number of locations in similar time frames.

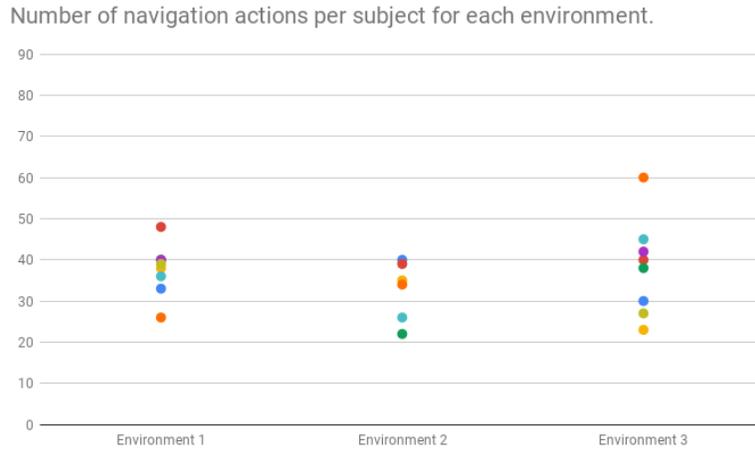


Figure 4-15: Number of navigation actions per subject in each environment. This figure shows that there is no significant difference among environments as to the number of navigation actions taken within them.

Changing Context by Turning

Turning action, quick and definite change in the participants orientation, was frequently observed in all sessions. In fact, turning was the most frequent action overall. Two different functions of turning emerged: in the first type, a turning action was followed by a navigation action. Participants turned to some specific direction just before they wanted to move to a new location. In most cases, participants turned directly towards the designated target. In fewer cases, subjects moved in the general direction of a target and then adjusted their orientation along the way. The second function of turning was to allow the participant to observe more from her or his location, similar to a looking-around action. In both cases, the turn action indicated a change in the context and was often followed by a remark on a new location or a move to a new direction. For example, a person describing what she saw in the kitchen would sometimes turn away and begin describing another part of the room, remarking that "On that side of the room ..." Contextually relevant objects, such as objects in the kitchen, were found in close proximity to each other, and participants might have found it easier to describe relevant objects spatially and contextually in one episode (e.g. without turning away). Therefore,

turns usually indicated a transition from one context to another, both in terms of verbal descriptions and the participant's location in the environment.

In the final section, I will return to the idea of anchoring and show that an anchoring framework enables modeling the verbal descriptions made by the participants. I will provide answers to the following two questions: What are the common language elements used by the participants when they describe the environment? Can we discover domain-specific representations that enable anchoring these language elements and generating descriptions?

4.3 Anchors and Domains

The anchoring framework I introduced in the beginning of this chapter suggests that descriptions are generated from various domain-specific representations, such as depth, color or vertical relations. Anchors connect symbols to these representations and generate descriptions by combining those symbols. With this approach we can generate inner stories, in which various story elements (e.g. objects or material qualities) are connected to each other. In fact, the descriptions generated by the participants include many such connections. These connections are made based on spatial and temporal relationships, or visual properties such as shapes, colors and materials. Consider the following transcript:

So, exiting the bathroom at this time. At the front entrance, there is another large mirror and there's a front door. I assume I can't go up the stairs. Sort of in the area underneath the stairs, there is a little bench with a few books on it. A picture hanging on the wall that looks like sort of a baroque interior with huge chandeliers and a piano. There's other images on the walls.

I emphasized various aspects and relationships that are described by the participant, which include objects (blue), visual properties (purple), spatial relationships (red), temporal relationships (green), and grouping (orange). Apart

from those, I also emphasized the first sentence (brown) which implicitly signifies an observed spatial relationship such as "I am *in* the bathroom," and grayed out a description that concerns a painting within the virtual space. Notice how each of these spatial, temporal, and visual elements establishes a particular type of relationship among the observer and the objects: both the mirror and the front door are *at* the same location; individually perceived books are grouped as *a few* and all of them are *on* the bench, which is *underneath* the stairs. After observing a picture on the wall, the observer sees *another* one. The mirror and bench have different size qualities: one of them is *large* and the other is *little*.

As in this example, the participant uses visual, spatial, and temporal information in her experience to form a variety of relationships between observations. These relationships were prevalent among all verbal descriptions I collected in the study. The most frequent ones were spatial relationships. The *on* relationship was observed 588 times and the *there* relation was observed 776 times in 3908 statements. In total, there were 41668 individual words with 2551 unique instances in the dataset. More than a third of the most frequent 50 words were those that explicitly refer to an environmental factor (excluding articles and pronouns). See Figure 4-16 for the most frequent words in the dataset.

One might argue that the high frequency words in the dataset, especially prepositions, might also be as frequent in other corpora. However, I will show that the words that relate to spatial, visual and temporal information in the environment are not as frequent in other corpora, such as the Brown Corpora (Francis, 1971). I present a comparison among my dataset, and two categories from the Brown Corpora: news articles and fiction. Figure 4-17 shows the normalized frequencies of the most frequent spatial words in those datasets. Except for the words "in" and "with", visual descriptions contain significantly more spatial words. A comparison of the most frequent visual words shows even more drastic difference between visual descriptions and the Brown Corpora (Figure 4-18).

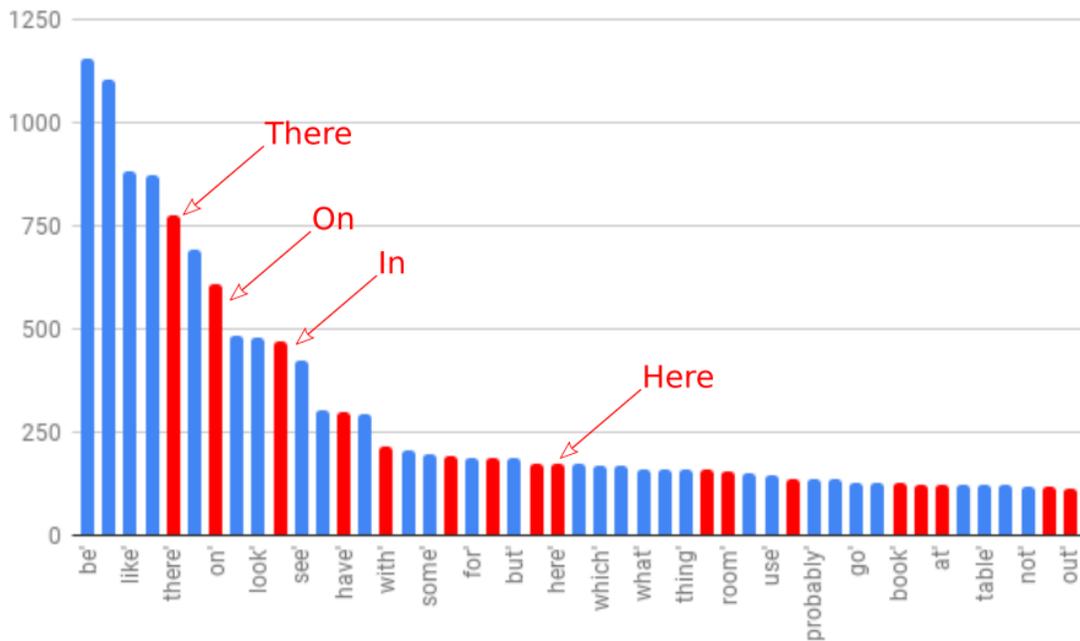


Figure 4-16: 50 most frequent words in the dataset. Red columns indicate a word related to an environmental factor.

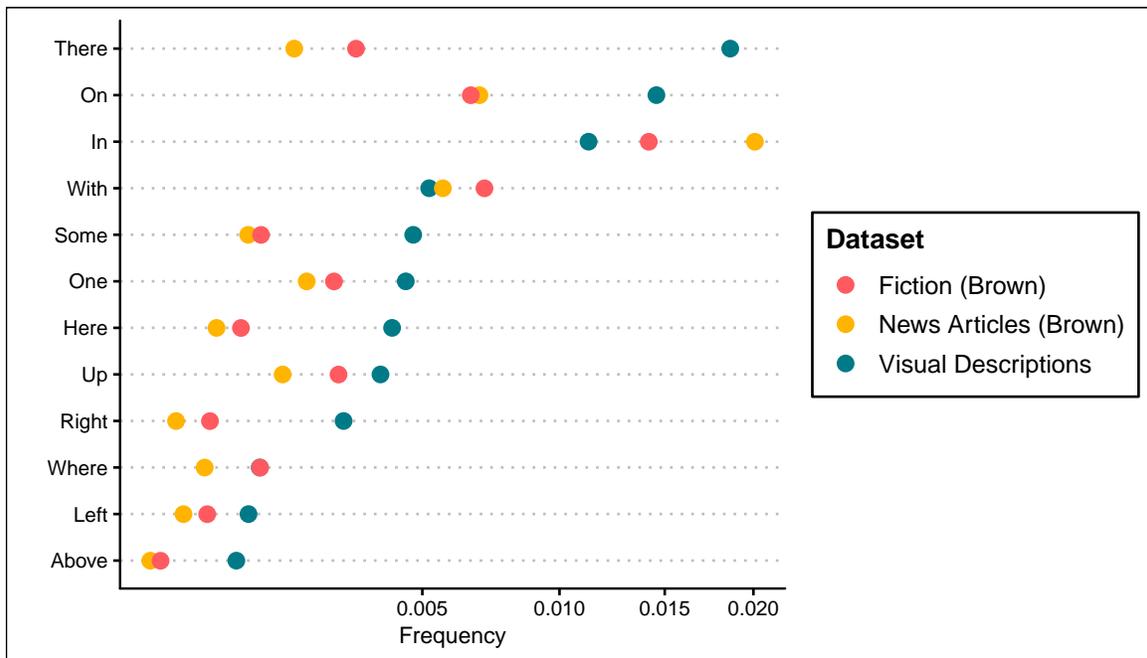


Figure 4-17: Normalized frequencies of the most frequent spatial words in three different corpora. Except for "in" and "with", visual description dataset set has much larger frequency of spatial words.

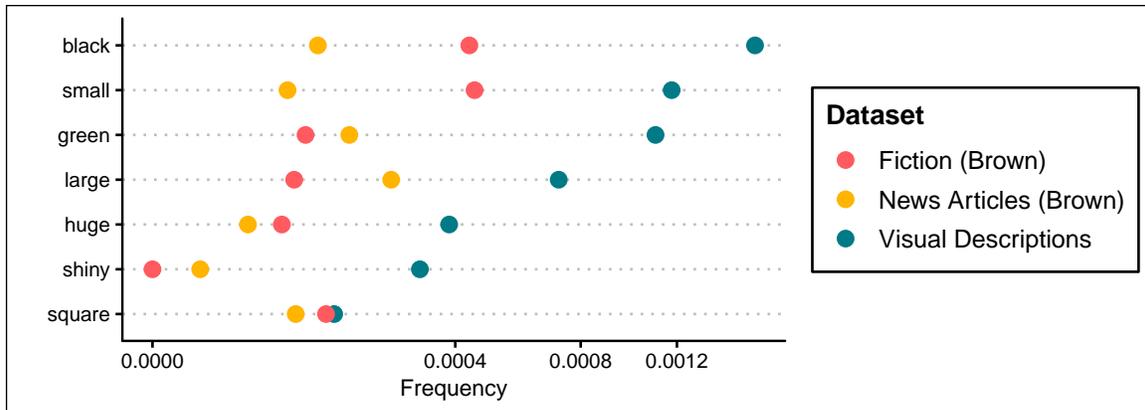


Figure 4-18: Normalized frequencies of the visual words in three different corpora. Visual descriptions contain significantly more visual words.

Summary

The analysis of the language elements used in visual descriptions support the idea that these elements form various relationships between the observed qualities and objects in the environment. In the following sections, I will examine these relationships using the anchoring framework introduced in subsection 4.1.2. I will consider each language element as a result of an anchor that operates on a domain-specific representation. I will examine the anchors under three general categories: *spatial anchors*, which identify spatial relationships among objects and the observer; *temporal anchors*, which relate two or more observations that are made in different times; and *visual anchors*, which identify objects and material qualities. This categorization is by no means complete —one can introduce more detailed categorization. For example, when describing an environment, people make connections to their personal memories, such as in the sentence "This looks like my brother's bike." The reference object in this example is outside of the observed environment and can be described as *a memory anchor*. For this present work, I consider this type of anchor under temporal anchors.

4.3.1 Spatial Anchors

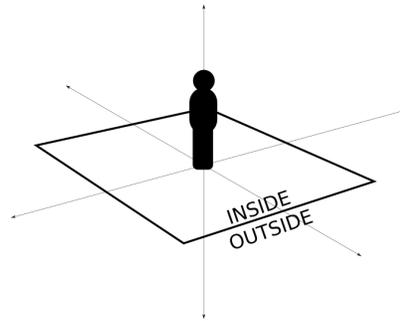
For identifying spatial anchors and related domain-specific representations, I will refer to the literature of spatial semantics (Talmy, 1983; Langacker, 1987; Clark, 1973; Landau and Jackendoff, 1993). Spatial semantics is a subset of semantics that studies spatial language. Different authors offer different framings of spatial language based on classes of expressions, communicative functions, or notional definitions (Zlatev, 2007). I will adopt Leonard Talmy's framing of spatial language, which considers spatial-language as a subset of closed-class grammatical expressions³ (Talmy, 1983). He argues that closed-class expressions operate in certain *domains* such as space, time, and perspective-point, and are limited in what they can express. In fact, they can express only the aspects within those domains, and can be thought of as structural elements (ibid). This definition is parallel to what I argue about the specificity of visual representations that correspond to symbols in language. Table 4.4 shows some of the closed-class words in English, with their functions.

Category	Function	Words
Demonstrative	Point to a noun	Here, There, This, That
Conjunction	Connect words, phrases, and sentences	and, or, with, but, yet, ...
Preposition	Relate nouns (time, location, etc.)	on, under, after, ...
Article	Provide context to noun	a, the
Cardinal Number	Indicate quantity	1,2,3,...
Quantifier	Indicate indefinite or relative quantity	all, most
Ordinal Number	Indicate rank and order	First, second, 5th, ...
Partitive	Indicate part relation to a whole	some of, all of, a lot of, ...
Difference	Indicate difference	another, other

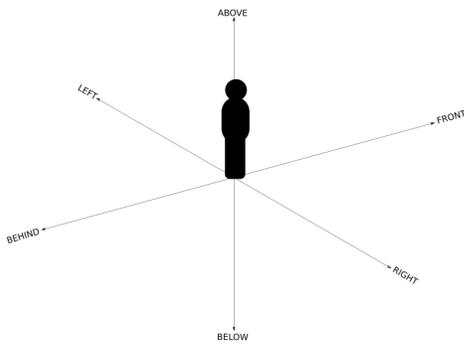
Table 4.4: Closed-class words in English.

Among the closed-class words in English, prepositions are those most frequently found in visual descriptions. In the article "How Language Structures Space," Talmy exclusively examines prepositions such as *on*, *under*, *to the left of*, *behind* and *across*, which he considers the building blocks of spatial language. In my visual exploration study, I discovered that in addition to prepositions, demonstratives,

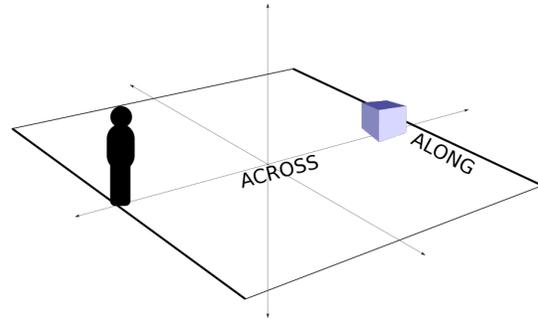
³Open-class refers to any set of elements, such as noun stems, while closed-class refers to the expressions with a small set of elements such as pronouns and prepositions. Closed-class expressions are fixed in membership. (Talmy, 1983)



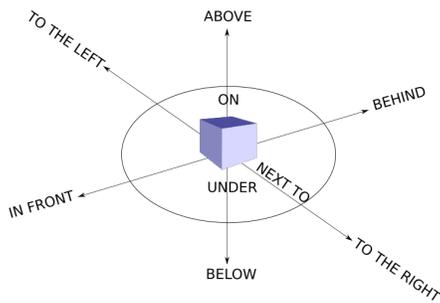
CONTAINMENT



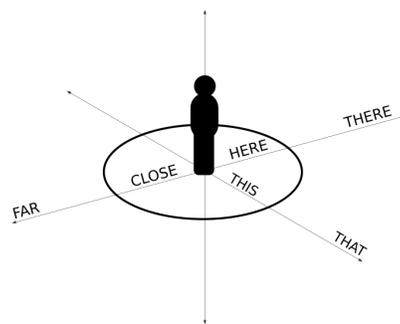
DIRECTION



PATH



LOCATION



DISTANCE

Figure 4-19: Illustrations of spatial anchor domains including containment, direction, distance, location and path.

cardinal and ordinal numbers, and partitives are also used in forming spatial relationships. Furthermore, prepositions themselves have a wide variety of uses, some of which are not related to spatial relations such as *before* and *after*. Among the ones that are related to the spatial relations, I introduce further categorizations for defining specific domains: vertical locations such as *on* or *above*, horizontal locations such as *left of* or *right of*, distance relations such as *front* or *behind*, distance relations such as *here* or *there*, paths such as *across* or *along*, and containment relations such as *inside* or *outside*. Figure 4-19 illustrates these domains and the relationships between various prepositions.

Demonstrative (Distance) Anchors

Demonstrative words such as *here*, *there*, *this* and *that* are considered distance anchors, because the information they use qualitatively discriminates an accessible location (*proximal*) from those that are inaccessible (*distal*). Just like CLOSE and FAR, relations of HERE and THERE to their domain-specific representation are relative. If someone is talking about a series of objects on a table, some of those objects would be HERE and some of them would be THERE. But if the person is talking in the context of the room he or she is in, then the table itself could be HERE and the door could be THERE, or vice versa. The context (the range) of the domain can be indicated in the domain function, such as $DEMOSTRATIVE_{room}$. Figure 4-20 illustrates the relevant information exposed by a domain specific representation for the demonstrative anchors.

Demonstrative anchors are necessarily egocentric, and qualify an object either as proximal or distal relative to the self. These anchors can be *existential*, that is, they can introduce an entity into the inner story. Consider the following anchor functions and corresponding descriptions:

[1] $THIS(x', p_{cup}) \rightarrow This\ is\ a\ cup.$

[2] $THAT(x', p_{table}) \rightarrow That\ is\ a\ table.$

[3] $HERE(x', p_{couch}) \rightarrow Here\ is\ a\ couch.$

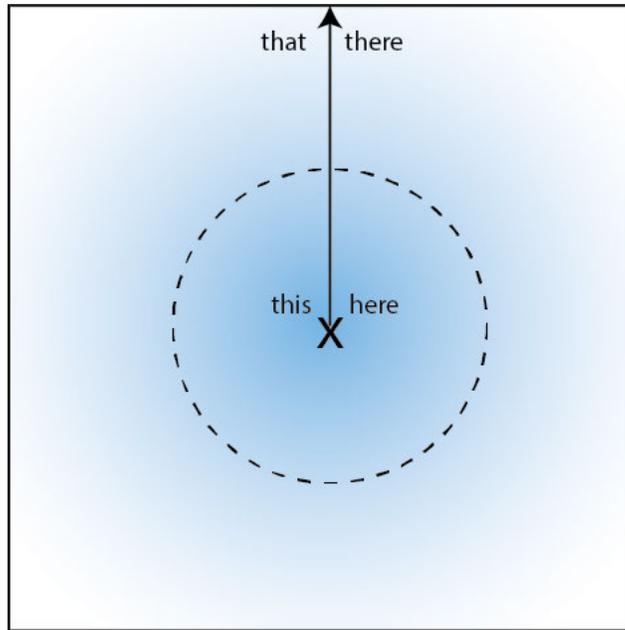


Figure 4-20: An illustration of domain-specific representation for demonstrative anchors. The intensity of the gradient represents the relative distance to the observer, and the dashed circle is a hypothetical division between proximal and distal anchors.

[4] $THERE(x', p_{door}) \rightarrow There\ is\ a\ door.$

In all of these examples, demonstrative anchors introduce a new entity to the story. Demonstrative words are sometimes omitted when introducing an element, such as "I see a window." An observer inevitably perceives the window either here or there, but he or she does not necessarily explicitly state this. Sometimes this omission introduces a challenge for analyzing stories, because an element without a connection to other objects or the self is essentially detached from the rest of the story. I will introduce a way to include an unanchored element in the story via temporal anchors.

Figure 4-21 shows a few examples of demonstrative anchors obtained in the study. A domain-specific representation is generated by assigning distance values to each pixel on the image. The distance values are obtained from the three dimensional model data. The descriptions corresponding to each image are as follows:

[1] And, there's a light **here**.

[2] On top of the bookcase over **here**, we got some paintings.

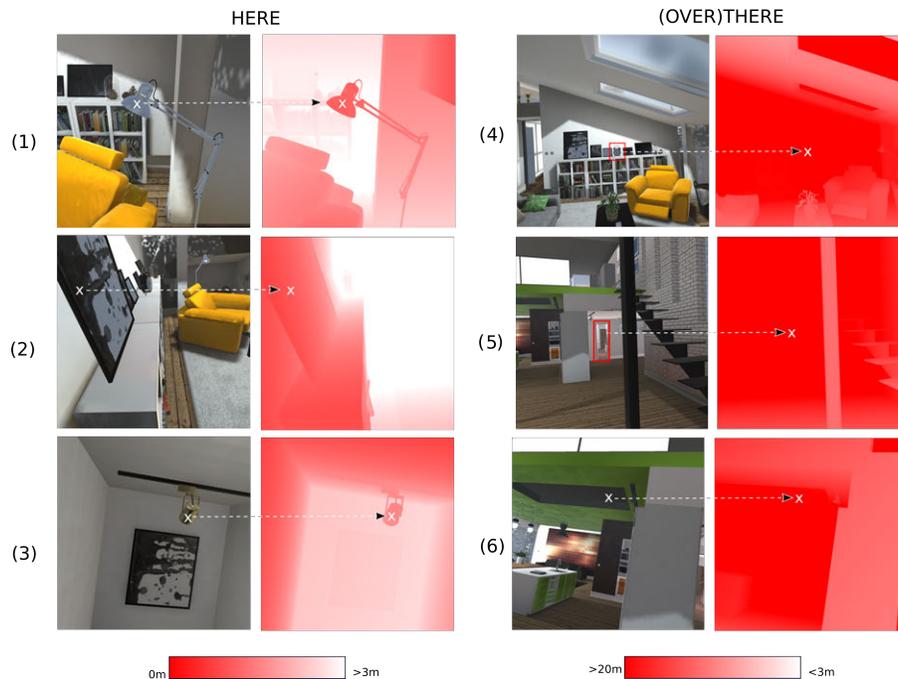


Figure 4-21: Domain-specific representations for demonstrative anchors from the study.

[3] A little light mounted in the ceiling **here**.

[4] You can use **that** to set alarms to read the time, to know what time it is

[5] Oh, is **that** room over **there**?

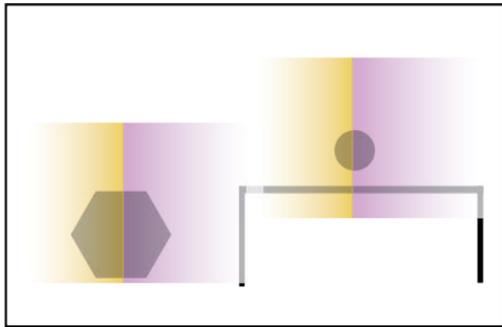
[6] ... if I go up **there** at some point, it would make more sense for me why it looks like that.

Location Anchors

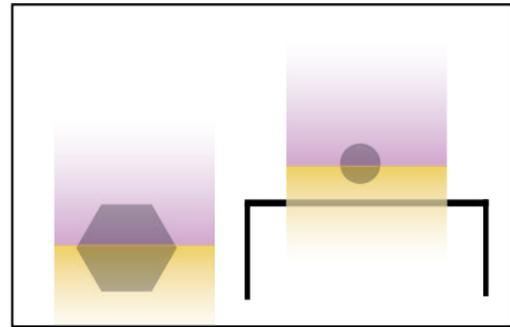
While closed-class demonstrative elements establish a location relative to the observer, they do not establish a location for the observer relative to the environment. A participant often made reference to where she or he was by either identifying a distance relationship with an object in the environment such as "I am near the front door" or by stating a containment relationship within a distinct part of the environment such as "I am in the kitchen". The function of these types of anchors are similar to that of demonstrative anchors in the sense that the observer identifies the part of the



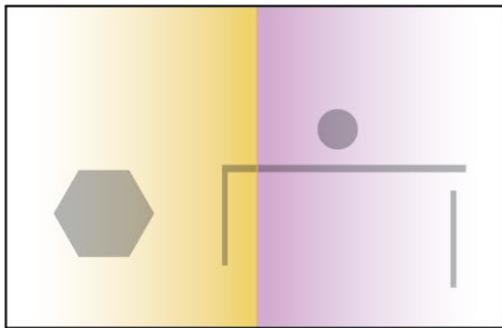
Figure 4-22: Names of rooms mapped on the locations in the Environment where participants verbalized them. While most of the time a participant verbalized the name of a room when she or he was in the vicinity of it, there were also alternative cases. For example, a subject made a reference to the windows of the living room when she was in the bathroom.



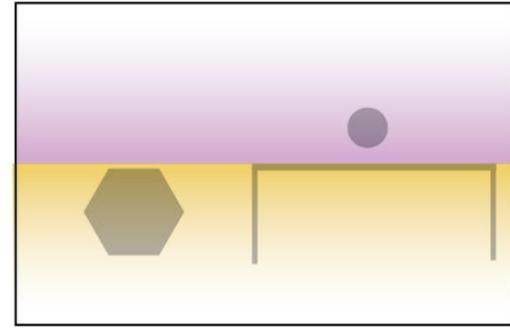
(a)



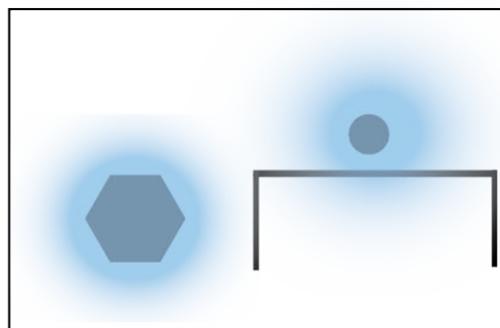
(b)



(c)



(d)



(e)

Figure 4-23: An illustration of domain-specific representations for proximity anchors. (a) Object centered horizontal, (b) object centered vertical, (c) egocentric horizontal (d) egocentric vertical (e) omnidirectional

environment she or he has access to. A mapping of the location names verbalized by the participants to the locations in the environment demonstrates that these words acts as location anchors for the observer, illustrated by Figure 4-22.

Proximity Anchors

The most frequently used type of spatial anchors are those indicate a spatial proximity between two objects. These anchors include attachment relations such as ON or AT, vertical relationships such as ABOVE or UNDER, horizontal relationships such as LEFT OF, and indeterminate proximity relationships such as NEAR, BY and AROUND. All proximity anchors take two object arguments for primary and secondary objects that are in a spatial relation. Below are some examples of proximity anchors:

[1] $ON(x', p_{cup}, p_{table}) \rightarrow$ *The cup is on (top of) the table.*

[2] $AROUND(x', p_{lights}, p_{mirror}) \rightarrow$ *There are lights around the mirror.*⁴

[3] $BEHIND(x', p_{window}, p_{couch}) \rightarrow$ *The window is behind the couch.*

[4] $RIGHT(x', p_{painting}, p_{stairs}) \rightarrow$ *The painting is to the right of the stairs.*

While all proximity anchors have similar functions, they have different domain-specific representations. The most common types are vertical relationships, attachment relationships, horizontal relationships, depth (front-behind) relationships, and indeterminate nearness relationships such as by, near, and around.⁵Figure 4-23 provides a few illustrations for proximity anchors.

Figure 4-24 and Figure 4-25 illustrate a few examples of domain-specific representations for vertical relationships and attachment relationships. For the vertical relationships of ABOVE and BELOW, I generated a domain-specific representation by assigning a value to each pixel in the image corresponding to the

⁴The word *lights* is plural, which means there are multiple lights around the mirror. This property is explained in grouping anchors section.

⁵Around has two different uses. The first case occurs when around is used as a synonym for near. The second case indicates that one object surrounds the other. Here I consider around as a synonym of near.

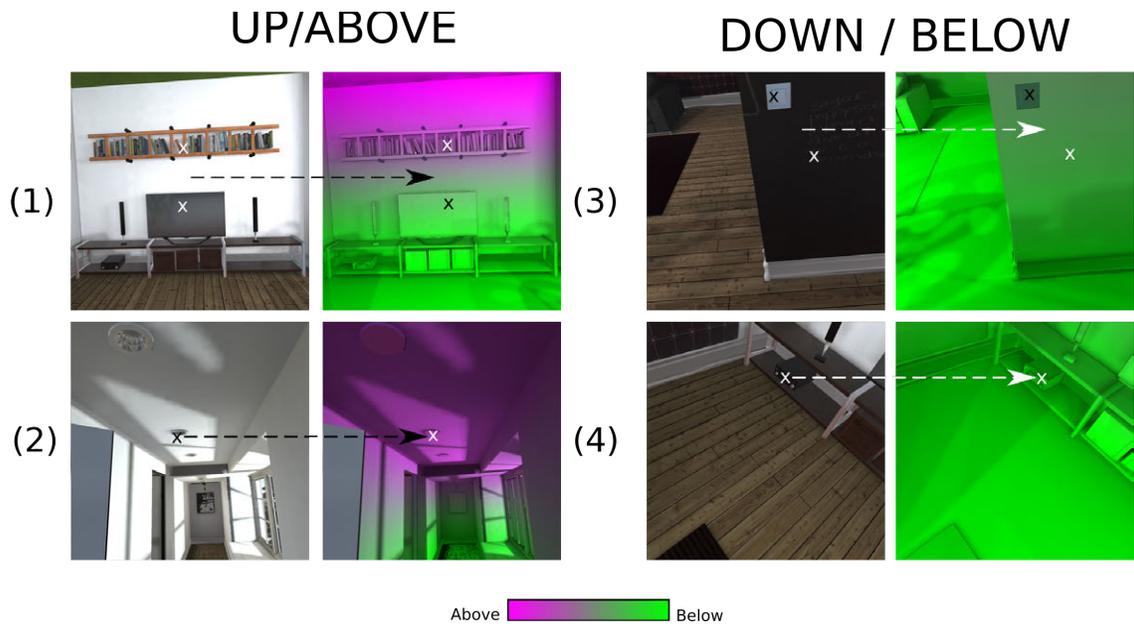


Figure 4-24: Domain-specific representation for vertical relations

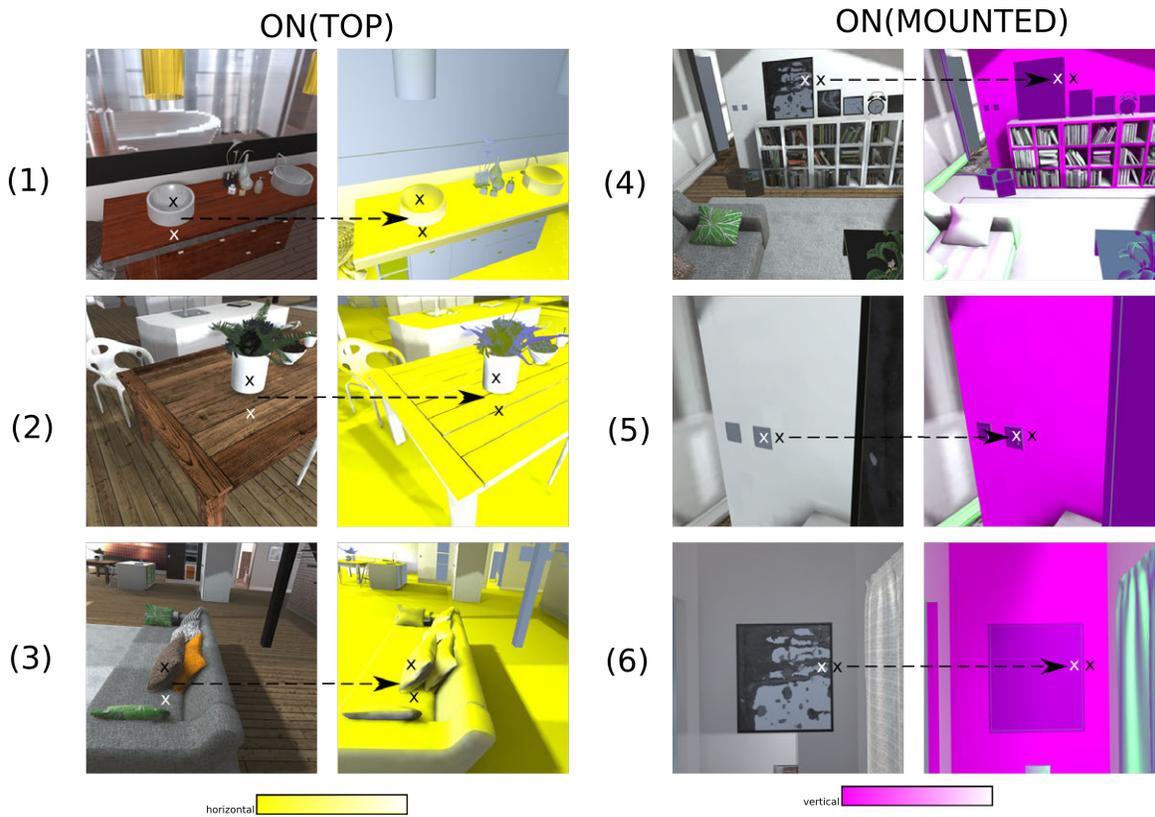


Figure 4-25: Domain-specific representation for attachment relations

relative height of the corresponding point in the environment. This representation makes evident which of these pixels may generate an ABOVE (purple) or BELOW (green) relationship. Additionally, pixel values in this representation can be compared to each other, the higher of which generates an ABOVE relationship.

Similarly, for the attachment relations, I generated a domain-specific representation by assigning a value to each pixel, such that each pixel represents the relative orientation of the surface that generated that pixel. The horizontal surfaces indicate an ON(TOP) relationship (yellow) while the vertical surfaces indicate an ON(MOUNTED) relationship (purple). Below are the corresponding descriptions for each image in the figures:

ABOVE/BELOW (Figure 4-24):

- [1] Um, and then, **above** the TV there's, like, this ladder that's on its side.
- [2] I see lights **above** all along the hall to light up the hall when it's dark.
- [3] There's another list **below** next to a light switch.
- [4] There's a little black box **down there** that looks like a VHS player.

ON(TOP)/ON(MOUNTED)(Figure 4-25):

- [1] And **on top of** it are two bowls
- [2] Um I don't know if I mentioned that there are plants **on** the table
- [3] This gray couch has a bunch of pillows **on** it
- [4] I see a painting **on** the wall.
- [5] I see a switch **on** the wall, turn the lights on and off.
- [6] I go further in, I see another painting **on** the wall.



Figure 4-26: Grouping anchors.

Grouping Anchors

Grouping anchors combine multiple objects into a single entity. The most common grouping anchors are counting anchors, both indefinite quantities including plurals, *a bunch of*, *a few*, and cardinal numbers. In the sentence "There are **THREE** chairs around the table," the group of chairs are anchored together with the cardinal number three. While there is no clear domain-specific representation for grouping anchors, grouped entities usually have some shared properties and are in proximity to each other. In the observation study, I observed 497 instances of counting. In most of those cases, the participants directly referred to the grouped entities such as "two boxes," and in some cases they explicitly counted individual entities. Below are a few example statements that involve grouping, corresponding to the images in Figure 4-26.

- [1] Uh, and underneath on the second level of the TV stand, we have these **three random wooden boxes**.
- [2] Um, there's **two perfume bottles** that are square.
- [3] This is a table with, let's see, **nine chairs** at it, and they're all made of wood and metal.

Partitive Anchors

Closely related to grouping, partitive anchors refer to a member of a group of entities. Implicitly, a partitive anchor acts as a grouping anchor. Consider the sentence "One of the chairs around the table is plastic." This time, the difference in one member of

the group enables it to be partitively anchored to the group.

4.3.2 Visual Anchors

Visual anchors enable identifying and describing objects and materials. In the beginning of this chapter, I discussed the identification of material qualities. I argued that there may be many different visual dimensions that collectively make up our perception of materials. This problem is inherently difficult because there are so many words that describe the types of materials and material properties. Developing a system with a full coverage of the domain of materials is still a challenge because one needs to consider all possible visual dimensions that contribute to the perception of materials. A system with an anchoring framework nevertheless can learn and use a limited set of domain-specific representations that help identify a series of material properties, even when those domains are incomplete. For example, material properties of reflectivity, smoothness, or opacity can be learned and used for describing a series of materials. Consider Figure 4-27 for illustrative purposes. In this figure, the image at the center is produced from a variety of *render buffers*, represented in smaller boxes around the final image. Among those buffers are metallic, which represent the level of metallic qualities of a material, and opacity. A vision system that can map the final image into those buffers can use those buffers as material domains for describing material qualities.

In the beginning of this chapter, I introduced a hypothetical OBJECT domain composed of objects like cups and tables for illustrative purposes. However, similar to the case of materials, a single domain-specific representation is less likely to cover the range of objects we can perceive. Objects have a wide range of visual and functional properties, and there might be many domains that differentiate one type of object property from others. For example, some objects afford seating, some objects are mobile, and some hold liquids. While enabling machines to learn these domains is an interesting problem, it is beyond the scope of the current work. I will instead use the state-of-the-art DNNs that perform object detection for a wide range of objects found in everyday environments.

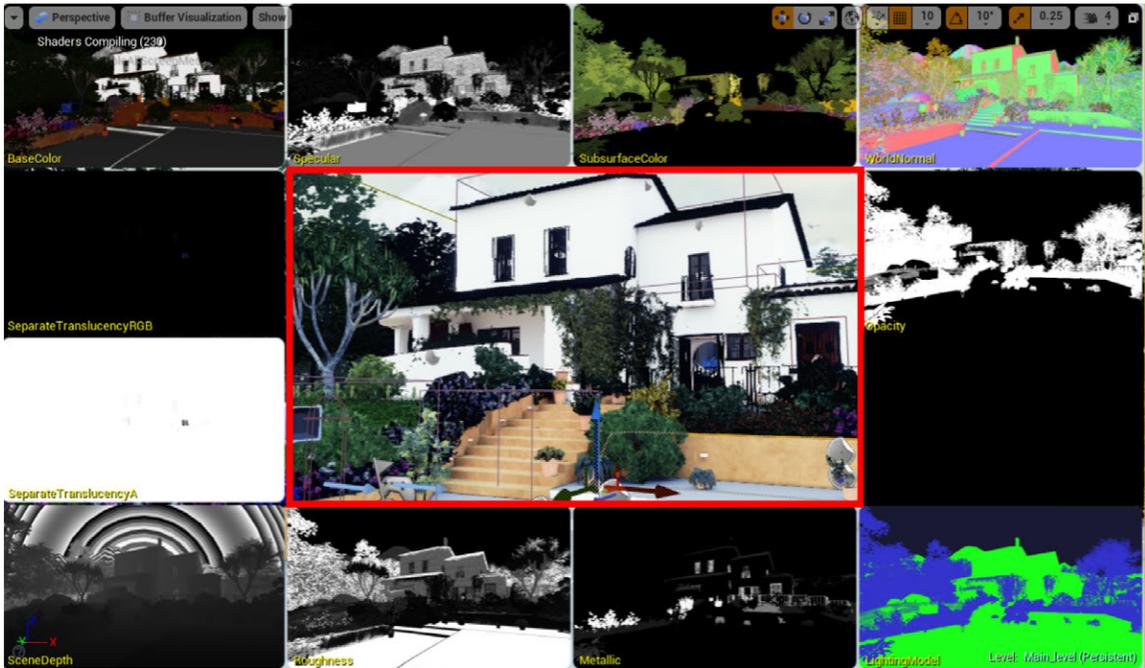


Figure 4-27: An illustration of material domains. This figure obtained from Unreal Game Engine shows a variety of render buffers, which collectively produce the final render at the center of the figure. Among those buffers are metallic, opacity and surface color. A vision system that can map the final image into those buffers can use them as domain-specific representations to describe materials qualities.

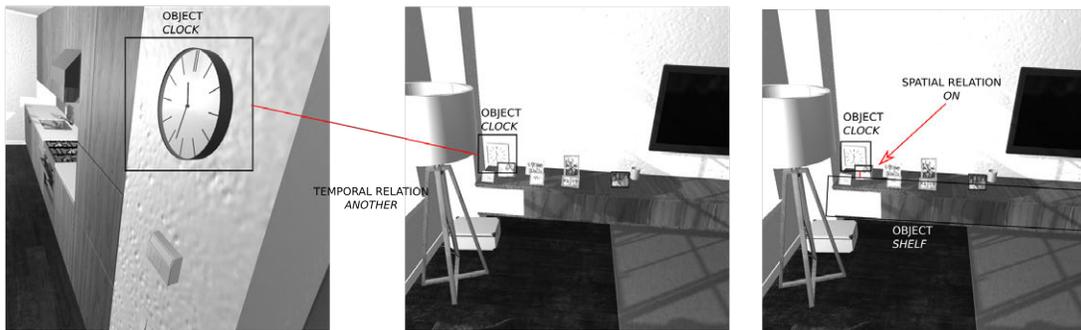


Figure 4-28: Temporal anchors relate a present observation to a previous one. This figure illustrates the ANOTHER (TEMPORAL) anchor, which contributes to the description "There is another clock on the shelf."

4.3.3 Temporal Anchors

The idea of temporal anchors emerged as a result of the observation study. Almost all of the participants (24/26) made an observation in which the object presently under observation was compared or related to an earlier one. The most frequent words used for making these observations were *another*, *more*, *different*, and *similar*. Anchoring a present observation to an earlier one enables a person to make connections between two elements when there is no direct spatial or visual relationship between them. On average, participants made five such connections during their explorations, most of them were comparisons between objects based on their identities such as "another clock" or their material qualities such as "a different green." Figure 4-28 shows an example from the study where a participant observed "another clock on the shelf." This observation followed an earlier observation "there is a clock on the wall," therefore the word *another* temporally connects these two observations.

4.4 Summary

In this chapter I introduced a novel approach to human spatial experience: I framed spatial experience as an inner story that we tell ourselves to interact with and understand our environment. In this context, I defined an inner story as a chain of observations that are connected to each other based on the spatial, temporal, and visual relationships among them.

The idea that an inner story acts as a binding mechanism among various language-independent spatial representations is supported by the cognitive studies I introduced in the beginning of the chapter. This motivated another question: How are these language-independent representations combined via inner stories? This is a problem that requires understanding how symbols (in our case, symbols in language) are connected to those language-independent representations. I framed this process as anchoring and introduced a computational mechanism that can generate symbolic descriptions (inner stories) from visual, spatial, and temporal information.

With the anchoring framework I proposed that symbols in language refer to locations in domain-specific representations. For example *close* and *far* refer to locations in a depth representation, and *red* and *green* refer to locations in a color representations. Having these representations allows us to answer questions about the presence of some environmental factor (distance or color) unambiguously, and to understand how two symbols in the same domain are related to each other.

My visual exploration study illustrated that the anchoring framework can model a large portion of the verbal descriptions made by the participants. In these descriptions, the frequency of the words that relate to an environmental factor were significantly higher than those in other corpora. Words that indicated spatial relationships, such as *here*, *there*, *on*, and *above* were notably abundant in the dataset. I grouped those words under three categories and examined each of them in relation to corresponding visual information: spatial anchors, which create spatial relationships, temporal anchors, which create temporal relationships, and visual anchors, which identify object and material qualities. I also provided a few examples of domain-specific representations that I generated from the 3D model. I suggest that those representations can be learned by a computer vision system when the required information is not available (such as depth).

In the next chapter, I will focus on two specific problems related to the anchoring framework in the context of artificial intelligence: learning domain-specific representations and connecting vision systems to symbolic systems with anchors.

4.5 Contributions

My contributions in this chapter are as follows:

- I introduced the anchoring framework. Building on Patrick Winston's Strong Story Hypothesis, I introduced the anchoring hypothesis, which states that spatial experiences are inner stories that we tell ourselves to interact with and understand our environment. I further suggested that this mechanism requires connecting perceptual features of environments to cognitive processes.

- I developed a methodology I call "See, Act and Tell" in which the participants explored and verbally described virtual environments. I analyzed language elements in relation to observations and showed that the verbal descriptions include a significantly high number of words that relate to environmental factors and establish spatial, temporal and visual connections.
- As a result of this study, I identified various anchors —functions that generate symbolic relationships from visual information. Anchors are related to property-specific representations that make evident a certain type of information. I developed a computational method for producing descriptions using these anchors and domain-specific representations. I showed that I can model a large portion of verbal descriptions using a limited set of anchors.
- I created an observation dataset that includes 6.5 hours of explorations in three different virtual environments. The dataset includes camera parameters, voice recordings, and timestamped speech transcripts (both at phoneme and word level). This dataset can be used for training artificial agents and vision systems. In a previous study, Gilpin et al. (2018) showed that it can be used as a ground truth for image captioning tasks.

Chapter 5

Implementing the Anchoring Framework for Developing Spatial Experience in Machines

In this chapter, I will present the steps I took towards developing artificial intelligence systems that can learn, understand, and communicate their spatial experiences by composing inner stories and relating their observations in terms of those stories. I will show that the anchoring framework enables an AI system to make connections between symbolic and perceptual systems in order to solve visual problems and generate verbal descriptions from images. There are two sub-problems involved in generating inner stories with anchors: learning domain-specific representations and generating anchors. Learning domain-specific representations is essential to an anchoring framework because it allows the system to isolate an environmental factor and answer questions regarding that factor unambiguously. Only after having a domain-specific representation can we develop specific anchor functions that generate symbolic descriptions regarding an environmental factor.

First, I will provide an example of learning a domain-specific representation in the context of spatial navigation. I will show that a robot can learn to extract information regarding its distance to environmental boundaries and its relative orientation in space from a camera image, using additional sensory information provided in the training

time. The resulting representations helps the robot to identify nearby boundaries and avoid them, while at the same time keeping track of its heading direction, represented by the cardinal directions of North, South, East and West.

I will then introduce my experiments on connecting symbolic systems with visual systems. I will introduce two examples, both of which use the Genesis Story Understanding System (in short, Genesis) (Winston, 2014) for symbolic reasoning and problem solving. In the first example, a robot replaces a cell-phone battery following instructions provided in simple English. In the second example, Genesis generates interpretable verbal descriptions from visual explorations, augments those descriptions with common-sense knowledge, and answers natural language questions regarding its observations. Both of these examples use the anchoring framework for connecting the visual information with inner story representations.

5.1 Learning Domain-Specific Representations

In the previous chapter, I provided a series of domain-specific representations, such as depth maps and vertical relationship representations, which made evident a particular type of environmental information and enabled the construction of symbolic descriptions. Those representations were constructed by using the three-dimensional model that was available in my dataset. However, in real-world scenarios, this type of information is generally unavailable, and we might only have access to two-dimensional image data. In those cases, we need to learn mappings from an image to domain-specific representations so that we can use anchors to generate descriptions. In this section, I will provide one example of such mapping in the context of spatial navigation. The task in this example is to identify strictly from images the environmental boundaries and the relative orientation of an agent navigating in the environment. This task is similar to a SLAM (Simultaneous Localization and Mapping) task, in which an agent constantly builds a map and locates itself within it. The difference between that and the current task is that the current task requires building a representation with which we can generate symbolic

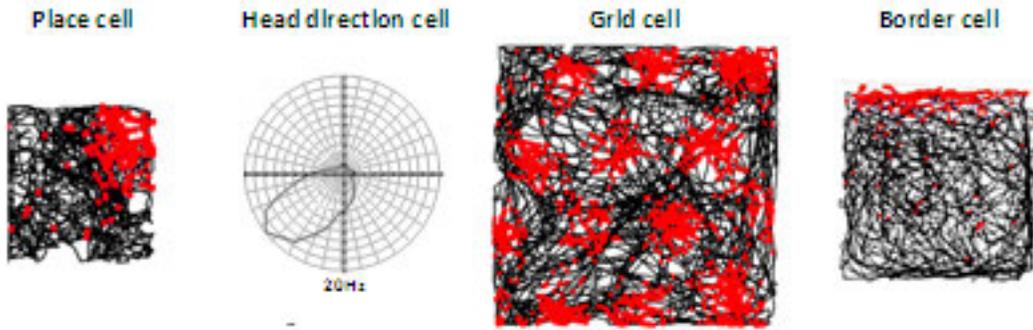


Figure 5-1: The spiking patterns of spatial cells of a rodent brain in a square environment. A place cell has a "place field" and is activated only in a particular part of the environment. A head direction cell is activated only when the rodent is oriented towards a particular direction in the environment. A grid cell has an activation pattern that has a particular distance interval that builds a "grid" when recorded in multiple locations. A boundary cell (border cell) is activated when the rodent is near a boundary that lies in a particular direction. In this example, the boundary cell is activated near the north wall of the room.

descriptions regarding distances and orientations.

Inspired by the studies on neural computations that are involved in spatial awareness introduced in chapter 2, I introduce a model that learns boundaries and orientations from images. This information is then integrated into "place" representations that enable the AI system to identify distinct locations in an environment. Spatial representations in the rodent brain includes place cells, boundary cells, head-direction cells and grid cells (Figure 5-1). I introduce a neural network architecture that can learn these representations from images. The project described in this section was developed in collaboration with Ege Ozgirin and Bilge Zeren Aksu. From now on, I will use the pronoun "we" to refer to myself and these collaborators.

5.1.1 Background

Visual space learning poses an important challenge for developing artificial systems. In a previous study, Oliva et.al. showed that perception of a scene relies on several components, including low-level visual features, previous experience about the

environment, and spatio-temporal history in the environment (Oliva et al., 2011). Perceptual information regarding a scene can be characterized as the spatial-envelope, an object level description of scene geometry that includes openness, roughness, perspective, and volume (Oliva and Torralba, 2001). Integrating these features into a place representation requires bridging multiple descriptions obtained from an active exploration of space. We are also inspired by a few successful implementations of Simultaneous Localization and Mapping (SLAM) methods, notably the RatSLAM (Milford et al., 2004), which uses a place-cell inspired model for generating maps in novel environments. RatSLAM uses place cell representations to assign camera images to unique places where the linear motion and orientation information is extracted from images through an optical flow method.

In this project, we used a deep learning architecture to extract boundary cell activations from camera images. The main difference of this approach from other SLAM models is that it does not generate holistic maps, topological representations, or geometric models. Instead, in any given time, we predicted boundary cell activations using only the provided camera image. This method allows efficient decision-making for robot navigation, as CNN tends to produce predictions very robustly. We tested our system for two tasks. First, we used boundary cell predictions to decide the most plausible path for the robot to take in the next frame so that it avoids obstacles. Results from three testing environments showed that our method allowed the robot to successfully avoid obstacles and randomly explore the environment without supervision. Second, we used boundary activations to determine unique locations in the environment. We showed that the predicted activations of our model are selective to obstacles in particular allocentric orientations,(e.g. north-facing walls), which presented similar characteristics to the recordings from boundary cells in the rodent brain.

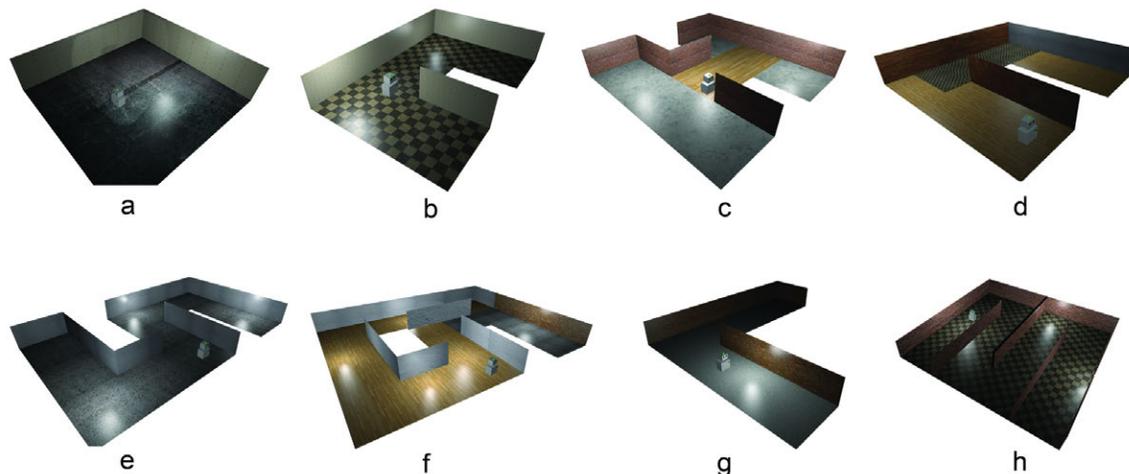


Figure 5-2: Our model was trained and tested in eight different simulation environments. Five environments were selected for training while three of them were used for testing. The test environments included novel textures that were not included in training environments.

5.1.2 Methodology

As a first step, we created a robot simulation environment with the MORSE library in Python/Blender (Echeverria et al., 2012). We designed five environments for training and three environments for testing (Figure 5-2). Our simulated robot (the agent) was supplied with an onboard camera and a laser scanner (SICKLMS500). We wrote several utility functions to form appropriate representations that are isomorphic to the spatial representations —boundary cells, head direction cells and grid cells in rodents. We created four different models by implementing a CNN architecture with four different types of label vectors. The first three types are 8, 12, and 24 dimensional vectors respectively, representing multiple boundary cell activations. The fourth label type is a 24-dimensional vector representing 24 grid cell activations. We trained our model separately with boundary cells and grid cells.

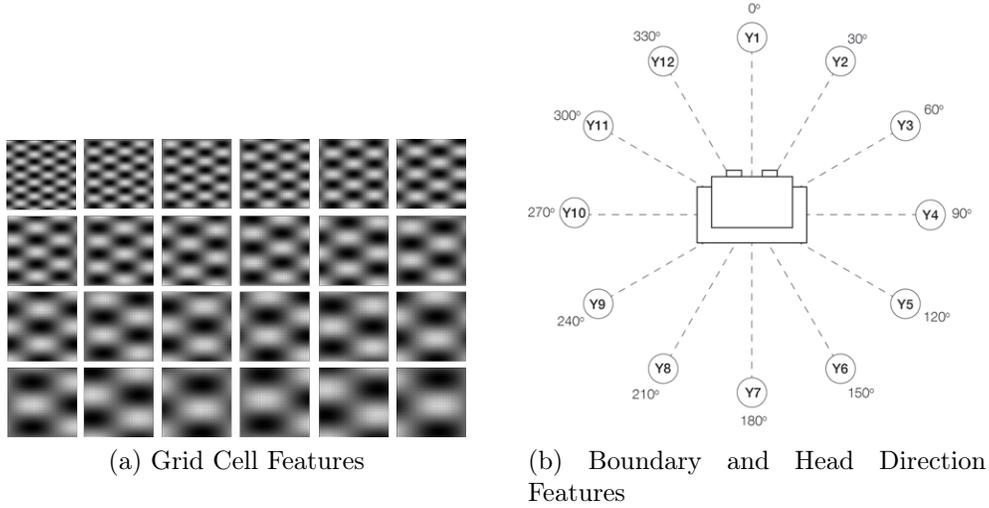


Figure 5-3: Feature Representations. a) Grid maps are used to train the system to predict grid activations. There are 24 different grids, each with a different phase and frequency. The final grid feature G was calculated with a function $F(G_t, x, y)$ where $t \in [1..d]$ and x and y are real valued positions obtained from simulation. b) Head Direction and Boundary features are selective to orientations. Activations of these cells were calculated with functions $F(B_t, \alpha)$ and $F(H_t, \alpha)$ where $t \in [0..d]$ and $\alpha = t * 2\pi \div d$

5.1.3 Spatial Cell Representations

Boundary Cells We used the output values of the SICKLMS500 laser scanner model to create boundary cell activations. The data from the sensor is provided in an array that stores the distance to the first obstacle received from each ray. Boundary cell activation values are generated by mapping the laser-scanner output between 1 and 0. The values are close to 1 if the agent is in close proximity to a wall and close to 0 if it is distant from a wall. The value of a boundary feature B is given by:

$$F(B_t, \alpha) = \begin{cases} 1 & \text{for } L(\alpha) \leq \sqrt{k} \\ k/L(\alpha)^2 & \text{otherwise} \end{cases}$$

where $L(\alpha)$ is a function that gives a distance reading from the laser scanner from the direction α and k is a constant value for the minimum distance for a cell to reach the maximum activation.

Direction Cells

The agent's pose could be gathered by using a pose sensor providing the rotation values around the axes of the sensor. Head direction cell activations are initially mapped from the radian values received from the sensor to the cartesian values. We calculated a gradient for each of the 12 cells by representing the most active cell with a value closer to 1 and assigning the less active cells with lower values that are closer to 0 relative to the most active cell.

Grid Cells

Finally, 24 grid cell activations are generated by computing a gradient of frequency and phase values that are consistent with biological data. 24 different grid cells are triggered only when the agent passes through grid activation zones in the environment based on each cell's grid spacing and phase shift parameters that vary between 10 cm to 100 cm. Again, the output values of grid cells are designed to be real values between 0 to 1, where 0 means the cell is dormant and 1 means the cell spikes at the maximum rate. Activation of grid features are calculated by:

$$G_t(F, P, x, y) = 0.5 * \sin(P_t + 2 * \pi * x / F_t) * \sin(P_t + 2 * \pi * y / F_t)$$

where phase delay $P_t \in [0.. \pi]$ and grid frequency $F_t \in [0.1..1]$ (Metric units between 0.1m to 1m).

Making the location information $[x,y]$ available for the grid cells does not contradict with the idea of inferring place information, as current biological evidence supports the fact that grid cells are innervated with neurons carrying proprioceptive information. The question of how the location information becomes available for the grid cells is particularly interesting, but it is outside of the scope of this study.

5.1.4 Model

We developed a learning model for the agent to identify unique places in a given environment. Our aim was to make the agent rely only on the visual input to 1) determine its relative location in the environment and 2) navigate to goal locations. Our system employs a custom CNN architecture in which the output layer produces a vector Vs that represents the activation of spatial layers - boundary cells, head direction cells and grid cells. Our CNN architecture consists of two consecutive convolution and max pooling layers. Convolution layers use 32 3×3 filters and ReLU activations. Max pooling layers have the size 2×2 . Finally, we applied two fully connected layers before the output layer. The output layer dimension varies according to the training variation. We tested $d = 8, d = 12, d = 24$ for the dimension of feature vector Y in three different training sessions. In all sessions, we used sigmoid activation for the output layers and root mean squared objective function.

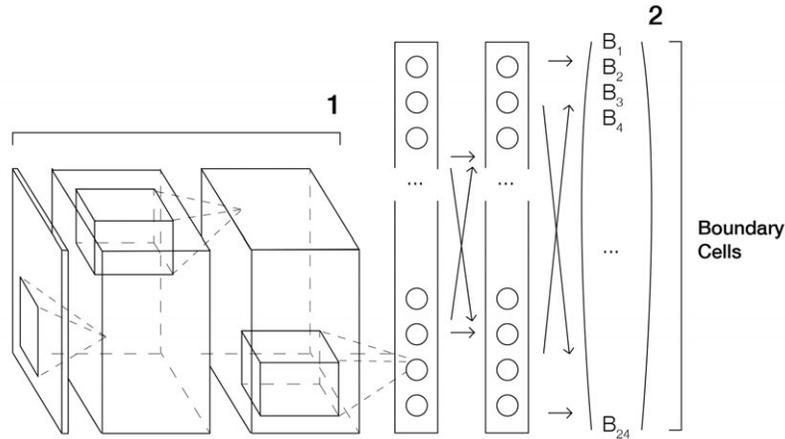


Figure 5-4: CNN Architecture for predicting Boundary Cell Activations. This model produces a vector of values in range $[0,1]$ that correspond to the level of activity for an individual boundary cell for a particular direction.

All labels were real valued numbers such that $Y_i \in [0,1]$, where each label Y_i corresponded to the level of activation of a particular cell. Each feature vector Y was independently generated and does not correspond to a predetermined class.

Therefore, unlike CNNs that perform classification tasks, probabilistic representations such as *softmax* were not suitable for our purposes. As our feature vector Y is a random vector that is continuously generated from the environment during the online training phase, we decided to use mean squared error for the loss function. Additionally, we experimented with different CNN architectures and learning parameter values.

The model is trained online with camera images that are streamed from the simulation environment with a rate of 2 frames per second. Images are converted to $256 \times 256 \times 3$ RGB format and normalized to floating point values such that $X \in [0, 1]$. In each frame, label values are calculated from the sensors, and CNN is updated with these labels and given the camera image. We performed two turns of back-propagation in each update.

5.1.5 Training

We trained the system in 5 different environments with the label vectors composed of 8, 12, and 24 boundary cells and 24 grid cells, respectively, in four different settings. We trained the model in each simulation environment for three minutes, for three epochs. In the first two epochs, the model was trained while the agent navigated in the environment using distance data provided by the laser-scanner. In the third cycle, the model was continued to be trained while the agent navigated based on the

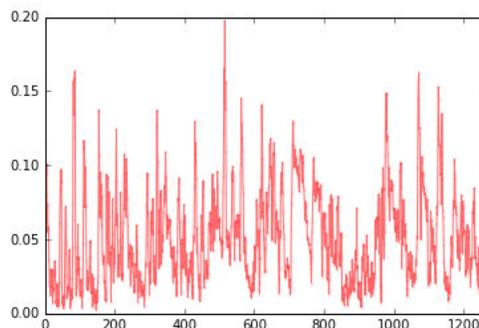


Figure 5-5: Loss value after each sample over the all training sessions. The final model has fluctuating loss values between 0.02 and 0.05.

prediction of the model. In the last cycle, we interfered with the process whenever the robot failed to detect a boundary and hit the wall.

The training procedure was performed consecutively in each environment in each cycle. We changed the environment every three minutes and continued training the network. Because there was no way of presenting randomly selected samples to the network as in an ideal scenario, we believed that altering the environments between cycles would allow us to prevent over-fitting to a particular setting and would provide a better generalization. Figure 5-5 shows the loss value after each sample over the all training sessions. We observed the loss value tends to get closer to a value 0.02 while fluctuating up to 0.08.

5.1.6 Results

Prediction

We trained three different networks using only boundary features of size $d = 8, d = 12, d = 24$. Each network was tested in three novel environments that were not included in the training sessions. The network trained with $d=24$ failed to avoid obstacles and presented a very poor performance for predicting boundary activation. This model was excluded from the analysis because the robot failed to navigate the environment and could not produce any prediction. Networks with $d = 8$ and $d = 12$ successfully navigated in the test environments. Please refer to this video for a sample test session.¹

In order to compare the performance of the models in terms of predicting environmental directions, we visualized the error rate for each direction in the text environment. Figure 5-6 shows a comparison of robot trails in two models, $d=8$ and $d=12$ for North, South, East and West selective cell activations predicted by the system. The figure shows that the 8-cell model (A) made a better discrimination

¹The video file can be found at: <https://vimeo.com/166680547>

then the 12-cell model(B). Both models failed to make correct predictions in certain parts of the environment, but the 8-cell model is significantly more accurate.

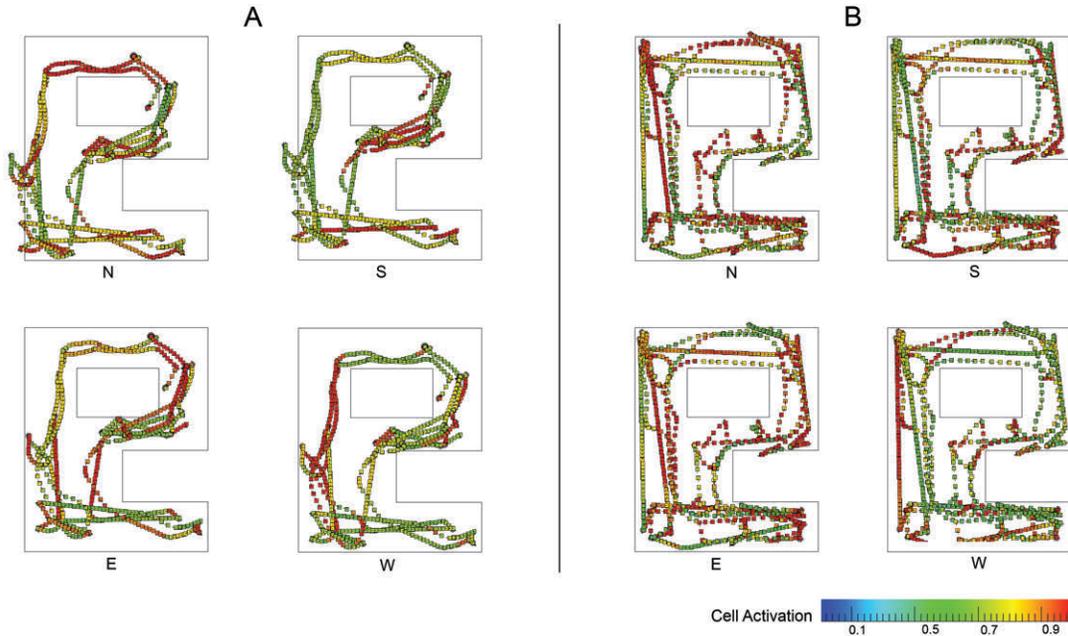


Figure 5-6: A. 8-Cell Model predictions for four different directions. It can be seen that activations are higher around the borders for the indicated directions. B. 12-Cell Model predictions. This model shows less accuracy for predicting activation values.

Figure 5-7 shows the loss value of predictions for two different training environments. We observed that the loss value becomes higher around the borders. We believe there are two possible explanations for this behavior. Because the feature values are generally smaller in central areas, the error in prediction should be getting relatively smaller as well. Around the borders there are more cells with maximal value which should produce a higher L2 distance between actual and predicted values. Additionally, the camera image changes rapidly near the borders because of the distance and gets more homogeneous, especially if the robot is facing the wall. The system should be more prone to errors when the agent is near the boundaries. We believe a better distance rating should be used for addressing this issue.

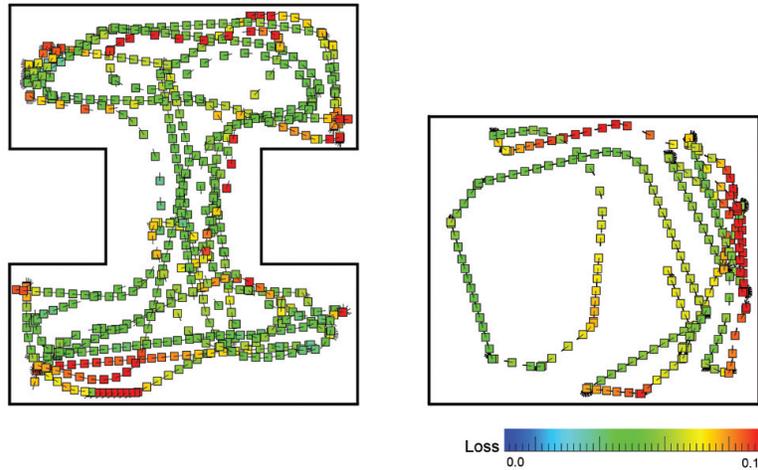


Figure 5-7: The training loss for the predicted activation values in two different training environments. The training loss increases near the borders.

Grid Predictions

Finally, we tested a model we trained with 24 grid features to see if we could predict grid cell activations with the same model. Figure 5-8 shows three different grid cell predictions from the model that is trained for 15 minutes in a square environment. As can be seen on the figure, the predictions did not converge to required grid patterns, although there are partial repetitions in the spike pattern. We found this poor result intuitive as grid pattern is orientation invariant and mostly depend on proprioceptive signals rather than visual ones.



Figure 5-8: . Predicted activations for 3 different grid features in a square environment. Refer to Figure 5-3 or corresponding input grid maps. The predictions do not reflect the expected patterns although they present some grid-like regularity.

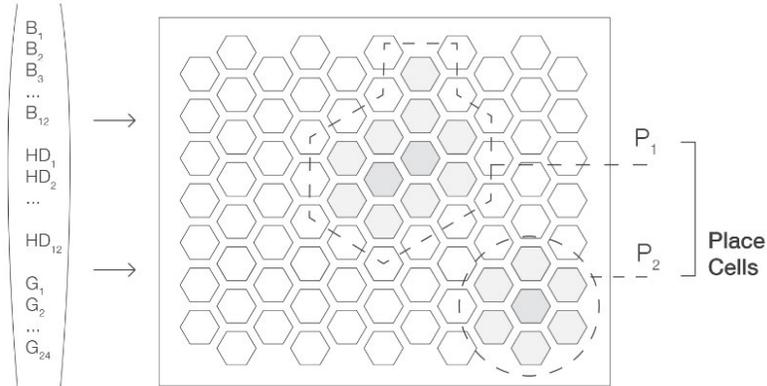


Figure 5-9: A Self Organizing Map (SOM) can be used to generate unique place descriptions from boundary, grid and head direction cells.

5.1.7 Using Self Organizing Maps as Place Descriptors

In the first part of section 5.1, I showed that boundary representations can be learned and used for guiding navigation. In this section, I will further explore the capacity of these representations for learning distinct parts of the environments, as has been observed in biological place-cells. For this purpose, I decided to train another neural network that can encode the boundary representations, introduced in the previous section, in a Self Organizing Map (SOM), generating a similarity space that clusters similar boundary features together. SOM appear to embody some of the properties of the human visual system and organizations in the cortical areas (Philips and Chakravarthy, 2017). SOM has a straightforward implementation that proved to be an effective way for describing an environment in terms of its distinct places.

A Self Organizing Map (SOM) is a type of artificial neural network that is trained in an unsupervised fashion. The objective of the training is to teach the SOM the similarities between the input data and encode them in a lower dimensional space. The particular way an SOM learns a lower dimensional representation causes similar data to cluster in similar locations in the encoding space. SOM uses competitive learning instead of error correction learning. Competitive learning works by increasing the specialization of each node in the network based on the topology of the encoding space. Therefore, this type of learning requires that the nodes be arranged in a

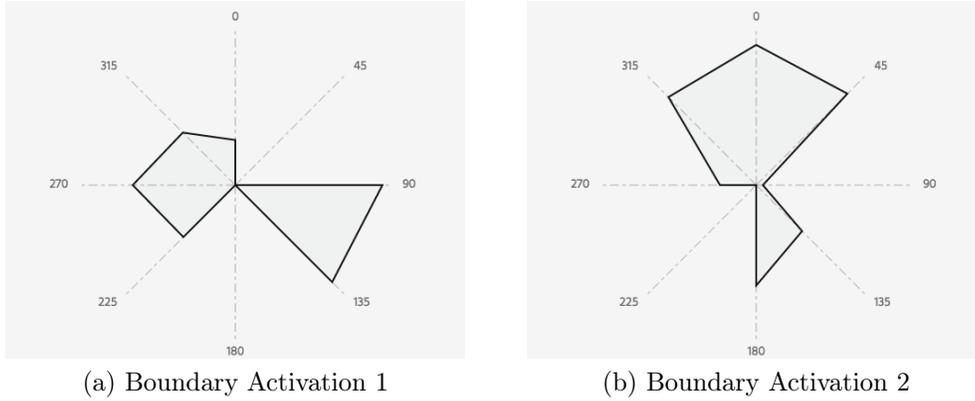


Figure 5-10: Spider Graph Visualization. This figure illustrates the domain-specific representation of environmental boundaries and their relative directions. Each axis of the spider represents a normalized distance to a boundary in a specific direction. A value of 1 corresponds to a boundary closer than one meter, and a value of 0 means there is no boundary within five meters. For example in (a) the agent detects no boundaries at 180 degrees (behind), while there is an immediate boundary at 90 degrees (on the right).

predetermined coordinate system where one can calculate the distance between two nodes. During the learning phase, the update on a node causes neighboring cells to have the same update multiplied by a distance coefficient. The farther away the neighbor cell, the less it is affected by the update operation.

In order to better illustrate the representations learned by SOM, I adopted a visual representation of a boundary similar to the illustration in Figure 5-3(b). In this representation, eight values obtained from boundary activations are mapped on their corresponding directions, creating a spider graph. Because each activation value signifies a nearby boundary or the lack of one, each is represented as a line segment towards its selected direction with a length from 0 (no boundary) to 1 (very close boundary). The spider graph therefore represents both the existence and distance of all boundaries around the robot. Figure 5-10 illustrates the spider graph visualization.

Model Training and Evaluation A 20 x 20 cell SOM is initialized with random values. Figure 5-11 illustrates the initial state of the SOM. A learning rate of $\alpha = 0.5$ and the neighborhood size of 2 are selected as the model's parameters. The learning rate determines how much of an existing value of a cell will be updated with

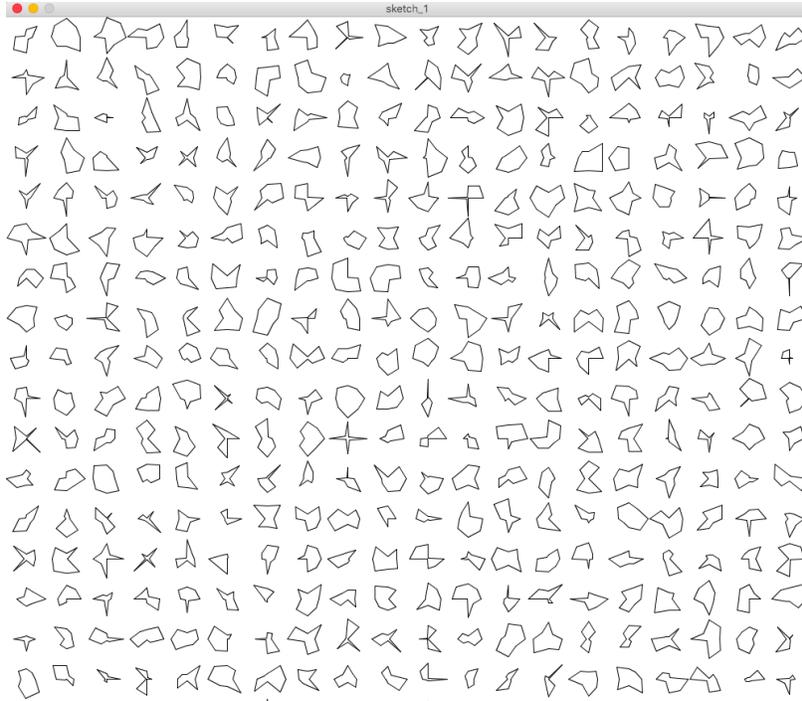


Figure 5-11: Randomly initialized SOM with 20x20 cells

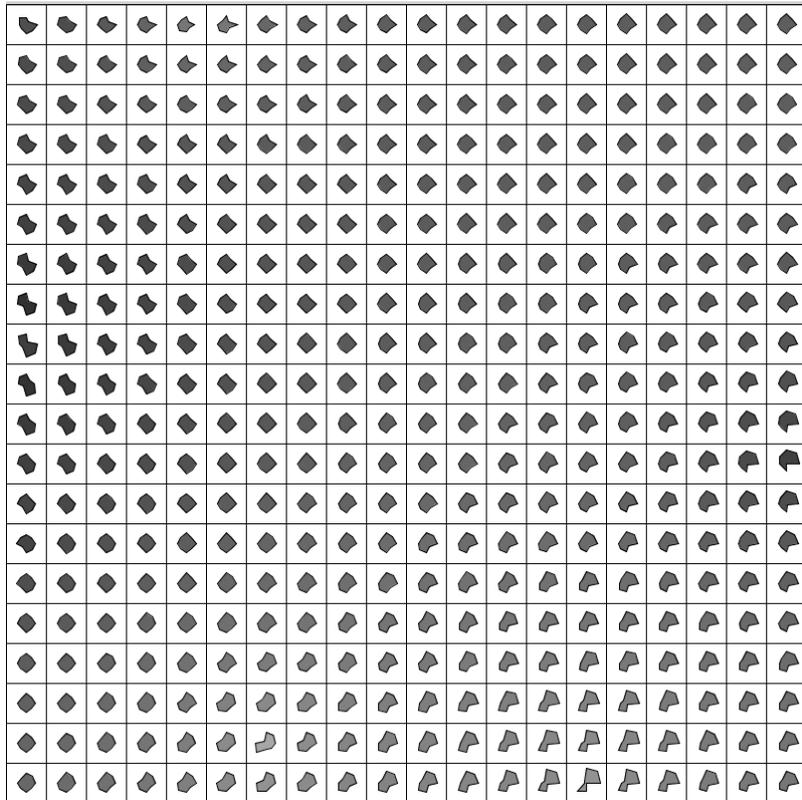


Figure 5-12: Final SOM trained with domain-specific representation of distances and orientations

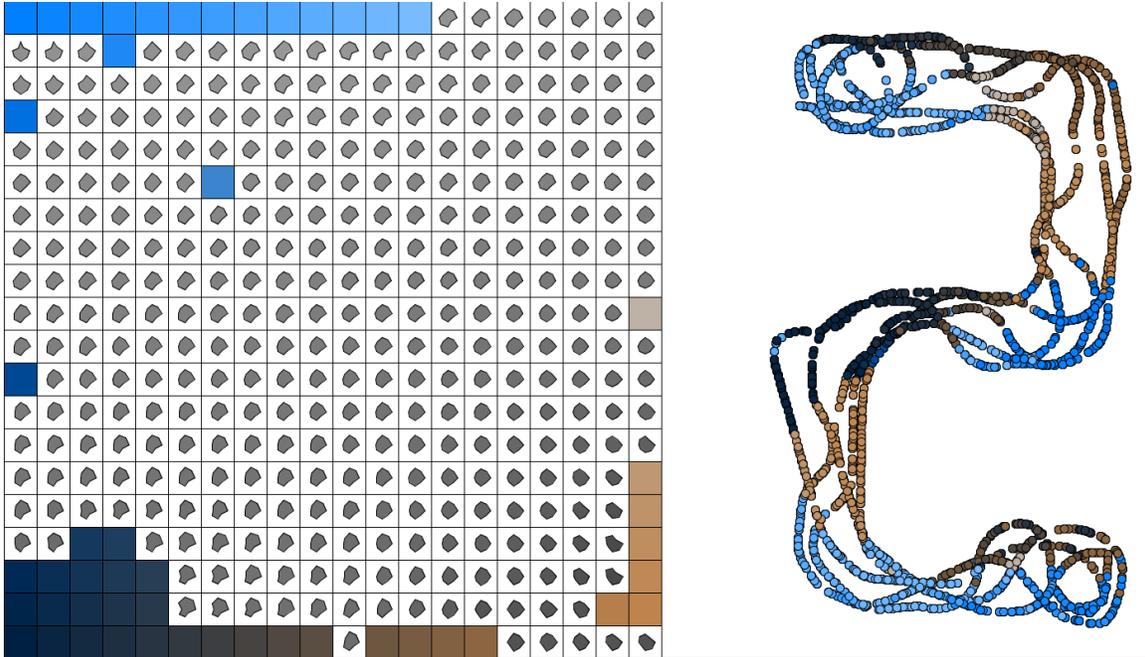


Figure 5-13: The SOM distinguishes distinct places in an environment. The colors represent the index of a cell, where blue correspond to the top left $[0,0]$ index, and brown represent the lower right $[20,20]$ index. This figure shows how each location in the environment on the right is identified by a cell in the SOM. Notice how different types of locations are distinguished by the SOM

the incoming data. The neighborhood radius determines the range of an update, from which no cell is updated. A Gaussian function is used to determine the decay of an update as it moves away from the best matching cell for an update. The SOM is trained with the data collected from the agent's navigation in all training environments until there is no significant update to the cells with the incoming data – that is to say, until the SOM is stabilized. Figure 5-12 illustrates the final, trained SOM. In this figure, each cell represents a value of activation towards eight different environmental direction (see Figure 5-10 for details of this representation).

Results We tested the trained SOM on different environments to see if it can successfully distinguish different locations. The results suggest that the SOM in fact identifies distinct locations in each environment. It appears that the SOM learns generic geometrical features such as straight paths, turns, or corners well enough to assign them different indexes in the map (Figure 5-13). This result is significant

because it shows that only a small subset of environmental information —distances to environmental boundaries— is enough to identify distinct locations in an environment. While this experiment takes place in relatively simplistic maze-like environments, the same approach can be used in real-world scenarios, given that the perception system can produce additional information such as materials and objects.

5.1.8 Summary

Although several biologically-inspired learning approaches exist for modeling spatial navigation and representation (Milford et al. 2009), our model differs in terms of learning strategy and architecture. Initially, our model employs a unique learning approach by training itself online with dynamic labels representing cell activation values that are sampled from a particular random distribution. Moreover, we demonstrated that a CNN with two convolutional layers provides a decent mechanism for predicting cell activation values from images.

In conclusion, through this project we have provided a model to understand mammalian spatial representation and inference mechanisms contributing to the field of computational neuroscience and cognitive science studies. Furthermore, we implemented a multi-modal visuospatial learning framework that could be useful for scene recognition and robotic navigation.

5.2 Connecting Visual and Symbolic Systems

In this section, I outline my approach for using an anchoring framework for visio-spatial reasoning and problem solving. I will present two projects that use the Genesis Story Understanding System —Genesis, in short— for symbolic reasoning and problem solving in coordination with a perception system that generates anchor representations. In the first project, following the instructions provided by Genesis Problem Solver, a simulated robotic arm replaces a broken cellphone battery. In the second project, Genesis generates a description of an environment based on a series

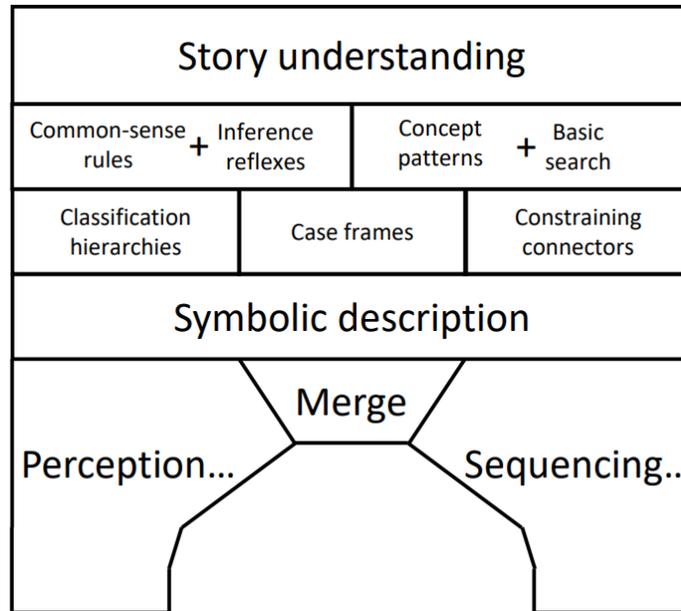


Figure 5-14: Building blocks of Genesis (Winston and Holmes, 2019). My current work focuses on connecting the perception system with symbolic descriptions.

of observations in simulated and real environments. These descriptions are generated in a manner to answer questions, such as "What is on the table?" or "Where can you sit in this room?". Anchors create these descriptions, which are then connected to a rule-based reasoning system and a common-sense knowledge base.

First, I will provide a brief description of the Genesis Story Understanding system, which provides the symbolic reasoning and problem-solving capacities for the anchoring system.

5.2.1 Genesis Story Understanding System

Genesis computationally models various aspects of human story understanding. The project builds on the core idea that story understanding is a key component of human intelligence, which separates humans from all other animals. According to Winston, we humans have a unique capacity to form inner stories, which in turn allow us to form complex and nested symbolic descriptions. An inner story is a "highly nested symbolic description of properties, relations, actions, and events." (Winston and Holmes, 2019).

Genesis translates stories written in English to inner stories, with which it performs various computations such as answering questions, solving problems, or comparing different stories (Winston 2011, 2012a,b). While Genesis primarily uses the external stories provided in natural language to generate inner stories, Winston notes that the outer stories can also be provided purely in visual form via images, pictures or diagrams. My current work on anchoring focuses on extending the capabilities of Genesis to use visual and spatial information to form inner stories. A detailed description about the foundations of Genesis can be found in (Winston and Holmes, 2019). Figure 5-14 shows the basic building blocks that make up Genesis (Winston and Holmes, 2019).

Genesis Uses Case Frames to Represent Inner Stories

Genesis uses Boris Katz's START system to translate English into a collection of entity-relation-entity triplets (Katz, 1968), and then further processes them into descriptions of story elements (Winston and Holmes, 2019).

Winston and Holmes introduce four different components used to express story elements:

An *entity* is an element that has a name and a unique index that distinguishes it from other elements with the same name. For example, the word "table" is translated into an entity in an inner story.

A *function* is an entity with an additional subject slot filled by another entity or entity subclass. For example, in the phrase "on the table", **on** is a function with a subject **table**.

A *relation* is a function with an additional object slot filled by another entity or entity subclass. A relation describes how two entities are related to each other. For example, in the sentence "A ball rolled on the table," the word **rolled** describes how the entity **ball** is related to the entity **table**, which is the subject of the function **on**.

A *sequence* is a set of story elements, which combines a series of entities, functions and relations.

With these components, we can represent story elements, which are also called

case frames. For example, the sentence "A ball rolled on the table" can be expressed in the inner story as:

```
(relation roll
      (entity ball)
      (sequence roles (function on (entity table))))
)
```

The story elements are connected to each other with various explicit connection constraints, such as *cause*, *means-ends*, and *enablement*. For example, the sentence "The ball rolled on the table because the kid dropped the ball" can be expressed in the inner story as:

```
(relation cause
      (sequence conjunction (relation drop
                             (entity kid)
                             (function object (entity ball))))
      (relation roll
        (entity ball)
        (sequence roles (function on (entity table))))
)
```

Genesis Uses Common Sense Rules to Connect Story Elements

As in the example above, a story is a sequence of story elements. Genesis connects story elements any time an explicit connection between story elements is expressed, such as by using the words *because*, *leads to*, or *in order to*. Genesis also uses commonsense rules to form connections between story elements. Winston calls these connections *inference reflexes*. Inference reflexes represent a relation or a fact that we infer from story element even if they are not explicitly stated. Inference reflexes automatically add story elements to inner stories. Winston introduces a series of commonsense rules that trigger inference reflexes, such as a deduction rule "if x murders y, y becomes dead," or an explanation rule "if y angers x, x may kill y." Using entities, case frames, and common-sense rules Genesis builds an elaboration

graph (Figure 5-15).

Inner stories enable Genesis to make inferences, answer questions, and solve problems. However, in its current state, Genesis can only use information provided in natural language or in the form of commonsense rules. Winston points out that much of our knowledge comes from our experiences in the world and is presented to us through our perception. Using the anchoring mechanism, I propose a novel method to connect perceptual information to inner stories. In the following two examples, I will show how anchoring enables Genesis to (1) solve a visuo-spatial problem —guide a simulated robot to replace a cellphone battery, and (2) verbally describe a visual scene, and answer questions regarding that scene.

5.2.2 Simulated robot replaces a cell phone battery

What will you see in this example?

- A simulated robotic arm replaces a cell phone battery under the guidance of Genesis Problem Solver.
- Anchors create vertical and horizontal relationships from visual information.
- Genesis communicates with a vision system, using anchors.
- Genesis discriminates visually identical objects, using anchors.
- Genesis understands where to place objects and if they can be picked, using anchors.

Project Setup

In this example, Genesis communicates with a simulator to guide a robotic arm to replace a cellphone battery. The knowledge required for replacing a cellphone battery is provided in simple English. Genesis parses those instructions and converts them into individual steps which can be communicated to the robotic arm. The simulator performs object identification, and builds a representation of the scene using a series

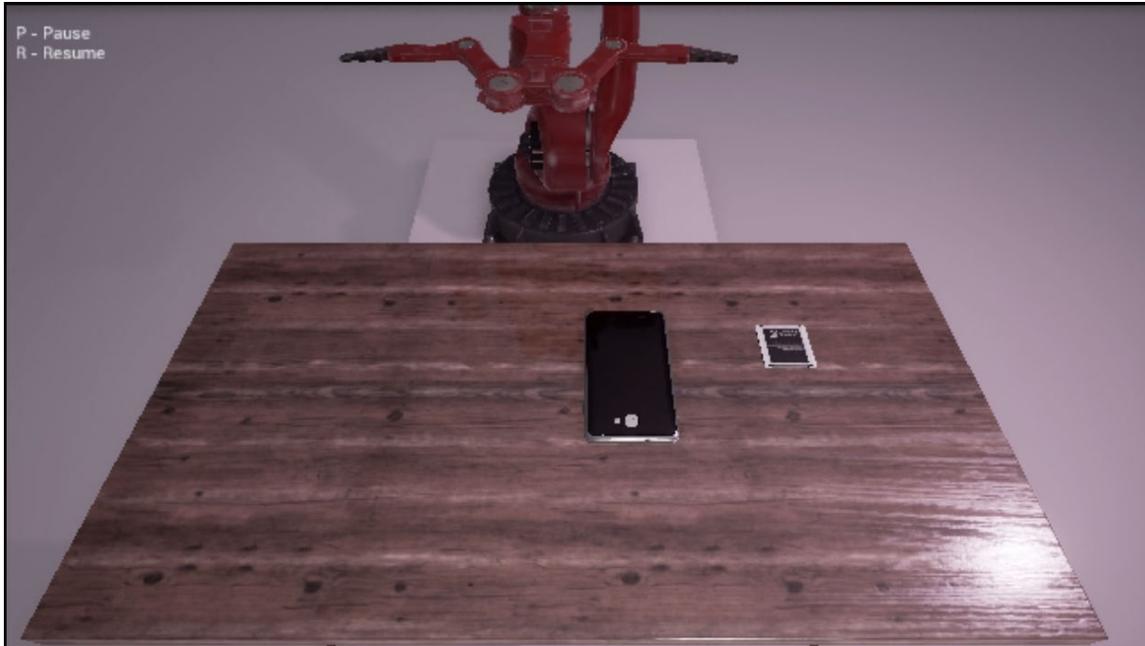


Figure 5-16: Robotic Arm Simulator. The simulator environment consists of a robotic arm placed behind a table, a cellphone and a replacement battery on the table. The simulated perception system sees the environment from the same point of view as in this image.

of anchors. The simulator uses the anchors to answer the questions asked by Genesis, such as "Is there available space on the table?". The communication protocol enabled by the anchors allows Genesis to successfully follow the instructions and guide the robotic arm to perform the task.

The simulator environment is created in the Unreal Game Engine. Figure 5-16 illustrates the simulator environment. The environment consists of a robotic arm located behind a table on which a cellphone and a replacement battery are placed. The simulated perception system sees the environment as depicted in Figure 5-16. There are two main responsibilities of the simulated perception system: identify objects and their locations and inform the robotic arm to take an action.

The robotic arm has two hard-coded actions: (1) pick up an object and 2) place an object on a target location. Both of these actions are guaranteed to succeed; that is, the robotic arm never fails to pick or place an object. The object and target locations are provided as three dimensional coordinates in the simulated environment.

The simulated perception system carries out the three tasks as they are requested

by Genesis. The first task is to identify the objects provided in the instructions, which consists of a cellphone body, a cellphone cover, a broken battery, a replacement battery, and a table. In other scenarios, the perception system might also notify Genesis about other objects it sees (as, for example, in the next example). In this example, however, the problem solver only needs to interact with this limited set of objects.

Similar to the actions of the robotic arm, the simulated perception system is guaranteed to identify this limited set of objects; it is not guaranteed to discriminate between the broken battery and the replacement battery. The batteries are visually identical, and both of them are recognized as *battery* by the system. They are discriminated solely based on their locations on the table and with respect to the other objects. Anchors allow the simulated perception system to represent those relationships as needed. For example, in the beginning of the task, Genesis informs the simulated perception system that the battery on the table should be known as the *replacement battery*. The attachment anchor (ON) generated by the perception system such as "cellphone on the table" and "battery on the table" enables the simulator to identify the battery on the table as the replacement battery. Anytime Genesis sends a request involving the replacement battery, it is matched to this anchor so that the simulated perception system can correctly identify the replacement battery..

The second task is determining if an object can be picked up, and the third task is determining if there is available space on a target object. Both of these determinations are made using the anchor representations (i.e. if a location is part of an ON anchor, that location is not available for placing an object on it).

Parsing the Instructions

The Genesis Problem Solver follows a special instruction written in English for replacing the cellphone battery. The instruction consists of a beginning condition, steps, and "The end." marker, at which point the task is performed:

If the intention is "Replace cellphone battery".

Step: Identify cellphone parts.
Step: Disassemble cellphone.
Step: Install replacement battery.
Step: Reassemble cellphone.
Step: Inspect cellphone.
Step: Test cellphone.
The end.

Genesis parses each set of instructions and converts them into individual steps. For example, the first instruction is parsed into a series of identification procedures. In order to parse the instructions into more detailed steps, Genesis uses a series of commonsense rules. Note that the Genesis Problem Solver has multiple levels of reasoning, starting by first identifying the problem, then identifying the intention and, finally, proposing an approach. The details of the problem solver can be found in (Yang and Winston, 2017).

Each step is further broken down to individual steps using the common-sense rules. For example the instruction to disassemble cellphone is broken down as follows:

If the intention is "Disassemble cellphone".
Step: Remove cellphone cover.
Step: Remove suspect battery.
The end.

At this stage, the Genesis Problem Solver has two subtasks to perform: remove the cellphone cover and remove the battery. These steps are also broken down as follows:

If the step is "Remove cellphone cover".
Solve: Put the cellphone cover on the table.
The end.

If the step is "Remove suspect battery".

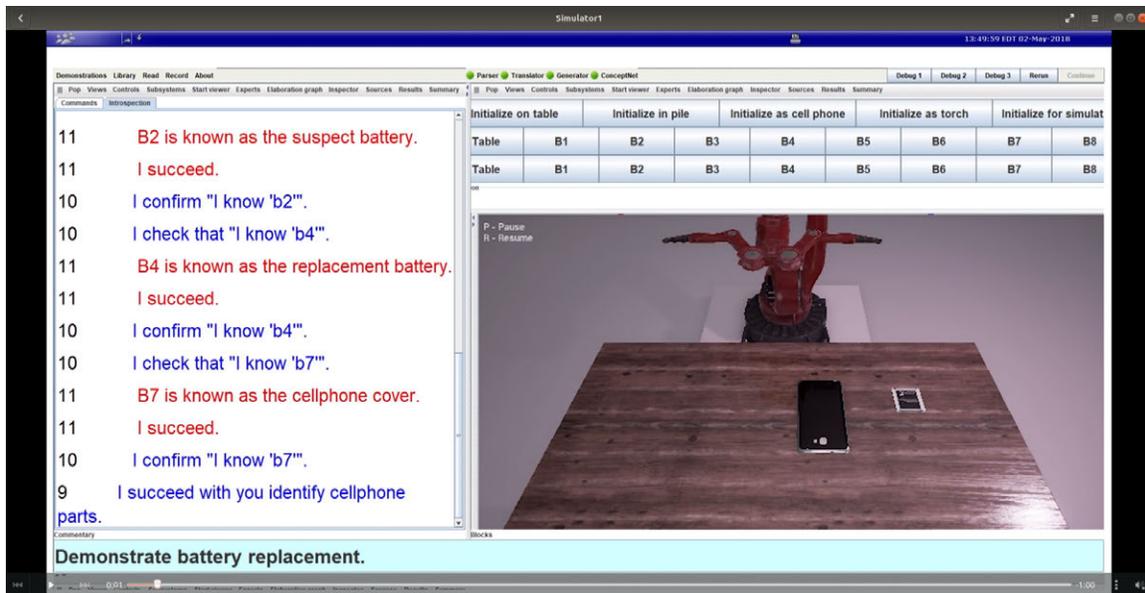


Figure 5-17: An overview of the Genesis Problem Solver and the simulator. This figure illustrates a screen from the complete system. On the left, Genesis logs each step it takes. The blue text corresponds to a step that Genesis internally parses and follows, and the red text represents steps handled by Just Do It methods, which communicate with the simulator.

Solve: Put the suspect battery on the table.

The end.

Whenever Genesis reaches a "Solve" marker, it attempts to take the action provided by the instruction. This is when Genesis needs to consult with the perception system. The problem solver needs to determine (1) the locations of objects in the instructions (2) whether it can pick the object, and (3) whether there is available space on the table. All of these are questions to be answered by the perception system. The Genesis Problem Solver has a generic function to carry out the tasks that require the perception system, which is called *Just Do It*. Simply stated, Genesis cannot symbolically determine the presence of an object or its spatial relationship with other objects, but it can instruct the perception system to act on its behalf. A Just Do It function transmits a message that conveys a particular instruction for the perception system, such as "put x on y." In return, the perception system notifies Genesis that a task is completed or that it failed.

Figure 5-17 illustrates an overview of the system. On the left of the figure, there is

a series of logs generated by Genesis in real time, corresponding to the steps it takes. A blue colored text represents a step that is symbolically parsed and followed, while a red colored text represents a Just Do It method that is handled by the simulator. The numbered letters (i.e. B2) in the logs represent unique indexes to each object the simulated perception system identifies. In this way, Genesis internally keeps track of the unique identities of the objects in the simulator scene.

Handling *Just Do It* Methods in the Simulator

The most critical component of the battery replacement system is handling the Just Do It methods. Apart from parsing instructions and following predetermined common-sense rules, Genesis cannot carry out the task without these Just Do it methods. There are three main steps in handling a Just Do It method:

1) Transmit message to the simulator:

When Genesis runs a Just Do It method, a message is constructed and transmitted to the simulator. There are three different types of possible messages that can be constructed in a Just Do It method. The first one is an action message, which tells a robot to pick up an object, such as "pick up cellphone body" or place an object at a target location, such as "place the cellphone cover on the table". An action message consists of a an "ACTION" marker, and corresponding parameters. The first parameter is the *source object* and the second parameter is the *destination object*. Some examples are as follows:

```
\\Syntax -- Message(MARKER, COMMAND, SOURCE, DESTINATION)
```

```
Message ("ACTION", "pick", "cellphone cover",null)
```

```
\\this message is parsed by simulator as "Action: Pick up cellphone cover".
```

```
\\null parameter indicates that there is no target object for this action.
```

```
Message ("ACTION","place", "cellphone cover", "table")
```

```
\\this message is parsed by the simulator as "Action: Place the cellphone  
cover on the table".
```

The second type of message is an identification message, such as "Identify cellphone cover" or "The battery on the table is a replacement battery". An identification message can either instruct the simulator to construct an anchor representation or disambiguate an object with an existing anchor. The following are examples of identification messages:

```
\\Syntax -- Message(MARKER, COMMAND, SOURCE, DESTINATION)
```

```
Message ("IDENTIFY", "Generate", "cellphone cover", null)
```

```
\\This message is parsed by the simulator as "Identify: Generate cellphone  
cover".
```

```
Message ("IDENTIFY", "Disambiguate", "battery on table", "replacement  
battery")
```

```
\\This message is parsed by the simulator as "Identify: Disambiguate the  
battery on table as replacement battery".
```

The third type of message is a spatial relation message, which ensures either that there is available space in a target location, or a target and source objects are in a certain spatial relationship. For example, the instruction "Verify that nothing is on cellphone cover" or "Verify that replacement battery is in the cellphone body" are both converted into spatial relation messages as follows:

```
\\Syntax -- Message(MARKER, COMMAND, SOURCE, DESTINATION)
```

```
Message("VERIFY", "ON", "cellphone cover", null)
```

```
\\This message is parsed by the simulator as "Verify that there is nothing  
on the cellphone cover"
```

```
Message ("VERIFY", "IN", "cellphone body", "replacement battery")
```

```
\\This message is parsed by the simulator as "Verify that the
```

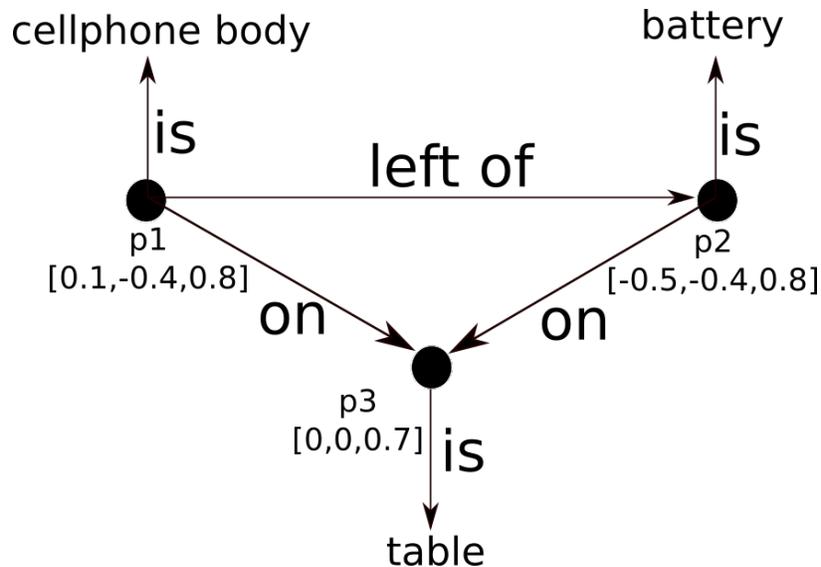


Figure 5-18: An anchor graph consists of nodes representing three dimensional locations in space that are connected to each other with particular relationships identified by anchors. An anchor graph is a directed graph, where a relationship is directed from the source node to the target node. In this example, a recognition anchor generates "is" relationships, which assigns an identifier symbol to the node. A vertical anchor generates "on" relationships and a horizontal anchor generates "left of" relationships.

replacement battery is in the cellphone body"

I use JSON (Java Script Object Notation) format to construct the messages and transmit them over the local network to the simulator. An example message in JSON format is as follows:

```
{"Marker": "VERIFY", "Command":"On", "Source:"cellphone cover", "Destination":null}
```

Once this message is transmitted to the simulator, it is parsed and a corresponding action is taken by the simulator, based on the marker.

Parse messages and generate anchors:

Once a message is transmitted to the simulator it is parsed into one of the possible actions identified by the marker. The first type of action is an IDENTIFY action, which is also the first step in the phone replacement scenario. The simulated perception system can generate three different types of anchors for this process:

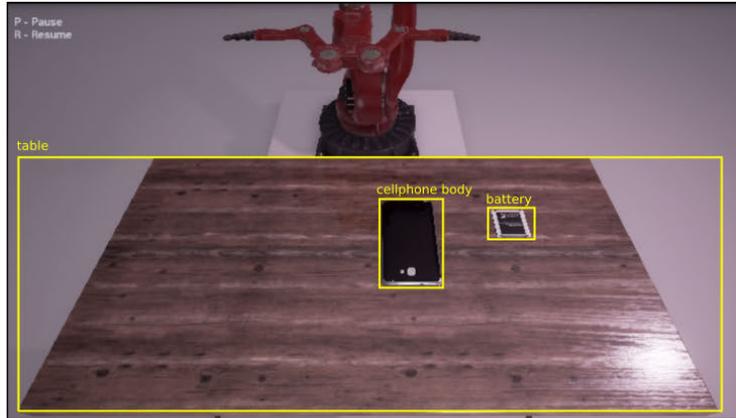
recognition, vertical relation, and horizontal relation. All of the anchors are stored in an *anchor graph*. An anchor graph is an internal representation of the simulator, which stores all identified relationships in the scene. The anchor graph consists of location nodes that are connected to each other with particular relationship identified by an anchor. Figure 5-18 provides a visual representation of an anchor graph. This graph corresponds to the result of the initial anchor generation process. The simulator scene at this stage is illustrated in Figure 5-16.

The anchor graph, *A*, is internally stored as a list of triplets representing anchors. The graph in Figure 5-18 is represented as follows:

```
A = ([p1, IS, "cellphone body"], [p2,IS,"battery"], [p3,IS,"table"], [p1,
LEFT_OF, p2], [p1,ON,p3], [p2,ON,p3])
```

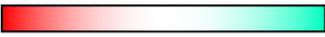
A *recognition anchor* is generated by an object recognition procedure (hard-coded for this example) that recognizes an object and gathers its three-dimensional location. A *vertical relationship anchor* identifies an attachment relationship ON between two objects. Because the current context does not allow checking for containment relationships, IN is also considered a vertical relationship when it occurs between the phone body and a battery. Therefore any time an ON anchor is generated between the cellphone body and a battery (i.e. [p1, ON,p2] in the previous graph), it is automatically replaced with an IN anchor. Because I had access to the physics engine of the simulator, I validated an ON anchor by checking whether two objects collided. *Horizontal relationships anchors* LEFT and RIGHT are generated based on their horizontal locations in the scene. Figure 5-19 illustrates visual processes for generating all three types of anchors. Illustrations of domain-specific representations are also provided for vertical and horizontal relationships.

Using these three anchors, the simulator can identify objects in the scene and generate an anchor graph. Any time Genesis sends an identification message, the simulator can check if corresponding anchors are present in the graph or generate them if necessary. In algorithm 1, a pseudocode is provided for generating anchors when Genesis transmits an identification message. There are two utility functions in this

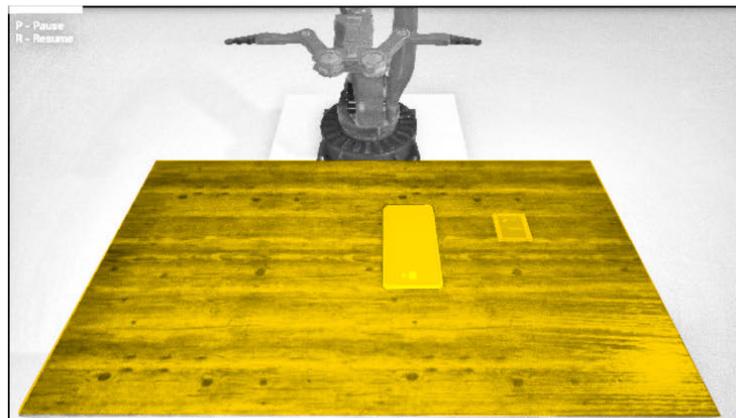


(a) Object Recognition



y=-1.5m  y=1.5m

(b) Horizontal Relations



(c) Vertical Relations

Figure 5-19: Visual processes that generate anchors.(a) Object Recognition (b) Horizontal Relationships (c) Vertical Relationships.

algorithm: *GenerateAnchor* and *AssignAnchor*. *GenerateAnchor* function iterates over all three domains(recognition, vertical and horizontal relationships) and attempts to generate corresponding anchors for the provided source object. An *AssignAnchor* function first generates a candidate anchor from the provided source object and then matches it to the anchor graph. If there is a successful match, it assigns the target symbol to the match.

Result: An anchor identifying an object added to the Anchor Graph

R -> List of objects recognized by the system.

A -> Anchor graph. Each node of the graph is a three dimensional location.

Function *Identify(Command, Source, Target):*

```

if Command == "Generate" then
  foreach r in R do
    if r == Source then
      Generate Anchor(Source)
      return True
    end
  end
  return False //Object not found
end
if Command == "Disambiguate" then
  if Source not in A then
    Generate Anchor(Source)
  end
  Assign Anchor (Source, Target)
  return True
end

```

Algorithm 1: Generate Anchors for Identification.

Most of the required information for the task is generated in the identification phase, because in addition to the list of objects, the simulator exhaustively searches for all possible anchors in the scene and then builds an anchor graph. The second type of action that Genesis might request is a VERIFY action. A VERIFY action establishes whether or not a target location is available. Except for the table, which can have multiple vertical relationships, all vertical relationships are stored in the anchor graph after the identification. Therefore, the simulator directly verifies an available space by searching the anchor graph. If a location (represented by an object) is in the anchor graph as the source object, it means that the location is not available. If the

source object is the table, then four possible predetermined locations are checked. These locations are hard-coded in the system in order to reduce the complexity. In this case, the algorithm counts the number of ON relationships the table is in as the primary object and ensures that number is smaller than four.

The third type of message is an ACTION. When the simulator receives a message with an ACTION marker, it instructs the robotic arm to take the corresponding action provided in the parameters. The robot can either pick an object or place an object. A pick action is always necessarily followed by a place action, therefore these two actions can be grouped as a move action.

Update the Anchor Graph:

Any time a move action takes place, the simulator updates the anchor graph because some of the anchors are no longer valid. The process for updating an anchor graph consists of (1) deleting horizontal and vertical relationship anchors for the original anchor and storing identity anchors in a temporary list (in case it is a battery), (2) running an identification procedure, and (3) running an assignment procedure for the stored identity anchors.

Respond to Genesis:

Each request from Genesis expects a success or failure message. After all IDENTIFY, VERIFY, and ACTION messages, the simulator sends either a success or failure message.

Figure 5-20 illustrates various stages of the battery replacement scenario. Notice how different messages from Genesis (represented in red) guide the robot to carry out the task as requested.

Summary

In this battery replacement scenario, I illustrated how an anchoring framework enables Genesis to carry out visual and spatial tasks in collaboration with a perception

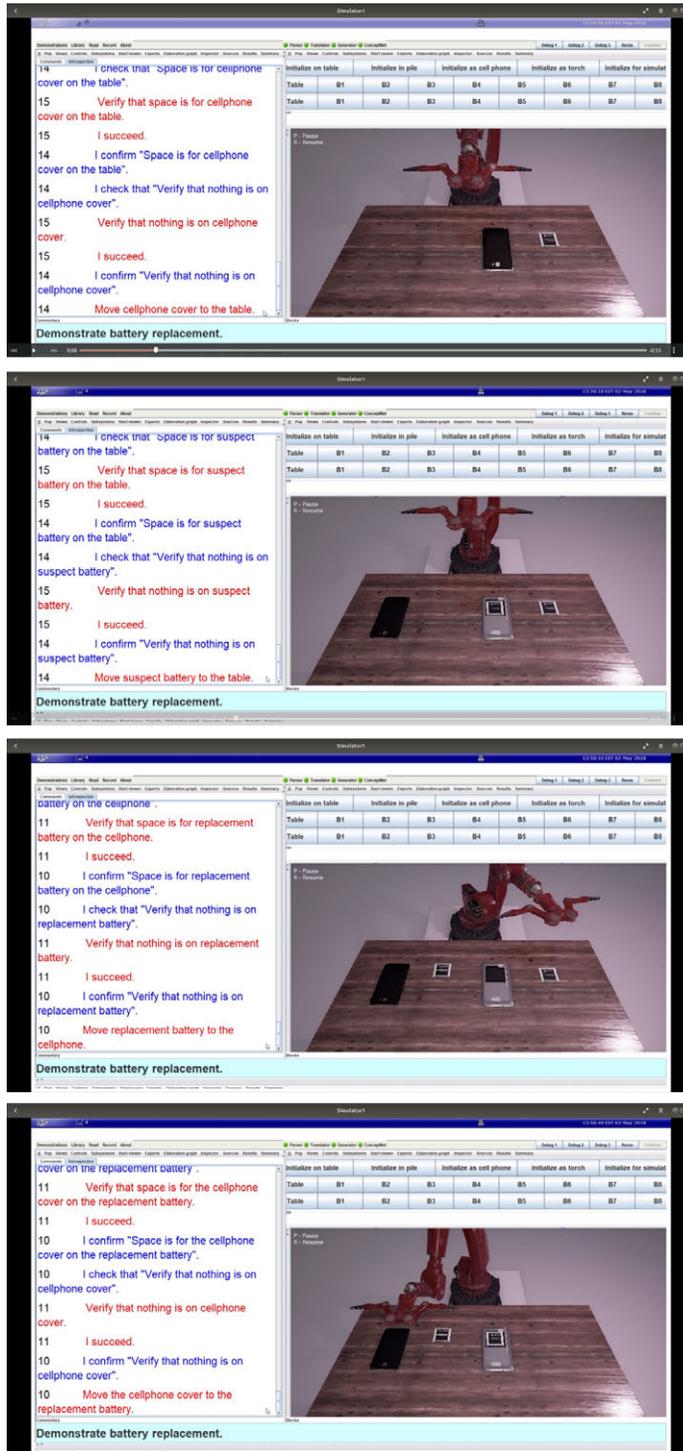


Figure 5-20: Different stages of battery replacement task

system. This framework is flexible enough to carry out a completely different set of instructions as long as the provided domains are sufficient. In fact, Yang and Winston (2017) provide a series of problems that can be solved by the Genesis Problem Solver in a similar fashion. In the next example, I will briefly introduce another important advantage of using the anchoring framework: the ability to chain together observations and answer questions.

5.2.3 Genesis Describes the Environment

What will you see in this example?

- Genesis builds an elaboration graph with anchors generated by a simulator system that explores a virtual environment.
- The simulator generates anchors and transmits them to Genesis in real time.
- Genesis chains together anchors and identifies additional relationships using commonsense rules.
- Genesis generates a description of the environment.
- Genesis answers questions about the environment.

In the previous section, I provided an analysis of observations in virtual reality environments using the anchoring framework. I showed that the majority of descriptions can be modeled using a limited number of anchors that identify visual, spatial, and temporal relationships. In this example, I will demonstrate that we can make Genesis construct similar descriptions using anchors. This example is similar to the previous example in the sense that Genesis communicates with a simulator. However, this time, instead of Genesis making queries about the environment, the simulator constantly sends anchors to Genesis, where they are integrated into an elaboration graph (see section 5.2.1). By constructing an elaboration graph Genesis can discover relationships, augment descriptions with commonsense rules, and answer questions.



(a) Test Environment 1: Residential

(b) Test Environment 2: Office

Figure 5-21: The test environments for the scene description task

Project Setup

The simulator consists of a residential environment, similar to those explored by the subjects in the previous chapter. The simulated perception system identifies objects from the camera, which is controlled by a human experimenter to look into different parts of the environment. Observations are sent in real time to Genesis to be integrated into an elaboration graph.

Generating Anchors in the Simulator

The first task in describing the environment is to generate various anchors using the simulated perception system. I used the same simulator employed in the previous example, which can identify objects, horizontal and vertical relationships. In addition to these, I also introduced a NEAR anchor that identifies neighborhood relationships between objects regardless of their particular horizontal or vertical relationships. This anchor is used by Genesis to apply commonsense reasoning to identify a location context as well as to discover additional relationships between objects even when those were not identified by the perception system.

In addition to anchor representations extracted from the images, the simulator

also reports on its actions. It can identify any time the observer turns towards a direction such as LEFT, RIGHT, UP or DOWN. When an action takes place, the simulated perception system constructs a message with an ACTION marker and the type of action.

Figure 5-21 shows two test environments for this task. The first environment is a residential environment that was not included in the observation study, and the second one is an office environment that was used in the observation study.

Transmitting Anchors to Genesis

Genesis listens for an observation event that may occur in the simulator. Once an anchor such as ON(computer, table) is generated, the simulator constructs a message to be transmitted to Genesis. Similar to the messages constructed by Genesis in the previous example, the simulator constructs two types of message markers: a PERCEPTION message and an ACTION message. The following are examples of messages that are transmitted to Genesis:

```
\\Syntax -- Message(MARKER, COMMAND, SOURCE, DESTINATION)
```

```
Message ("PERCEPTION", "on", "computer", "table")
```

```
Message ("ACTION", "turn", "self", "left")
```

These messages are then converted into JSON format and sent over the network to Genesis. Genesis's story processor parses the incoming messages into story elements based on the marker it receives.

Building the Elaboration Graph

Genesis first translates the message into an English statement, such as "computer is on table" and sends it to the START parser. The START parser converts the statement into a story element which could be integrated into the elaboration graph. When Genesis receives an ACTION message, it updates its current orientation with

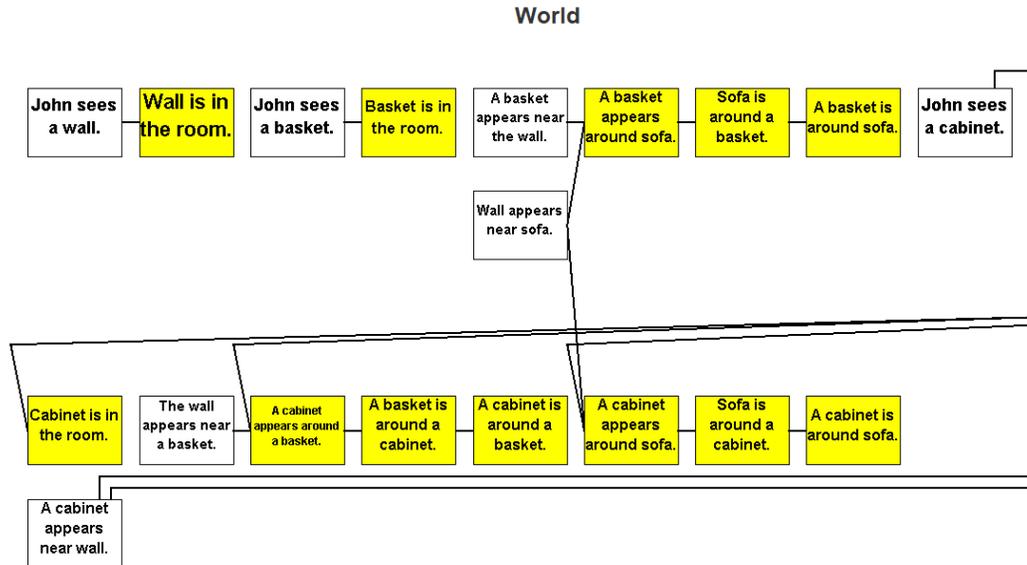


Figure 5-22: Elaboration Graph generated from observations in the residential environment

the data received in the message. By so doing, Genesis can infer relative horizontal relationships between objects, which are then used in summarizing the observation.

One additional step before adding a story element to the elaboration graph is to check if there is an existing observation in the story graph. Genesis matches an object, such as a computer, to an existing object with the same name if two observations are made consecutively. For example, two consecutive statements such as "a computer on table" and "a keyboard in front of computer" enable Genesis to identify the computer in these statements as the same.

Genesis also uses a series of commonsense rules to infer additional relationships between the observations. For example, two statements such as "table is near by the couch" and "lamp is near the table" lead to an additional observation of "lamp is around the couch." Notice that the perception system might not construct this relationship but commonsense rules allow Genesis to add it to the story. Below I provide some example commonsense rules that are used in inferring additional relationships from the observations:

`xx,yy,zz are things.`

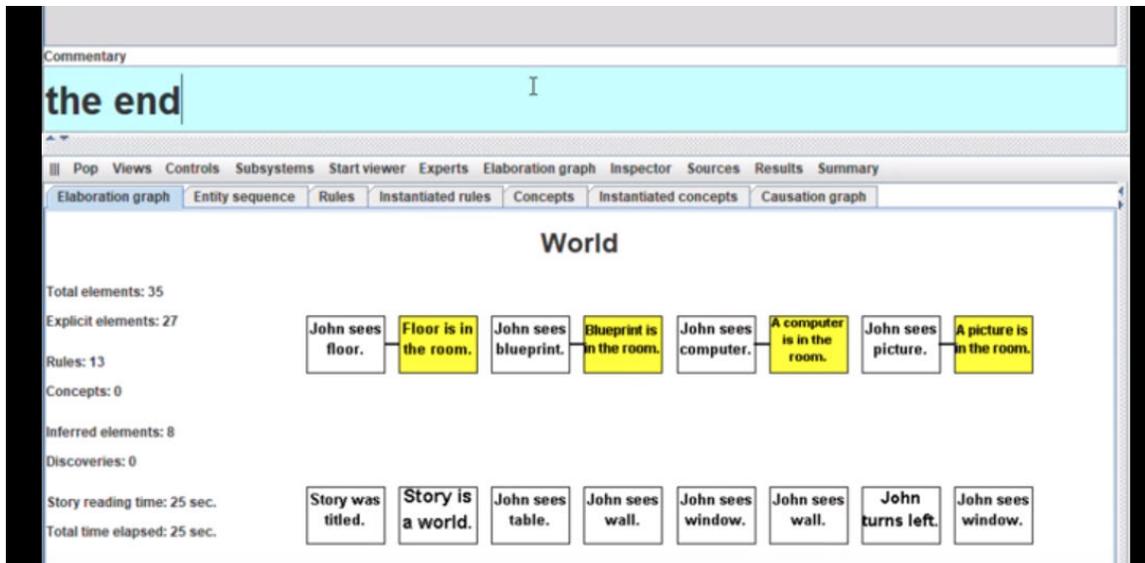


Figure 5-23: Elaboration Graph generated from observations in the office environment

Whenever xx is in the room and yy is in the room then xx may be near yy.

Whenever xx is near yy and yy is near zz then xx is around zz.

Whenever xx is near yy then xx cannot is around yy.

if xx is around zz then zz is around xx.

Genesis builds an elaboration graph after matching existing elements and identifying additional relationships using the common sense rules. Figure 5-22 and Figure 5-23 illustrate elaboration graphs generated at the end of observations in two test environments. Yellow elements are those added by Genesis to the story using the commonsense rules. Note that the "self" marker in a message is converted to "John" due to the limitations in the START parser.

Below is the complete list of statements used in generating this elaboration graph. The observations in the left column are obtained from the residential environment, while the ones in the right column are obtained from the office environment.

Start story titled "The House".

Start story titled "The Office".

John looks forward.

John looks forward.

John sees a window.

John sees a window.

John sees a wall.

John turns up.

Window is near wall.

John sees a ceiling.

John turns right.	John sees a skylight.
John sees a television.	John sees a lamp.
John sees a cabinet.	John looks forward.
Television is on cabinet.	John turns left.
John turns left.	John sees a cabinet.
John sees a sofa.	John sees a window.
John sees a pillow.	John sees a wall.
Pillow is on sofa.	John turns right.
John looks down.	Johns sees a blueprint.
John sees a floor.	John sees a table.
John sees a carpet.	John sees a computer.
The end.	Computer is on table..
	John sees a chair.
	John sees a picture.
	Picture is on wall.
	The end.

Summarizing Observations

The elaboration graph enables Genesis to reason about the observations and answer specific questions. However, the elaboration graph itself is not a description of the environment. A description is generated as a summary of all observations, which includes where the objects appear relative to the observer, and explicit and implicit relationships between objects (not repeated for every pair of object). In the summarization process, all "John sees" statements are replaced with "There is" for clarity.

The observations in the residential and the office environments are summarized by Genesis as follows:

The House Story . *I am in a house. There is a wall in front of me. There is a window on the wall. There is a sofa near the wall. There is a pillow on the sofa. There is a cabinet on the right. There is a television on the cabinet. There is a floor*

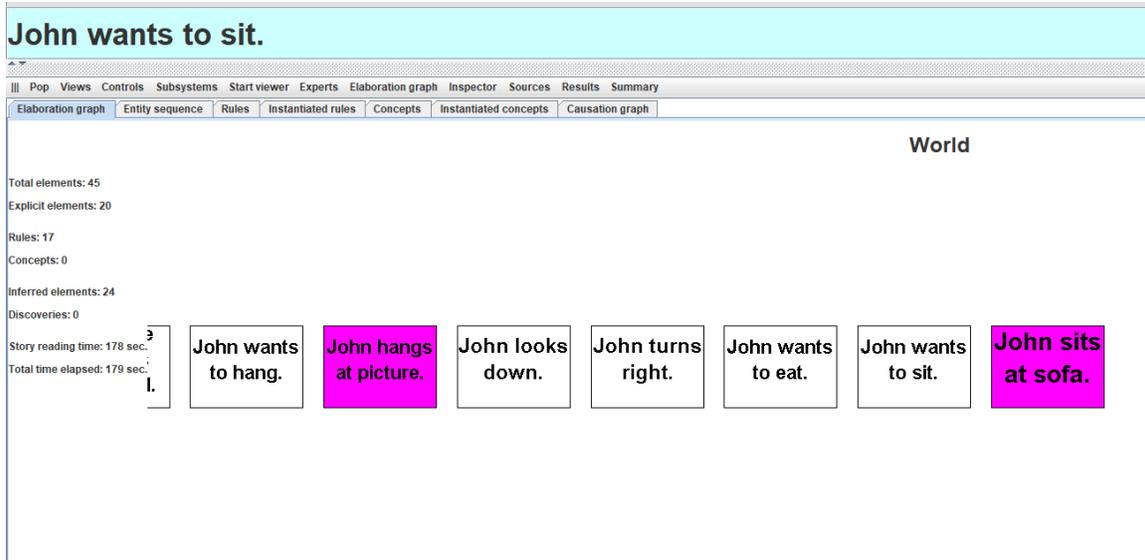


Figure 5-24: Answering questions using ConceptNet

below. There is a carpet below. The end.

The Office Story . *I am in an office. There is a window in front of me. Window is on the wall. There is a cabinet near the wall. There is a ceiling above. There is a skylight on the ceiling. There is a lamp near the skylight. There is a blueprint on the right. There is a table near the blueprint. There is a computer on the table. There is a chair at the table. There is a wall. There is a picture on the wall. The end.*

Answering Questions

Genesis can answer questions about the spatial relationships among objects, their presence in a story, as well as their function. In order to answer questions about the functions of the objects, Genesis consults ConceptNet, an online commonsense database. ConceptNet provides a list of connections between concepts, such as "a type of," "used for," and "located at." When Genesis receives a question regarding the function of the object, it looks for a "used for" relationship in ConceptNet for each object observed in the story. For example, if we ask, "Where can I sit?", (1) Genesis infers that the question is about the sitting action, (2) sends ConceptNet a query to obtain a list of objects used for sitting, (3) and then matches the list of objects sent

by Concept Net to those observed in the story. If there is a match, Genesis provides the answer to the user. Whenever Genesis answers a question using ConceptNet it adds a purple element to the story (Figure 5-24).

Summary

In this example, I illustrated how anchor representations can be transmitted to Genesis and integrated into inner stories. Sending anchor representations to Genesis enables it to construct complex symbolic descriptions —inner stories— from the observations: using these inner stories, Genesis can summarize an observation and answer questions. Genesis also uses commonsense rules to infer additional relationships such as AROUND, which helps it identify how two objects are related to each other even when the perception system fails to capture such a relationship. Answering questions with the help of ConceptNet also implicitly adds new elements to the story. The answer to the question "Where can I sit?", can be added to the elaboration graph as "You can sit on a chair." We can imagine asking many questions about objects, how they appear and what they are used for to augment initial anchor representations.

One immediate next step for this project is to enable Genesis to guide the simulation to make additional observations based on its common sense reasoning. For example, if Genesis cannot answer the question "Where can I sit?", it should tell the simulated perception system to observe the environment further in order to carry out the task.

5.2.4 Contributions

In this chapter, I presented the implementation of the anchoring framework for spatial problem solving and visual descriptions. I identified two sub-problems related to the anchoring framework: (1) learning domain-specific representations, and (2) connecting perception systems with symbolic systems. My contributions in this chapter are in two areas:

Contributions towards learning domain-specific representations:

- Within the context of spatial navigation I presented a biologically inspired model that can learn domain-specific representations that in turn make evident information regarding distances and orientations from images.
- I showed how the model successfully predicts an agent's orientation and distance to the environmental boundaries and help it avoid obstacles.
- I further demonstrated that these representations can help us identify another type of environmental information: distinct locations within the environment. The idea of discriminating locations based on boundaries and orientations is supported by the studies on place cells in the rodent brain.

Contributions towards connecting symbolic systems with perceptual systems:

- I demonstrated that the anchoring framework enables a symbolic story understanding system (Genesis) to carry out a visuo-spatial task. With anchors, I enabled Genesis to discriminate visually identical objects and keep track of spatial relationships among objects.
- I demonstrated that anchors can be used to generate verbal descriptions of environments and answer questions. Building stories with anchors enables commonsense reasoning and the construction of rich visual descriptions.

Chapter 6

Conclusions

In this dissertation I introduced and tested the anchoring hypothesis, which states that spatial experiences are inner stories that chain together perceptions and actions by identifying spatial and temporal relations among them. I further suggested that this mechanism requires connecting perceptual and symbolic systems. In order to test my hypothesis and to show how perceptual and symbolic systems can be connected, I conducted several experiments and case studies. With these studies, I investigated the relationships between stories, perceptions, and actions.

6.1 Contributions

This dissertation contributes to the fields of design, artificial intelligence, and computational cognitive science.

The case studies introduced in chapter 3 explore the idea of creating and representing spatial experiences. While revealing the inherent relationships between inner stories and perception, these case studies also provided insights on the creative use of immersive media. In the first case study, I presented a virtual reality documentary, “September 1955”, which tells a story within a digitally reconstructed historical building. I introduced several story elements including the *objects of interest* and the *spatial soundscape*, in order to capture and guide the viewer’s attention during the experience. The contribution of this project to the overall

thesis is that (1)it enabled understanding the role of attention and actions in spatial experience: each viewer had a slightly different experience because they each attended to different parts of the environment during the experience, and (2) it illustrated connections between spatial experiences and stories. Many viewers reported that they understood the historical story better because they could spatially explore it and thus felt they were part of it.

In the second case study, I presented the results of a class“Computational Ethnography and Spatial Narratives”. Through this case study, I described how designers represent their spatial experiences using a variety of media, including drawings and physical models. I identified common spatial and temporal relations that the students integrated into their representations, including directions, paths and boundaries. Construction of inner stories, as it was revealed in this case study and the previous one, varies significantly from person to person, and is driven by active interaction with an environment.

In chapter 4, I presented the anchoring framework. In order to better understand the relationships between language and perception, and to discover a variety of anchors, I developed a methodology I called “See, Act and Tell”. In this methodology, the subjects explored an immersive environment, verbally describing what they perceive as they explored it. I provided an analysis of language elements in relation to observations. As a result of this study, I introduced various anchors, including spatial anchors, which are related to domain-specific representations that make evident a certain type of information. I also introduced a computational method for producing descriptions using these anchors and domain-specific representations. I showed that I can model a large portion of verbal descriptions using a limited set of anchors.

In chapter 5 I proposed solutions to the problems related to (1) learning domain-specific representations, and (2) using anchors for connecting symbolic and perceptual systems. For the first problem, I developed a system in which a simulated robot can learn to identify environmental directions and distances to boundaries from images, using range sensors. For this purpose the robot learned a

domain-specific representation, in which the distances to eight different cardinal directions were identified. This representation therefore could be used to symbolically describe both the distance and the relative direction of a point in the input image. I further illustrated that the same system could identify different locations within the environment by learning a similarity space among the direction and distance readings for each location (the property specific representation of the first task).

For the second problem, I demonstrated that the anchoring framework connects a symbolic system (Genesis Story Understanding System) to a vision system in order to solve a spatial problem. In this example, I showed how a simulated robotic arm can follow the instructions provided in English for replacing a cellphone battery. Again for the second problem, I showed that an AI agent can generate verbal descriptions from an environment, similar to those made by the participants of the observational study. I showed that anchor representations can be combined and processed by the story understanding system, so that it can then answer specific questions such as "Where is the red cup?". I also showed that having symbolic representations enables us to augment the initial descriptions with commonsense knowledge, so that a description generated by the anchors such as "There is a chair at the table" can be augmented to "There is a chair at the table. You can sit on the chair. You can eat dinner at the table."

The software I developed for the above examples enabled Genesis Story Understanding system to communicate with a vision system to understand an environment and solve spatial problems.

Finally, I created an observation dataset that includes 6.5 hours of explorations in three different virtual environments. This dataset can be used for training artificial agents and vision systems. In a previous study we showed this dataset can be used as a ground truth for image captioning tasks.

6.2 Open Questions and Next Steps

In this research, I focused on the limited scope of relationships between visual perception and inner stories. The idea that inner stories can model spatial experience requires further research in other sensory modalities such as aural and haptic senses. Another study can focus on the comparison of activities beyond visual exploration and verbal descriptions. How much an activity affects the types of inner stories we construct?

The artificial intelligence experiments demonstrated that the anchoring framework enables solving visual and spatial problems in simulated environments. A set of experiments designed in real environments can evaluate the anchoring framework in a broader context. Although I argue that the experiments designed in virtual environments are compatible with the ones that take place in physical environments, such an argument needs to be supported with further experiments.

Spatial experience is one of the fundamental aspects of human mind, so are inner stories. In this dissertation, I showed that framing human spatial experience as an inner story advances our understanding of how we humans understand and interact with our environment. Integration of symbolic and visual computing, as I suggested through the anchoring framework, will enable artificial intelligence systems to answer the question of "Where am I?".

Appendix A

Appendix: See, Act, and Tell Methodology

A.1 An overview of methodologies for observing spatial experiences

How can we obtain data from someone's spatial experience? Spatial experience is a particularly challenging process to observe because unlike other human activities, it requires access to what a person is *perceiving*, *thinking* and *doing* in an environment. It is only the *doing* part, if we take a very generous definition of doing, that appears as an externally visible aspect of spatial experience. Existing tools and procedures for capturing data from one's subjective experience operate under the assumption that there is a parallelism between observed behavior and underlying cognitive and perceptual processes. Taking this assumption to another level, I consider verbal descriptions that people make during their explorations a proxy for their cognitive processes. There are certain limitations to this approach, which I discuss in the next section.

In his study on daydreaming and artificial intelligence, Erik T. Mueller remarks on the challenge of obtaining data from the stream of consciousness (Mueller, 1990). He considers three relevant approaches: *retrospective reporting*, *think-aloud* and

event recording. In retrospective reporting, subjects are instructed to describe a previous experience (in Muller’s case their daydreams). This method is effective for understanding a person’s cognitive processes when the report is generated immediately after the experience (i.e. immediately after a person daydreams). However, the method becomes susceptible to distortions and omissions when the report is generated some time after the experiment, because it requires the subject to recall details from long-term memory (Mueller, 1990). In this context, event recording and the think-aloud protocol, both of which attempt to gather the data as the experiment unfolds, are more suitable methods.

A.1.1 Event Recording

Event recording or *experience sampling method* (ESM) is a methodology that allows sampling of an ongoing experience at a certain interval (Reis and Gable, 2000; Csikszentmihalyi and Larson, 2014) In this methodology, participants report their experiences, thoughts, or activities when a signaling device, such as a beeper, asks them to do so. Alternatively, participants may report whenever a predetermined activity occurs (for example, whenever a subject daydreams). Researchers use different coding techniques to categorize events or thoughts, from which the participant selects the most representative category for her or his activity. This method is particularly effective for studying long-term activities and for understanding fluctuations in experiences based on an activity context. It has also been argued that this method is effective for discovering intra-subject differences and correlations.

ESM requires the participant to actively report his or her ongoing experiences. This introduces challenges to obtaining data because not all participants may cooperate equally. Moreover, in certain contexts participants may introduce a bias to the study (for example when an employer uses this method to assess productivity of workers) (Csikszentmihalyi and Larson, 2014). In another variation of this methodology, instead of the participant reporting his or her own activities, an observer takes note of the events that have occurred at a predetermined interval.

This method is more suitable for observing behaviors of participants, and resembles another frequently used approach, the think-aloud protocol.

A.1.2 Think Aloud Protocol

Think aloud is one of the most frequently used methods for understanding people's cognitive processes when they are engaged in a certain activity such as problem solving or designing. This method involves participants thinking aloud, that is, verbally expressing what they are thinking during their activities. For example, a person solving a math problem might read the problem aloud, enumerate each step she or he is planning to take in order to solve the problem, express her or his confusion, or talk about some other idea that comes to her or his mind. During the process, an observer takes notes based on a predetermined protocol, codifying distinct events that might occur, such as the participant having difficulty drawing a diagram, or achieving a certain step in the problem. The well structured nature of the think-aloud protocol enables the observer to quantify individual events and perform comparisons among subjects. However, the process relies heavily on the participant's ability to verbalize his or her thoughts and the observer's ability to create a comprehensive protocol.

A.2 The use of Virtual Reality as an experimental tool

Virtual Reality is a technology distinct from traditional media in that it provides spatial perception and the sense of proprioception to the user (Olson et al., 2019). This creates a distinct feeling of being immersed in the presented media rather than being an external observer to it. There are multiple levels of immersion that can be provided with the range of devices and technologies that exist today. In the simplest version, the device provides a 360 degree spherical image and allows the viewer to gaze around the content. In this form, there is a limited feeling of immersion due to the lack of depth information and limited movement capacity (often denoted as

3-degrees-of-freedom, or 3DOF). The degree to which we can consider VR technology as a research tool for understanding human spatial experience is directly related to the level of immersion provided by the technology. 3DOF devices are not suitable for this task because they do not allow people to move around within the presented content. At the other end of the spectrum, there are 6DOF devices which provide additional freedom of movement as well as stereoscopic vision. These devices provide a level of immersion that is suitable for performing spatial experiments. In a previous study, in collaboration with Julia Litman-Cleper, I showed that spatial perception in VR and physical spaces are compatible (Zaman and Litman-Cleper, 2015). In this study, we compared a participant's distance judgment of simulated objects with the object's digital distances to the camera in the simulated environment and discovered that these two generally matched.

Several questions need to be addressed in the context of using VR as a research tool in the context of conducting experiments in the "wild" instead of conducting them in a laboratory setting. The first such question concerns using VR to perform visual exploration tasks instead of conducting experiments in actual physical environments. The most important reason for preferring VR over a physical environment is the ability to control the environmental properties and ensure that every user is presented with exactly the same environment. Real-world environments are highly dynamic and are easily affected by external variables such as time of day, other people using the space and interacting with the subjects, privacy of the subjects and others, and so on. Studies in the past have used physical spaces that focused on the use of a specific environment by its inhabitants (Ittelson, 1960). These studies often produced limited and highly contextualized results (patient behavior in hospitals, customer behavior in malls) and depended on subjective interpretations of the observer. In those settings, the observer necessarily shares the environment with their subjects, and thus introduces a bias into his or her experiment.

As an alternative to VR, researchers sometimes create mock-up environments for isolating subjects from the external world. This is a highly unwieldy and expensive method especially for experiments that require multiple, large-scale environments. In

contrast, VR enables researchers to create multiple isolated environments where one can capture a subject's experience from his or her own field of view. In this current research, where I explore the relationship between cognitive and perceptual processes, accessing the field of view (FOV) is critical to making assessments. This leads to second important advantage of using VR as a research tool: the breadth and precision of data collection. While one can record a video from a subject's FOV, for instance, by using a head mounted camera, the resulting video has to be interpreted by the experimenter. For example, in order to count how many times a subject looked outside the window, the video data has to be manually investigated by the researcher. On the other hand, simulated environments can encode any information, before or after the experiment, thus providing many opportunities for researchers to make queries and analysis. For example, one can reenact multiple experiments concurrently in a simulated environment and compare the behaviors of different subjects.

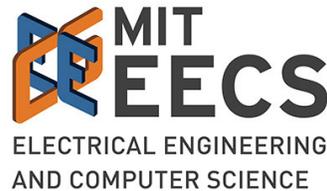
A.2.1 Limitations

One apparent limitation of a simulated environment is the perceived fakeness or the lack of realism in the presented content. In the past, together with the technical challenges and impracticality of VR devices, this phenomenon caused skepticism among researchers as to the use of virtual reality in the research settings. However, with the advent of computer graphics and VR technology, this problem has been mostly overcome. Today's systems can create highly realistic simulated environments, which greatly reduce the tendency of subjects to pay attention to the medium itself instead of the presented content. In my experiments, only a small portion of subjects (4/27) made remarks about the realism or the seeming fakeness of the environment. Even in those cases, I believe that subjects attended to these aspect mainly due to their personal interests in understanding the technology. I also observed that stereo-vision and 6DOF immersion provided a powerful sense of existence within the simulated environment.

Another limitation of the current VR technology is that participants's movements are restricted. Although 6DOF systems allow free movement in space,

there is a physical limit on how far a participant can move before the sensors lose tracking or the participant hits an obstacle in physical space. Currently, this problem is being overcome by introducing a secondary navigation method using a remote controller. Using a remote controller, a participant can either move towards a direction or "teleport" to another location. However, none of these methods provides the same experience as that of actually walking. The limited navigation capabilities of VR introduces challenges for developing experiments that require subjects to explore large-scale environments such as urban settings.

A.3 Consent Form Sample



CONSENT TO PARTICIPATE IN NON-BIOMEDICAL RESEARCH

Forming Contextual Descriptions of Virtual Reality (VR) Spaces

You will be asked to participate in a research study conducted by Cagri Hakan Zaman, Ph.D. Student in the Design and Computation group, Ainsley Sutherland, Director of Mediate VR, and Danielle Olson, Ph.D. Student in the Imagination, Computation, and Expression (ICE) Lab at the Massachusetts Institute of Technology (M.I.T.).

You were selected as a possible participant in this study because you have affiliation with MIT and/or the researchers running this study. You should read the information below, and ask questions about anything you do not understand, before deciding whether or not to participate.

- **PARTICIPATION AND WITHDRAWAL**

Your participation in this study is completely voluntary and you are free to choose whether to be in it or not. If you choose to be in this study, you may subsequently withdraw from it at any time without penalty or consequences of any kind. The investigator may withdraw you from this research if circumstances arise which warrant doing so.

- **PURPOSE OF THE STUDY**

The purpose of this study is to better understand how humans describe three-dimensional (3D) spaces in virtual reality (VR) using verbal descriptions.

- **PROCEDURES**

If you volunteer to participate in this study, we would ask you to do the following things:

1. We will ask you to sign an electronic consent form and return it to us via email before participation in the study.
2. Next, we will ask you to schedule a time during the week of the study to come into the lab, put on a virtual reality (VR) head-mounted display (HMD), look around a 3D space and describe what you see aloud, while we record you.

A.4 Verbal Description Samples

Below are samples of raw transcripts of verbal descriptions made by the participants.

Sample 1: Environment 1 Alrighty. Um. I see a weird I don't know if it's like a light. Oh yeah. It's definitely a light. Um. What could I do with it? I mean probably there's a light switch to turn it on. I could try to jump to touch it if I want. Um. I see this like painting or not painting but art piece that looks like a science photo wave. Um. With little lights above it. So I could probably turn that on. I could probably touch the thingy if I wanted to. I kind of want to, but I can't. Um. I see an outlet on the wall. So I could plug in my phone to charge it. Um. I see a weird swivelly chair thing. This is so weird to navigate.

Um. I could sit in the chair. I could knock the chair over. I could stand up in the chair.

Okay. Um. What else can I do with chair? I could put stuff in the chair, 'cause this set up actually kinda reminds me of my apartment 'cause of the windows and the like L-shaped couch thingy. Um. And we usually just throw all our crap in the chair instead of using it like an actual chair. Um. I see a carpet. I could lift up the part carpet and see what's underneath. Um. I could, if it were gross, vacuum the carpet. Is this like what you mean by like what I could do with it?

Okay. Uh. I see all these photos of blurred out faces. Which freak me out a little bit, 'cause there's no faces. Um. I could put the thing down so I don't have to look at them then 'cause they're creeping me out. Um. I could open these drawers. Is that what they are? On this like platform thing. It looks like there's a laptop. I could open that, and turn it on, and type, and it maybe connect it to the TV. Like if there's a cable or something in one of the drawers. Um. There's a TV mounted on the wall. I could turn on the TV. I could try to like move the TV to see what's behind it. So you get a better look at like cords and stuff behind it. Um. There is a cozy, co- What are these, those thingies called? Is that a cup or a cozy? I can't tell. Like you know those things that you like put on your like-your soda cans. Okay. I can't remember what they're called, but that thing I think, I could take that. I could

throw it. I could squish it. Um. I could throw it at the weird light thingy. Um. I could put it in the kitchen over there. Uh. There's a shoebox. No? Shoebox? Maybe kind of. Oh. Um. I could open the shoebox. I could crush the shoebox. I could put all these very creepy pictures in the shoebox. Um. What else? I could put the shoe box on that thing on the, uh, dresser drawer. There's a lamp in front of me that has a cord it looks like, but the cord is that going into the shoebox? It looks like it's going into the wall, but there's no outlet. So I'm a little confused by that. So I would try to like pull on the cord to figure out where it's actually going to 'cause I'm confused about that. Um. Is there a light bulb? There is a light bulb. I could take out the light bulb.

Uh. Is there a switch though? Uh. I don't see a switch of any kind. So I don't know how I would turn this on. So I'd probably just get confused by it and like ignore it. Um. There's a shelf thingy over here. Oh. I can like move through objects. Ha. I am standing inside the couch right now. Um. What is this? Is it just like a block of wood? Um. I would touch this, or push it, or like try to move it to figure out like maybe that something's behind it or something underneath it. Or maybe there's like a lid to it. I just like can't m- tell. Um. Now I'm looking outside the window. Oh my god. This is kinda terrifying.

No. Okay. I'm a little too close to the sun. Um. I see like windows and it looks like maybe a street, but I can't tell. Um. Okay. There's another weird lamp thing over here connected to the wall. Oh, this is making me feel woo. Uh. There's a what are you?

Okay. Uh. There's like a stereo over here. I could turn it on. Um. There's like a little compartment thing underneath it. So I probably open that to see if there's like speakers in there 'cause it it looks like just like the stereo part but not like any speakers. So I'd probably turn it on and then try to find the speakers. Um. There's a crap ton of books. Lots and lots of books. Uh. I Can I read what they say though? This does that say br- breathe king? What does it the mont- the monton maker? I don't know what any of that means. Um. So I'd probably open that. I could open up the books. I could move the books. Um. I could stack all those books and then

I could reach the weird light thingy better. Um. I also see that there's like little light thingies above the bookshelf, but that confuses me because if there's a light thingy, then it's probably illuminating something but that angle doesn't look right to illuminate the book shelf. So is there something up there maybe? So I'd want to climb up the book shelf and see if there's something up there. Um. There's a book called I Am a Cat, and it's very different color than all those other books. So I want to take that book and read it. Uh. And then more books, and books, and books. There's also oh, there's also a coffee table that I'm standing inside of. And okay, so now I'm not. So that's good. Um. It looks like magazines. I could take those magazines. I could close this one. I could reorder them, restack them so they look nicer, and I can put them on the book shelf.

Um. There's a couch. I could take all the pillows and stack them together to make one big pillow. Um. I could take all the cushions off the couch, uh, and make like a f- fort in this lovely space. Um. What else? I could fall asleep on the couch, which is what I'd probably do. Um. And, okay. I think that's it for the couch. Okay. Um. The Is this a door? I see, I feel like I'm getting t- twisted. I see like this, uh, black rectangle that might be a door, but there's no handle. So maybe it's like one of those fancy schmancy doors. Where you like push it instead of twist something or I don't know. Maybe it's a door. I'm confused by that. I don't like that. Um. Oh no, that's the door. Okay. I don't know what that thing is then. So now I see a door. Oh shit. And, uh wait. On the backside of the small rectangular door is like a big black wall. So maybe I don't know, someone got I'm confused.

Um. I see towels. I could take all those towels off the shelf. I could put them all on one shelf. I could put the towels into the washing machine. Um. I could take a towel with me for some reason. Um. I also see a washing machine which I could open. I could turn it on. I could unplug it. I could push a bunch of buttons and see Are there buttons? There's like a knob which I could push and do some stuff with. Um. It looks like there's, uh What are they called? Shelves. Doors. Little Whats' the word? Shelf thingies. Door thingy. Handle thingy that you could open up? Huh? Cabinets. Thank you. Cabinets in here. Or maybe that's a fridge. No it's definitely

a cabinet. Um. I see, uh, bottles of stuff. I could s- um, push the top of the bottom, bottle to like make stuff come out to see what's inside of it. I could open it up to see what's inside of it. Then I could use the towel to wipe off my hand of whatever stuff was inside of it if it's gross.

Um. I see a white box it looks like on a higher shelf that I can't reach. As well as a bunch of wicker baskets that I can't reach. So I could climb up the shelves to get to the baskets if I really needed to, but probably don't want to do that. Um. I see I don't know what to call this. There's like a like a it looks shiny. Like a like plastic maybe or porcelain or I don't know. And it's like a rectangle and it's like slightly jutting out of the wall. Um. But then it also continues like through the doorway. And I don't know what that is. So I could kick that. Um. I could like, like flick it or something to see like what it is or maybe like try to move it back and forth 'cause maybe it slides and that's why its' partially through the wall.

Um. Oh. I see a very big bathroom. Uh. Which is awesome. It's like bigger than my room, room. Um. And I see two what look mirrors. Except there's like gray. I can't see my face in them. It actually looks like a book shelf inside of the mirrors. Oh. Is it like I can see through into the room? Maybe.

Um. But it's like the wrong orientation I think. So, um, there's that and, uh, there's like a oblong sink with two faucets. I could turn on the faucets. Probably one's hot, one's cold. Um. Except there's no like hole in the sink. I think that that thingy's a hole but it's like it looks sealed. So maybe I could overflow the sinks if I wanted. Um. There's soap it looks like. So I could pump that and then there's like a cup. Which I could take with me to hold stuff if I wanted. Um. There's another wicker basket underneath the counter with nothing in it. I could take that to hold stuff in it like the towels or something else. Uh. It looks like there's bar soap on this one instead of the pump soap. And a thingy to hold your toothbrushes. I don't know what I could do with that. Um. There's a toilet. I well I could use the toilet. That's an obvious thing. Uh. There's toilet paper. Which I could take with me to use for something.

There's a plunger. Um. Which I could use. I could open up the toilet. I could

take the plunger. I could throw the plunger. There's like all these little bathroom mats which I could take with me and maybe put in the wicker basket to carry around for some reason. Um. I don't know what this is. There's a . is maybe it's art. Uh. It's like how would I describe this? You're like it's like a rectangular thing made of smaller rectangular things. Like slats almost but some of the slats are taken out. Maybe it's a weird radiator. I don't know. Uh. I could touch it. I could like try to break one of the slats off to figure out what that is.

Sample 2: Environment 2 Okay, so I'm in the middle of a, um, of a space with a large communal work desk in the center of it. Um, surrounded by Aeron chairs and computers, um, computer screens that are Macintosh, brand. Um, the walls of the space are brick, with some exposed, um, black painted steel I-beams, um, inset into the wall, surrounding the windows, um, supporting the ceiling. And, uh, and a few other locations. Um there is exposed duct work, here as well. Um, so there's a main line, running from the enclosed m-, conference room at one side. The one, at the interior corner of the space, um, running in an S, uh, an S pattern to the m-, farthest s-s, ex-extent of the space, which is a, um, an area with three raised steps and what appears to be a, window although there's nothing outside the window. It's just whiteness, it could also be a sliding door of the Japanese variety with paper instead of glass. Um, ceiling, uh, in this area as well as the rest of the space is, is wood with, um, small with, um, small rafters ab-, supported by the, um, steel, um, beams in the space.

Um, in terms of light fixtures there are, um, a series of, um, appendant fixtures with lamp, with, uh, a examination lamp head style, um, fixtures hanging from them. Um, they're black as well and they're suspended at about ten above the ground. Uh, more or less above the work area. Um, above the chairs in the work area. Um, there's an old fashion coffee machine here. Um, coffee machine has a large coffee cup graphic on it, with a small cookie on the, um, saucer. Um, and there is a series of, of buttons, um, that enable you to choose the type of coffee that you want to have. Um, there's also a copy machine, next to the coffee machine. Um, that, um, looks like it can do

collation and other complex tasks. Um, there's, there's a series of low filing cabinets that are black, uh, powder coated metal. Um, on the, on the floor above them are three, um, sequential photographs, um, mounted at the same height of, um, one, which appears to be like a, a pool and a sort of a, a Florida, um, tropical house.

Um, the other one is a rustic farm scene and the third one is another view of like a, a house with a pool. Um, farther down toward the conference room, there's a-another series of images on the wall arranged more sporadically. Um, that have similar depictions. Um, on the floor in this area there's another filing furniture, um, this one's subdivided by, um, by vertical pieces of wood into smaller cubbies that hold folders of various sizes. Um The, uh, meeting area over here, um, is separated from the main work area by, um, a double sl-, uh, face mounted sliding door. Um one, and so the left one opens left, the right one o-opens right. The, all the hardware for this door including the rails, um, the wheels, the handle and the strips, uh, separating the door from, um, the frame are, very shiny chrome metal. Um, inside the room, um, there's a s-, a low meeting table surrounded by eight blue Eames chairs with wire feet.

Um, the walls in this room look like they're created of plywood, um, strips, uh, that cover most of the surface area of the wall except for the corners, which are in brick. Um, light fixtures are the same black hanging lamps that are shown in the, it, uh, that are featured in the, other part of the, of this environment. Um, the table has a few books on it, um, the table itself is white and, um, it's f-, uh, it's shaped in such a way that, uh, it's s-s-somewhere between a rectangle and, uh, an ellipse. There are a few bicycles leaning against the wall. One by the window and one just behind that. Um, behind the two bicycles there's a motorcycle. Um, motorcycle is propped out on its kickstand. Behind the motorcycle, um, is what appears to be a, um, garage door style, uh, rolling door, segmented door. Um, that would retract and open to the outside in order to let the people on the bikes and, and motorcycle, um, exit.

There's a, a small hand rail, um, there's a small hand rail in the, um, just to the right of the, of the entrance garage door. Um, which has, uh, some n-, two with

gloves, gloves at the top. And intermediary, uh, vertical styles. Um, just by the entrance door, t-the man door, there are a series of green lock, uh, l-, cubby style lockers. Three wide by five tall, um, and in between the two, um, sets of lockers there are two s-, low file cabinets, um, that are brown in color. The door itself is, uh, a case metal door, with a blank gray panel of metal mounted about it. Um, there are several skylights in the space. Um, that are letting direct sunlight down, uh, wash down into the space. So, and the skylights are about, uh, six feet by six feet and they are mounted, er, and they are inserted into the space, above the, um, above the, the metal beams.

Um, the floor is either a carpet or a, or probably a, a r-, exposed concrete. Um, and this little informal meeting area there are three designer chairs also, of sort like bent plywood, um, with some up-upholstery. They're three legged and they're recline, they're recliner, no, they're, um, sort of relaxing positions sort of chairs. And there's a, um, a wire base, um, circular table in between the three chairs with a book on it. Um, in this area also there's a white carpet, a blank white carpet. Um At the end of the, of the large worktable there's a trash can with a series of architectural drawings in it. Um, the drawings are rolled and they have some markings on them that I cannot easily see. Um, there's one drawing unrolled on the table that looks like architectural plans. Um, yeah. That's about it for this space, I'd say.

Sample 3: Environment 3 Mm-hmm. Okay. So in the living room or seating area. This house is really open concept. Um. So all the rooms are basically connected there's no hallways between any two rooms. Uh. They sort of just lead into each other. Oh. I haven't described the floor. The floor is like, um, hardwood, uh, panels. It's really nice.

Um. Over here in this corner there's a little chair by itself. Um. It seats one person. Its yellow. Like a mustard yellow to match the recliner over there. Um. It appears to have like a wire base and, um, it's also like a modern style. It's very round. Um. Like if you sat in it, it just kind of cups you.

Uh. Behind it is like a braided like ficus tree I think. Um. I don't know if it's

fake or real. Obviously, everything in here is fake, but in the context of this room I don't know if it's supposed to be a living tree or not. But if it's a living tree, uh, I could water it. Uh. I could, uh, prune it and I could pick leaves off of it. Uh. If it's a plastic tree, I probably wouldn't want to do any of those things. Um. I could move it to a different location though. Um. It's in a black pot that's round, um, about a foot and a half tall and there appears to be dirt or, um, like fiber that the tree is growing out of or positioned in, in the center of the pot.

Um. The corner of this room is interesting, uh, because this room is not cubic. Um. It seems to have a a rounded ceiling, um, that arcs up onto the second floor. Oh. I've lost my place. Um.

Okay. Well I may as well note that there are stairs in this direction. Uh. They're black and then lead up to the second floor. Um. Which I can see is behind this balcony you here. And, uh, there's one, two, three, four, five, six, seven. There's like 10 steps. Um. There's no hand rail so these stairs would be dangerous to ascend. Um. But if I were allowed to in this environment, I could walk up them to get to the second floor. Um.

I might mention that the ceiling here is a green and it almost looks carpeted, but it might just be a textured paint. Um. I've never seen a green ceiling before. Uh. But it, that is interesting. I don't know if I like it or not.

There's a sky light above the island in the kitchen. Um. It appears dark. Um. So it might be tinted glass. But I can look up through this window here in the ceiling and, uh, see the sky a little bit. Um. It's about six feet long by like four feet wide, and, uh, it's embedded, sort of, in the ceiling by about half a foot. Oh. I guess if the second floor is up there, this isn't really a sky light is it? This is just some kind of indentation in the ceiling here so disregard what I said about being able to look up through this.

Um. There's a table here behind the kitchen island. Um. It's a solid wood table. Uh. It looks really nicely made. Um. It has nice, like thick legs on it, and, um, it's pretty big. It looks like it would seat about, mm, eight people. So there are three chairs around here that are white, very modern looking. They look like they came

out of like the early 2000s and, uh, they're round. Uh. Each chair seats one person. So I could sit in these chairs. I could move them back and forth. I could scoot them under the table to get them out of the way. Um. I could probably stack them on top of each other for easier storage.

Um. And then next to me, right here, are two different chairs. Uh, the don't have back on them. They're just tall stools. Uh. I think these are meant less for sitting at the table and more for sitting at this counter. I probably wouldn't be able to stack these on each other, but I can move them around the room, or I could sit in them. Um. They don't look very comfortable.

There is a cup of black liquid here on this table. Um. I'm guessing it's coffee. There's a spoon in it. Uh. I could pick this cup up and drink the black liquid out of it if I felt like that, or I could pour it out in the sink. Um. I could use the spoon to stir this black liquid, or, um, take some of it out and put it somewhere else. Um. I could pick up this cup and move it anywhere I wanted to, or move the spoon somewhere else.

Um. There's a magazine next to the cup. It says it has some advertisements in it for resorts or like summer homes. I can't tell where they are. Uh. But the pages say, "Visit in June," and "Vacation time." Uh. These look like very expensive homes here in this magazine. I could probably pick this magazine up and flip through the pages. Um. I could probably tear some pages out. I could do that with any book in this house, but with a magazine it seems more likely that I would be tearing pages out of that than a book. Um. I could close it. Uh. This magazine has a soft cover. Um. Not soft like the blanket, but soft as in, um, not rigid. Uh. I could roll this magazine up, um, I could fold it in half. Um. That's about it.

Uh. There's a bowl of what look like cherries here on the table in the very center of the table. It's a white bowl. Um. It's about like eight inches round. Uh. These cherries are huge, um, and blackish. I could take one of these cherries and eat it. Um. I could use these to cook with. Uh. I could dump them out anywhere. I could move this bowl anywhere. Um.

Oh. I might mention that the cup, and this bowl, and also this, uh, planter here

on the table are all ceramic. Um. And if I dropped them, they would shatter and break. Um. There is a planter here on this table. It has a flower arrangement in it. With some ferns and blue flowers that are sort of like sunflowers except blackish-blue. Um. It's not a very pretty arrangement. It's sort of neutral colors. I've gotten myself mixed up. The planter is about a foot tall and again, like eight inches round. Um. It's pretty basic. It's just a cylinder.

Uh. Outside. So one wall of the room here is just a big window, uh, with a huge sliding door. So this window is taller than I am. Um. I would say like seven feet tall, and it stretches like 12 or 13 feet across this wall. Uh. The sliding doors are two big squares of glass, and, um, there's a handle right here that I could pull to move the door, uh, left to open it, or back to the right to close it again. Um. Outside, uh, looks like I'm in the mountains. It's winter so there's snow everywhere on these evergreen trees. There's a lot of trees outside. Um. So I could open this door and I could step outside into the snow. Um. Or I could just leave it open and feel the breeze come in. Um. If this door's glass is breakable, I could break this glass. Probably wouldn't want to. Um. It has a metal border.

Um. Behind me here is another couch. But this couch looks like it could seat um, whoa. This couch looks like it could seat six people comfortable. Uh. It's a darker gray than the other couch. But it's a dark gray with like a textured pattern. It's also a modern, uh, blocky looking design, and doesn't look very comfortable. But the pillows on this couch are, um, they match the couch material. Uh. Unlike the other one. I could take these pillows and throw them around. Uh. They're soft, so I would probably want to put some of them behind me before I sat down, uh, to cushion myself. Um. I could sit on this couch. I could lay down on this couch. I could sleep on this couch.

Um. And behind it on the wall is, um, a picture. Painted onto the wall or, um, or it might be a decal that someone stuck to the wall, but it's a bunch of squares. Uh. Each square is a different shade of red, or orange, or yellow, or brown. Um. It doesn't appear to depict anything. Um. It just is pretty. Um. At the intersection of each corner of the squares there's like white Xs. Um. Here in front of the couch is a

purple rug, uh, that's about as wide as the couch is. Uh. Except it's a little offset. That's kinda bothering me actually. I might want to shove this couch over about a foot to make the edges match up with the rug.

Uh. But there's a glass table on top of the couch. Which is interesting. Um. It looks like a single piece of glass or plastic, um, that just has the edges folded down to form, uh, something like legs. But it's like solid walls instead of legs, and it's, uh, sort of rounded at the edges. It's a pretty cool table. There's nothing on it. I could move this table back and forth as well. Um.

Here to the left of the couch is a television on top of a set of shelves. Um. The television is pretty big. Um. Might say, this is just an estimate, like 40 something inches diagonally. Um. It's a flat screen. It's modern. It's not wall mounted. It's just sitting on top of like a a metal leg. Um. To either side of the television are what look like thin speakers, uh, also very modern looking. Uh. There about as tall as the TV. Um. And under the television on the shelf below it are three boxes. Uh. They're made of wood. The same wood as these, uh, shelves and they're cubes about and about like a foot wide, and long, and deep. Um. It doesn't look like you could store anything in them. They seem purely decorative, but I could probably pull them out and examine them to see if they open at all for storage.

Um. There's like three shelves here. So on the left shelf, there's a, a small black box that has knobs on the front. Um. It looks like this is for controlling he channel and volume on the TV, uh, where it might control the speakers, um, or be some kind of radio. Um. It's like three inches tall and like a foot square on top.

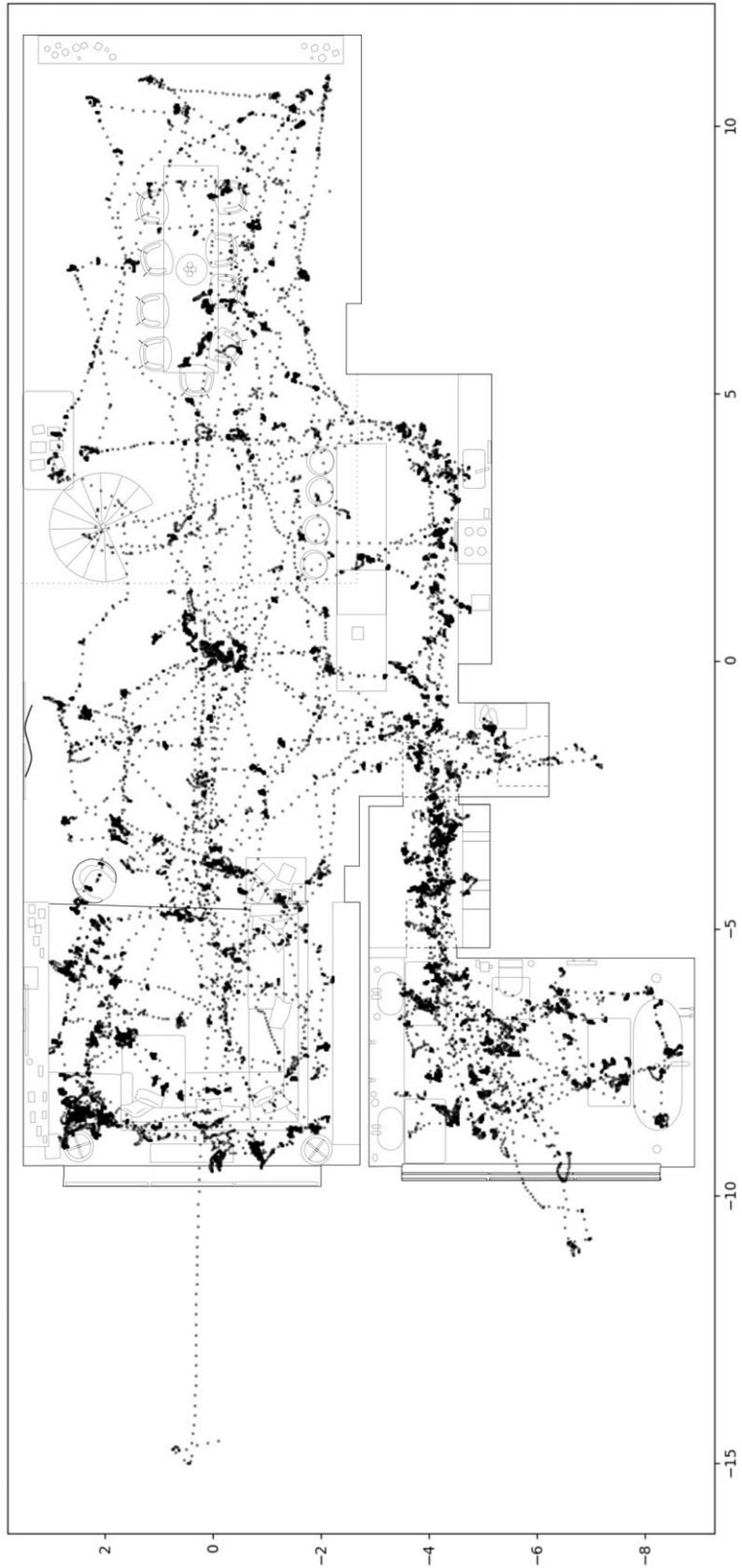
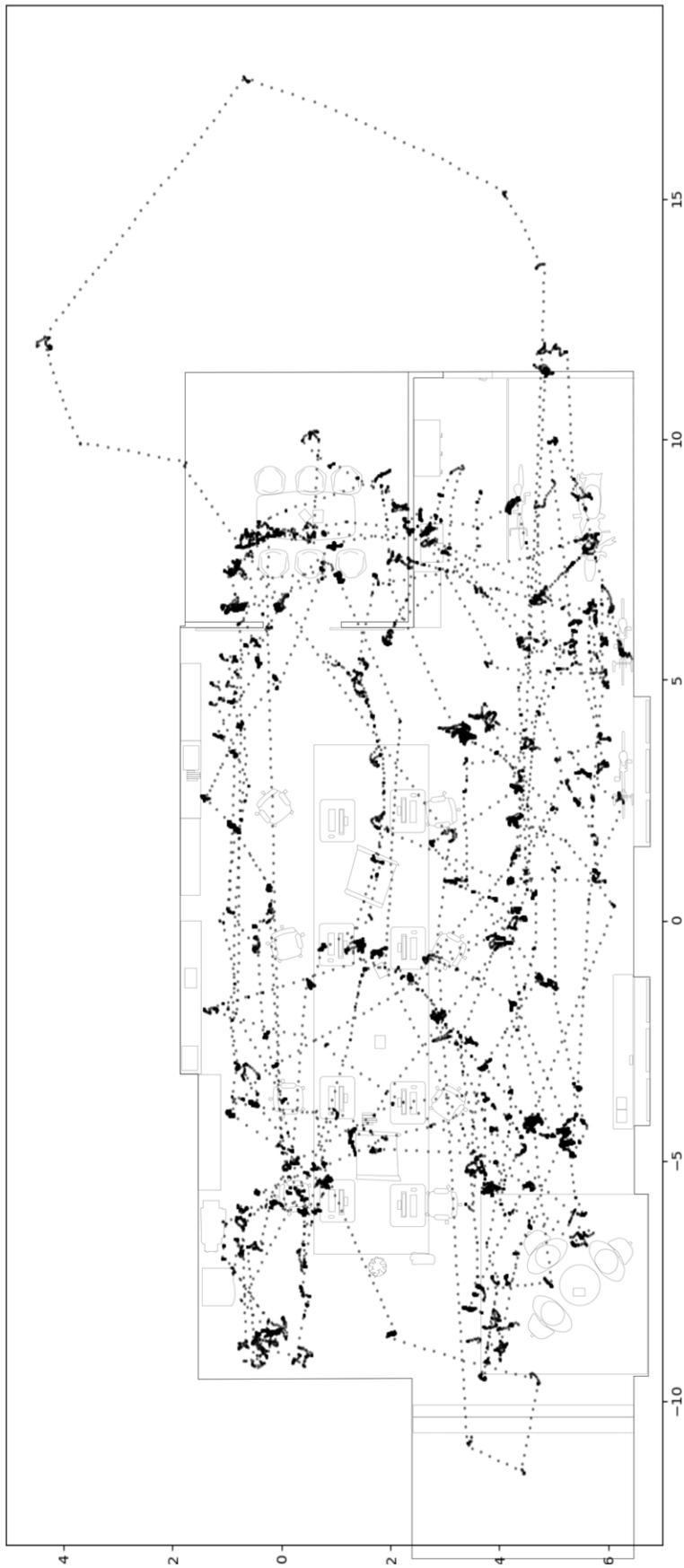


Figure A-1: Environment 1 Movement Overlay



214
Figure A-2: Environment 2 Movement Overlay

Bibliography

- Angela Andersen. Empathy and the creation of virtual space: Review of september 55, kellery gallery, mit. *Architectural Histories*, 6(1), 2018.
- Rudolf Arnheim. *Visual Thinking*. University of California Press, 2004.
- Richard C Atkinson and Richard M Shiffrin. The control of short-term memory. *Scientific American*, 225(2):82–91, 1971.
- Jon Barwise and John Perry. Situations and attitudes. *The Journal of Philosophy*, 78(11):668–691, 1981.
- Serge Belongie, A Rabinovich, A Vedaldi, C Galleguillos, and E Wiewiora. Objects in context. In *Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV 2007), Rio de Janeiro, Brazil, 14-20 October 2007*, 2007.
- Michael L. Benedikt. To take hold of space: isovists and isovist fields. *Environment and Planning B: Planning and Design*, 6(1):47–65, 1979.
- George Berkeley. *A New Theory of Vision and Other Select Philosophical Writings*. Number 483. Dent, 1922.
- Robert C Berwick and Noam Chomsky. *Why only us: Language and evolution*. MIT Press, 2016.
- 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. ISSN 0010-0285. doi: [https://doi.org/10.1016/0010-0285\(82\)90007-X](https://doi.org/10.1016/0010-0285(82)90007-X). URL <http://www.sciencedirect.com/science/article/pii/001002858290007X>.
- Otto Friedrich Bollnow. Lived-space. *Philosophy Today*, 5(1):31–39, 1961.
- Pierre Bourdieu. *Outline of a Theory of Practice*, volume 16. Cambridge University Press, 1977.
- Rodney A Brooks. Symbolic reasoning among 3-d models and 2-d images. *Artificial intelligence*, 17(1-3):285–348, 1981.
- Warner Brown. Spatial integrations in a human maze. *University of California Publications in Psychology*, 1932.

- Yuanzhouhan Cao, Tianqi Zhao, Ke Xian, Chunhua Shen, Zhiguo Cao, and Shugong Xu. Monocular depth estimation with augmented ordinal depth relationships. *IEEE Transactions on Image Processing*, 2018.
- Zhe Chen, Fabian Kloosterman, Emery N Brown, and Matthew A Wilson. Uncovering spatial topology represented by rat hippocampal population neuronal codes. *Journal of Computational Neuroscience*, 33(2):227–255, 2012.
- Herbert H Clark. Space, time, semantics, and the child. In *Cognitive Development and Acquisition of Language*, pages 27–63. Elsevier, 1973.
- Silvia Coradeschi and Alessandro Saffiotti. An introduction to the anchoring problem. *Robotics and Autonomous Systems*, 43:85–96, 2003.
- Mihaly Csikszentmihalyi and Reed Larson. Validity and reliability of the experience-sampling method. In *Flow and the Foundations of Positive Psychology*, pages 35–54. Springer, 2014.
- Michelle de Certeau. *The Practice of Everyday Life*. Number v. 1. University of California Press, 2011. ISBN 9780520271456. URL https://books.google.com/books?id=-Csl_AAoUT8C.
- John Dewey and Arthur Fisher Bentley. *Knowing and the Known*. Number 111. Beacon Press Boston, 1960.
- Halle R Dimsdale-Zucker, Maureen Ritchey, Arne D Ekstrom, Andrew P Yonelinas, and Charan Ranganath. Ca1 and ca3 differentially support spontaneous retrieval of episodic contexts within human hippocampal subfields. *Nature Communications*, 9(1):294, 2018.
- José P Duarte. A discursive grammar for customizing mass housing: the case of siza’s houses at malagueira. *Automation in Construction*, 14(2):265–275, 2005.
- Gilberto Echeverria, SÃlverin Lemaignan, Arnaud Degroote, Simon Lacroix, Michael Karg, Pierrick Koch, Charles Lesire, and Serge Stinckwich. Simulating complex robotic scenarios with morse. In *SIMPAR*, pages 197–208, 2012. URL <http://morse.openrobots.org>.
- TM Evermotion. Evermotion 3d models, 2019. URL <http://www.evermotion.org>.
- M Piedade Ferreira, Duarte Cabral De Mello, and José Pinto Duarte. The grammar of movement: A step towards a corporeal architecture. *Nexus Network Journal*, 13(1):131–149, 2011.
- Winthrop Nelson Francis. *A manual of information to accompany A standard sample of present-day edited American English, for use with digital computers*. Department of Linguistics, Brown University, 1971.

- Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.
- Katharina Galor, Donald H Sanders, Andrew Willis, David Cooper, Benjamin Kimia, Gabriel Taubin, and Oren Tal. Semi-automated data capture and image processing: new routes to interactive 3d models. In *Third International Conference on Remote Sensing in Archaeology*, pages 179–188, 2009.
- James J Gibson. *The Perception of the Visual World*. Houghton Mifflin, 1950.
- James J Gibson. *The Ecological Approach to Visual Perception*. Psychology Press, 2014.
- Leilani H. Gilpin, Cagri Zaman, Danielle Olson, and Ben Z. Yuan. Reasonable perception: Connecting vision and language systems for validating scene descriptions. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2018, Chicago, IL, USA, March 05-08, 2018*, pages 115–116, 2018. doi: 10.1145/3173386.3176994. URL <https://doi.org/10.1145/3173386.3176994>.
- Nelson Goodman. *Ways of Worldmaking*, volume 51. Hackett Publishing, 1978.
- Adolfo Guzmán. Decomposition of a visual scene into three-dimensional bodies. In *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I, AFIPS '68 (Fall, part I)*, pages 291–304, New York, NY, USA, 1968. ACM. doi: 10.1145/1476589.1476631. URL <http://doi.acm.org/10.1145/1476589.1476631>.
- Edward Twitchell Hall. *The Hidden Dimension*, volume 609. Garden City, NY: Doubleday, 1910.
- Feng Han and Song-Chun Zhu. Bottom-up/top-down image parsing with attribute grammar. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):59–73, 2008.
- Gary Carl Hatfield. *The Natural and the Normative: Theories of Spatial Perception from Kant to Helmholtz*. MIT Press, 1990.
- Martin Heidegger. Building dwelling thinking. *Poetry, Language, Thought*, 154, 1971.
- Bill Hillier. *Space is the machine: a configurational theory of architecture*. Cambridge University Press, 1999.
- David H Hubel and Torsten N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148(3):574–591, 1959.
- Edward Hussey. *Physics: books III and IV*, volume 2. Clarendon Aristotle Series, 1983.

- Edwin Hutchins. *Cognition in the Wild*. Number 1995. MIT Press, 1995.
- Tim Ingold. *The Perception of the Environment: Essays on Livelihood, Dwelling and Skill*. Routledge, 2002.
- Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1469–1482, 2013.
- William H Ittelson. *Visual Space Perception*. Springer, 1960.
- William H Ittelson. Environmental perception and urban experience. *Environment and Behavior*, 10(2):193–213, 1978.
- William James. *The principles of psychology*, 1890.
- Jürgen Joedicke. *Raum und form in der architektur. Space and form in architecture*. 1985.
- Mark Johnson. *Embodied mind, meaning, and reason: How our bodies give rise to understanding*. University of Chicago Press, 2017.
- Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. *CoRR*, abs/1610.01465, 2016. URL <http://arxiv.org/abs/1610.01465>.
- Russell Kahl. *Selected writings of hermann helmholtz*. 1878.
- Immanuel Kant. *Critique of Pure Reason, translated by NK Smith*. London: Macmillan, 1933(1787).
- Terry Knight. *Applications in architectural design and education and practice*. 1999.
- Terry Knight and George Stiny. Making grammars: from computing with shapes to computing with things. *Design Studies*, 41:8–28, 2015.
- Terry W. Knight. Transformations of de stijl art: The paintings of georges vantongerloo and fritz glarner. *Environment and Planning B: Planning and design*, 16(1):51–98, 1989.
- Terry W Knight. Color grammars: the representation of form and color in designs. *Leonardo*, pages 117–124, 1993.
- Kurt Koffka. *Principles of Gestalt psychology*. Routledge, 2013.
- Erland Körner. Working with physically-based shading: a practical approach. <https://blogs.unity3d.com/2015/02/18/working-with-physically-based-shading-a-practical-approach/>, 2015. [Online; accessed 16-September-2019].

- Stephen Michael Kosslyn. *Image and mind*. Harvard University Press, 1980.
- Adam Davis Kraft. *Vision by Alignment*. PhD thesis, Massachusetts Institute of Technology, 2018.
- Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016.
- Barbara Landau and Ray Jackendoff. “What” and “where” in spatial language and spatial cognition. *Behavioral and brain sciences*, 16(2):217–238, 1993.
- Ronald W Langacker. *Foundations of Cognitive Grammar: Theoretical Prerequisites*, volume 1. Stanford University Press, 1987.
- Jean Lave. *Cognition in practice: Mind, mathematics and culture in everyday life*. Cambridge University Press, 1988.
- Hermann Lotze. *Outlines of Logic and of Encyclopaedia of Philosophy: Dictated Portions of the Lectures of Hermann Lotze*, volume 6. Ginn, 1892.
- Setha M. Low and Denise Lawrence-Zúñiga. *Anthropology of Space and Place: Locating Culture*. Wiley Blackwell Readers in Anthropology. Wiley, 2003. ISBN 9780631228776. URL <https://books.google.com/books?id=XC7cnQEACAAJ>.
- Kevin Lynch. *The image of the city*, volume 11. MIT Press, 1960.
- Ernst Mach. *Contributions to the Analysis of the Sensations*. Open court publishing Company, 1897.
- Elizabeth Marozzi and Kathryn J. Jeffery. Place, space and memory cells. *Current Biology*, 22(22):R939 – R942, 2012. ISSN 0960-9822. doi: <https://doi.org/10.1016/j.cub.2012.10.022>. URL <http://www.sciencedirect.com/science/article/pii/S0960982212012079>.
- David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA, 1982. ISBN 0716715678.
- Mohan Matthen. *Seeing, Doing, and Knowing: A Philosophical Theory of Sense Perception*. Clarendon Press, 2005.
- Maurice Merleau-Ponty. *The Primacy of Perception: And Other Essays on Phenomenological Psychology, the Philosophy of Art, History, and Politics*. Northwestern University Press, 1964.
- Maurice Merleau-Ponty. *Phenomenology of Perception*. Routledge, 2013.

- Michael J Milford, Gordon F Wyeth, and David Prasser. Ratslam: a hippocampal model for simultaneous localization and mapping. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, volume 1, pages 403–408. IEEE, 2004.
- Susanna Millar. *Space and Sense*. Psychology Press, 2008.
- Marvin L. Minsky. *Society Of Mind*. Touchstone book. Simon & Schuster, 1988. ISBN 9780671657130. URL <https://books.google.com/books?id=bLDL11fRpdkC>.
- Daniel R Montello, Alinda Friedman, and Daniel W Phillips. Vague cognitive regions in geography and geographic information science. *International Journal of Geographical Information Science*, 28(9):1802–1820, 2014.
- Erik T Mueller. *Daydreaming in Humans and Machines: A Computer Model of the Stream of Thought*. Intellect Books, 1990.
- Ulric Neisser. *Memory Observed: Remembering in Natural Contexts*. Macmillan, 2000.
- Walter Netsch. Forms as process. *Progressive Architecture*, 50(3):94–115, 1969.
- Nora S Newcombe and Janellen Huttenlocher. *Making space: The development of spatial representation and reasoning*. MIT Press, 2003.
- Alva Noë. *Action in perception*. MIT Press, 2004.
- Christian Norberg-Schulz. *Existence, space & architecture*. New York: Praeger, 1971.
- Christian Norberg-Schulz. *Genius loci: towards a phenomenology of architecture*. Rizzoli, 1980. ISBN 9780847802876. URL <https://books.google.com/books?id=FLYkAQAAMAAJ>.
- Robert M Ochshorn and Max Hawkins. Gentle.
- MR O'Connor. *Wayfinding: The Science and Mystery of how Humans Navigate the World*. St. Martin's Press, 2019.
- John O'Keefe and Lynn Nadel. *The Hippocampus as a Cognitive Map*. Oxford: Clarendon Press, 1978.
- John O'Keefe, Neil Burgess, James G Donnett, Kathryn J Jeffery, and Eleanor A Maguire. Place cells, navigational accuracy, and the human hippocampus. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 353(1373):1333–1340, 1998.
- Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer vision*, 42(3):145–175, 2001.

- Aude Oliva and Antonio Torralba. Chapter 2 building the gist of a scene: the role of global image features in recognition. In S. Martinez-Conde, S.L. Macknik, L.M. Martinez, J.-M. Alonso, and P.U. Tse, editors, *Visual Perception*, volume 155 of *Progress in Brain Research*, pages 23 – 36. Elsevier, 2006. doi: [https://doi.org/10.1016/S0079-6123\(06\)55002-2](https://doi.org/10.1016/S0079-6123(06)55002-2). URL <http://www.sciencedirect.com/science/article/pii/S0079612306550022>.
- Aude Oliva, Soojin Park, and Talia Konkle. Representing, perceiving and remembering the shape of visual space. *Vision in 3D environments*, pages 308–339, 2011.
- Danielle Marie Olson, Elisabeth Ainsley Sutherland, Cagri Hakan Zaman, and D. Fox Harrell. Foundations of interaction in the virtual reality medium. In Newton Lee, editor, *Encyclopedia of Computer Graphics and Games*. Springer, 2019. doi: 10.1007/978-3-319-08234-9_174-1. URL https://doi.org/10.1007/978-3-319-08234-9_174-1.
- J Kevin O’Regan and Alva Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5):939–973, 2001.
- Mine Özkar. Visual schemas: pragmatics of design learning in foundations studios. *Nexus Network Journal*, 13(1):113, 2011.
- Ryan T Philips and V Srinivasa Chakravarthy. A global orientation map in the primary visual cortex (v1): Could a self organizing model reveal its hidden bias? *Frontiers in Neural Circuits*, 10:109, 2017.
- James B Ranck Jr. Studies on single neurons in dorsal hippocampal formation and septum in unrestrained rats: Part i. behavioral correlates and firing repertoires. *Experimental Neurology*, 41(2):462–531, 1973.
- Terry Regier, Paul Kay, and Richard S. Cook. Focal colors are universal after all. *Proceedings of the National Academy of Sciences*, 102(23):8386–8391, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0503281102. URL <https://www.pnas.org/content/102/23/8386>.
- Harry T Reis and Shelly L Gable. Event-sampling and other methods for studying everyday experience. *Handbook of Research Methods in Social and Personality Psychology*, 196, 2000.
- Eleanor H Rosch. On the internal structure of perceptual and semantic categories. In *Cognitive Development and Acquisition of Language*, pages 111–144. Elsevier, 1973.
- Deb Roy, Rupal Patel, Philip DeCamp, Rony Kubat, Michael Fleischman, Brandon Roy, Nikolaos Mavridis, Stefanie Tellex, Alexia Salata, Jethran Guinness, et al. The human speechome project. In *International Workshop on Emergence and Evolution of Linguistic Communication*, pages 192–196. Springer, 2006.

- Daniela Schiller, Howard Eichenbaum, Elizabeth A. Buffalo, Lila Davachi, David J. Foster, Stefan Leutgeb, and Charan Ranganath. Memory and space: Towards an understanding of the cognitive map. *Journal of Neuroscience*, 35(41):13904–13911, 2015. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.2618-15.2015. URL <https://www.jneurosci.org/content/35/41/13904>.
- D Schoen. The reflective practitioner. 1983.
- Donald A Schön. Designing: Rules, types and worlds. *Design studies*, 9(3):181–190, 1988.
- Maxine Sheets-Johnstone. *The Primacy of Movement*, volume 82. John Benjamins Publishing, 2011.
- Anna Shusterman and Elizabeth Spelke. Language and the development of spatial reasoning. *The Innate Mind: Structure and Contents*, pages 89–106, 2005.
- Herbert A Simon. The sciences of the artificial. 1969.
- Elizabeth S Spelke. Developing knowledge of space: Core systems and new combinations| 5. *The Languages of the Brain*, 6:239, 2002.
- Herbert Spencer. *The Principles of Psychology*, volume 1. Appleton, 1895.
- Roger W Sperry. Neurology and the mind-brain problem. *American Scientist*, 1952.
- Larry R Squire and Pablo Alvarez. Retrograde amnesia and memory consolidation: a neurobiological perspective. *Current Opinion in Neurobiology*, 5(2):169 – 177, 1995. ISSN 0959-4388. doi: [https://doi.org/10.1016/0959-4388\(95\)80023-9](https://doi.org/10.1016/0959-4388(95)80023-9). URL <http://www.sciencedirect.com/science/article/pii/0959438895800239>.
- George Stiny. *Shape: Talking About Seeing and Doing*. MIT Press, 2006.
- George Stiny. What rule (s) should i use? *Nexus Network Journal*, 13(1):15–47, 2011.
- George Stiny and James Gips. Shape grammars and the generative specification of painting and sculpture. In *IFIP Congress (2)*, volume 2, 1971.
- George Stiny and William J Mitchell. The grammar of paradise: on the generation of mughul gardens. *Environment and Planning B: Planning and design*, 7(2):209–226, 1980.
- Michela Tacca. Commonalities between perception and cognition. *Frontiers in Psychology*, 2:358, 2011. ISSN 1664-1078. doi: 10.3389/fpsyg.2011.00358. URL <https://www.frontiersin.org/article/10.3389/fpsyg.2011.00358>.
- Leonard Talmy. How language structures space. In *Spatial Orientation*, pages 225–282. Springer, 1983.

- Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022): 1279–1285, 2011.
- Edward Chace Tolman. *Purposive Behavior in Animals and Men*. University of California Press, 1949.
- Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.
- Yi-Fu Tuan. *Space and Place: The Perspective of Experience*. University of Minnesota Press, 1977.
- Alan M. Turing. Computing Machinery and Intelligence. *Mind*, LIX(236):433–460, 10 1950. ISSN 0026-4423. doi: 10.1093/mind/LIX.236.433. URL <https://doi.org/10.1093/mind/LIX.236.433>.
- R Steven Turner. *In the Eye’s Mind: Vision and the Helmholtz-Hering Controversy*, volume 227. Princeton University Press, 2014.
- Barbara Tversky. Spatial thought, social thought. *Spatial Dimensions of Social Thought*, 18:17–38, 2011.
- TM Unity. Unity game engine, 2019. URL <http://www.unity3d.com>.
- Carolyn Unsworth. Cognitive and perceptual dysfunction. *Physical Rehabilitation*, pages 1149–1185, 01 2007.
- Hendrik A.H.C. van Veen, Hartwig K. Distler, Stephan J. Braun, and Heinrich H. BÄijlthoff. Navigating through a virtual city: Using virtual reality technology to study human action and perception. *Future Generation Computer Systems*, 14(3):231 – 242, 1998. ISSN 0167-739X. doi: [https://doi.org/10.1016/S0167-739X\(98\)00027-2](https://doi.org/10.1016/S0167-739X(98)00027-2). URL <http://www.sciencedirect.com/science/article/pii/S0167739X98000272>. Virtual Reality in Industry and Research.
- David L. Waltz. *Generating Semantic Descriptions from Drawings of Scenes with Shadows*. PhD thesis, Massachusetts Institute of Technology, 1970.
- Max Wertheimer. *On Perceived Motion and Figural Organization*. MIT Press, 2012.
- William Hollingsworth Whyte. *The Social Life of Small Urban Spaces*. Conservation Foundation Washington, DC, 1980.
- Matthew A Wilson and Bruce L McNaughton. Dynamics of the hippocampal ensemble code for space. *Science*, 261(5124):1055–1058, 1993.
- Patrick H. Winston. *Learning structural descriptions from examples*. PhD thesis, Massachusetts Institute of Technology, 1970.

- Patrick H. Winston. The genesis story understanding and story telling system: A 21st century step toward artificial intelligence. Memo 019, Center for Brains Minds and Machines, MIT, 2014.
- Patrick H. Winston and Dylan Holmes. The genesis enterprise: Taking artificial intelligence to another level via a computational account of human story understanding, 2019. URL <https://dspace.mit.edu/handle/1721.1/119651>. Online, accessed 11 December 2019.
- Patrick Henry Winston. The strong story hypothesis and the directed perception hypothesis. In Pat Langley, editor, *Technical Report FS-11-01, Papers from the AAAI Fall Symposium*, pages 345–352, Menlo Park, CA, 2011. AAAI Press.
- Wilhelm Max Wundt and Charles Hubbard Judd. *Outlines of Psychology*, volume 1. Scholarly Press, 1897.
- Zhutian Yang and Patrick H. Winston. Learning by asking questions and learning by aligning stories: How a story-grounded problem solver can acquire knowledge, 2017. URL <https://dspace.mit.edu/handle/1721.1/119668>. Online, accessed 11 December 2019.
- Cagri Hakan Zaman. An inquiry into active spatial language using virtual reality. In *Cognitive Processing*, volume 19, pages S54–S54. Springer Heidelberg, 2018.
- Bruno Zevi. *Architecture as space: how to look at architecture*. Horizon Press, 1974.
- Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014a.
- Bolei Zhou, Liu Liu, Aude Oliva, and Antonio Torralba. Recognizing city identity via attribute analysis of geo-tagged images. In *European Conference on Computer Vision*, pages 519–534. Springer, 2014b.
- Jordan Zlatev. Spatial semantics. *The Oxford Handbook of Cognitive Linguistics*, pages 318–350, 2007.