UNIVERSITY OF CALIFORNIA

Los Angeles

Social Scene Understanding:
Group Activity Parsing, Human-Robot Interactions, and Perception of Animacy

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Statistics

by

Tianmin Shu

2019

ABSTRACT OF THE DISSERTATION

Social Scene Understanding:
Group Activity Parsing, Human-Robot Interactions, and Perception of Animacy

by

Tianmin Shu
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2019
Professor Song-Chun Zhu, Chair

This dissertation proposes new computational frameworks to address three core challenges for social scene understanding – group activity parsing, human-robot interactions, and perception of animacy. The goal of these frameworks is to represent the underlying structure of social scenes and to unify the perception and concept learning of both physics and social behaviors. For this, we first develop a joint parsing of group activities that yields a hierarchical representations of groups, events, and human roles, which provides a holistic view of a social scene. In a follow up work, the idea of joint parsing is also shown to be effective for boosting the performance of deep neural networks on group activity recognition. Second, we formulate social affordances as a hierarchical representation of human interactions, which can be learned from a handful of RGB-D videos of human interactions. Based on the symbolic plans derived from the learned knowledge, we further design a real-time motion inference to enable motion transfer from human interactions to human-robot interactions, which generalizes well in unseen social scenarios. Finally, we study human perception of animacy by designing new approaches to generate Heider-Simmel animations and by developing new computational models to account for human physical and social perception. Particularly, we propose a unified framework for modeling physics and social behaviors through i) a joint physical-social simulation engine, ii) a joint physical and social concept learning as the pursuit of generalized coordinates and their potential energy functions, and iii) a unified psychological space that integrates intuitive physics and intuitive psychology.

The dissertation of Tianmin Shu is approved.

Veronica Santos

Ying Nian Wu

Hongjing Lu

Song-Chun Zhu, Committee Chair

University of California, Los Angeles

2019

To my family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Prof. Song-Chun Zhu for introducing me to the field of social scene understanding, and for always encouraging me working towards the greatest challenges. As a young researcher, it is easy to get distracted by the most trendy topics in the field. The supervision given by Prof. Zhu strengthened my determination of solving important and hard problems in social scene understanding.

I have been fortunate enough to work with and to be advised by many great mentors: Prof. Hongjing Lu, Prof. Ying Nian Wu, Prof. Sinisa Todorovic, Prof. Michael S. Ryoo, and Prof. Tao Gao. The work presented in this dissertation has benefited tremendously from their insightful discussions. Many of them have also spent a great amount of time to help me edit papers, slides, and posters, which has been a valuable learning experience for me.

My sincere thanks to Dr. Caiming Xiong, Dr. Richard Socher, and Dr. Yuandong Tian for the two wonderful internships. Working in industry labs has offered me a very different view of research and engineering, which I was not able to get in an academic lab. It was also trilling to have the chance to mess with dozens of GPUs while conducting interesting and risky projects.

I have enjoyed much of my time at UCLA thanks to the ever growing VCLA family. Working, traveling, dinning, and joking with my lab mates have been a fun part of my life in the past five years. Specifically, I want to thank Dr. Dan Xie, Dr. Yibiao Zhao, Dr. Jianwen Xie, Prof. Quanshi Zhang, Prof. Tianfu Wu, and Prof. Xiaobao Liu for mentoring me on research and on career. I am also blessed to have had so many wonderful past and present colleagues in the lab – Dr. Yixin Zhu, Dr. Hang Qi, Dr. Xiaohan Nie, Dr. Brandon Rothrock, Prof. Jungseock Joo, Dr. Seyoung Park, Siyuan Qi, Yuanlu Xu, Yang Liu, Xiaofeng Gao, Lifeng Fan, Ruiqi Gao, Tao Yuan, Siyuan Huang, Mitch Hill, Mark Edmonds, Erik Nijkamp, Feng Gao, Xu Xie, Hangxin Liu, Tengyu Liu, Baoxiong Jia, Feng Shi, Shu Wang, Zhixiong Nan, Dr. Bo Li, and many other people.

A large part of this dissertation involves human experiments, which were made possible by the brilliant Yujia Peng, as well as by many participants and research assistants. Thank you all for your irreplaceable contributions to my dissertation work.

Finally, I dedicate this dissertation to my parents and grandparents, particularly my late grandmother, Zhiying, who devoted most of her time in her final years to raising me. Without the endless support of my family, I may never have the courage to pursue my academic dreams.

VITA

2014-2019	Graduate Research Assistant, Department of Statistics, UCLA.
2018	Research Intern, Facebook AI Research.
2017	Computational Modeling Prize (Perception/Action), Cognitive Science Society.
2017	Research Intern, Salesforce Research.
2010-2014	B.S. in Electronic Engineering, Fudan University.
2013	Research Intern, Center for Vision and Cognition, Learning and Autonomy, UCLA.

PUBLICATIONS

(* indicates equal contribution)

T. Shu, Y. Peng, H. Lu, S.-C. Zhu. Partitioning the Perception of Physical and Social Events Within a Unified Psychological Space. *41st Annual Meeting of the Cognitive Science Society (CogSci), 2019*.

X. Gao, R. Gong, **T. Shu**, X. Xie, S. Wang, S.-C. Zhu. VRKitchen: an Interactive 3D Environment for Learning Real Life Cooking Tasks. *ICML RL4RealLife Workshop, 2019*.

T. Shu, Y. Tian. M^3RL: Mind-aware Multi-agent Management Reinforcement Learning. *7th International Conference on Learning Representations (ICLR), 2019*.

T. Shu, C. Xiong, Y. N. Wu, S.-C. Zhu. Interactive Agent Modeling by Learning to Probe. *NeurIPS Deep Reinforcement Learning Workshop, 2018*.

P. Wei, Y. Liu, **T. Shu**, N. Zheng, S.-C. Zhu. Where and Why Are They Looking? Jointly Inferring

Human Attention and Intentions in Complex Tasks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.*

D. Xie, **T. Shu**, S. Todorovic, S.-C. Zhu. Learning and Inferring "Dark Matter" and Predicting Human Intents and Trajectories in Videos. IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI), 40(7): 1639-1652, 2018.

T. Shu*, Y. Peng*, L. Fan, H. Lu, S.-C. Zhu. Perception of Human Interaction Based on Motion Trajectories: from Aerial Videos to Decontextualized Animations. *Topics in Cognitive Science (TopiCS), 10(1): 225 - 241, 2018.*

T. Shu, C. Xiong, R. Socher. Hierarchical and Interpretable Skill Acquisition in Multi-task Reinforcement Learning. *6th International Conference on Learning Representations (ICLR), 2018.*

T. Shu*, Y. Peng*, L. Fan, H. Lu, S.-C. Zhu. Inferring Human Interaction from Motion Trajectories in Aerial Videos. *39th Annual Meeting of the Cognitive Science Society (CogSci), 2017.*

T. Shu, S. Todorovic, S.-C. Zhu. CERN: Confidence-Energy Recurrent Network for Group Activity Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.*

T. Shu, X. Gao, M. S. Ryoo, S.-C. Zhu. Learning Social Affordance Grammar from Videos: Transferring Human Interactions to Human-Robot Interactions. *IEEE International Conference on Robotics and Automation (ICRA), 2017.*

T. Shu*, S. Thurman*, D. Chen, S.-C. Zhu, H. Lu. Critical Features of Joint Actions that Signal Human Interaction. *38th Annual Meeting of the Cognitive Science Society (CogSci), 2016.*

T. Shu, M. S. Ryoo, S.-C. Zhu. Learning Social Affordance for Human-Robot Interaction. *25th Internation Joint Conference on Artificial Intelligence (IJCAI), 2016.*

T. Shu, D. Xie, B. Rothrock, S. Todorovic, S.-C. Zhu. Joint Inference of Groups, Events and Human Roles in Aerial Videos. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.*

CHAPTER 1

Introduction

In the past a few decades, physical scene understanding has been one of the most fundamental areas in AI research. Thanks to the recent progress, we now can build machines that achieve impressive (sometimes even super human) performance for detecting and recognizing objects [HZR16, HLV17, HGD17] and their relations [XZC17, JHM17, MGK19], for parsing, reconstructing, and reasoning 3D scenes [ZZ13, SGS13, HZC13, KLD14, LZZ17, ZLH17, ZCS18, HQZ18, HQX18], for learning intuitive physics [WYL15, LGF16, WLK17], and for many other aspects of understanding physical worlds.

However, understanding physical scenes by no means provides the full picture of modeling and reasoning the real human world, as humans not only see and reason about the physical objects in the world (i.e., physical perception), but also watch other humans' behaviors and try to understand their minds (i.e., social perception). In fact, as J. J. Gibson argued, "The richest and most elaborate affordances of the environment are provided by ... other people." [Gib79] In other words, we as humans need to understand the behaviors and mental states of other humans for our everyday activities. This also applies to machine agents (e.g., a service robot) if we want them to interact with humans, which calls for research on social scene understanding, an under-explored area of AI research focusing on building machines that are capable of understanding humans' behaviors and minds.

Consider the coffee shop scene captured in Figure 1.1. Various physical scene understanding models can offer us a large amount of details about the objects and their 3D layout in the room. By running state-of-the-art computer vision models [JKF16], we may even obtain reasonably accurate text descriptions of the physical properties or attributes of these objects as shown in the figure. There are also descriptions of simple actions taken by the humans, but they were treated in the

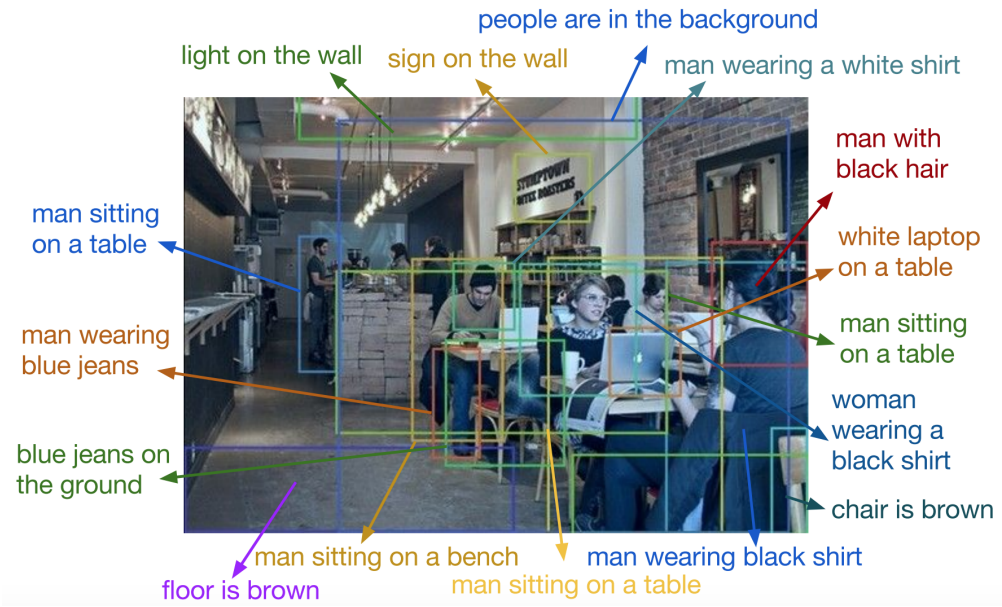


Figure 1.1: A coffee shop scene with dense captions generated by a computer vision model (from [JKF16]).

same manner as listing the attributes of objects. However, instead of simply accumulating these basic facts of this scene, humans usually future infer less obvious information hidden in the image – e.g., *two friends/co-workers are chatting while drinking coffee*; *a brista is serving an customer while another customer is waiting in line*. Such hidden messages constitute a social understanding of the observed scene. We can clearly see from this example that different from the extremely detailed picture drawn by the physical scene understanding, social scene understanding composes more abstract and high-level descriptions about what we can see.

In addition to the ability of constructing concise and structured interpretation of social scenes from rich visual information demonstrated in the above example, decades of studies on social perception reveals that humans can also extract rich social signals from very little input. The most well-known study is probably [HS44]. In 1940s, two psychologists, Heider and Simmel, created a short movie of three simple geometric shapes (a big triangle, a small triangle, and a circle) moving around a box, which is known as the Heider-Simmel animation (Figure 1.2). In their original experiments, they showed the animation to human subjects and asked them to describe it. Almost all of the subjects gave an anthropomorphic description – they were able to tell vivid and diverse

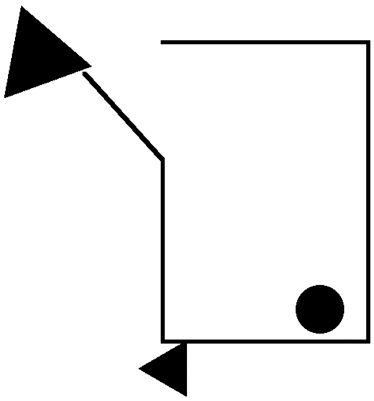


Figure 1.2: A snapshot of the original Heider-Simmel animation (from [HA04]).

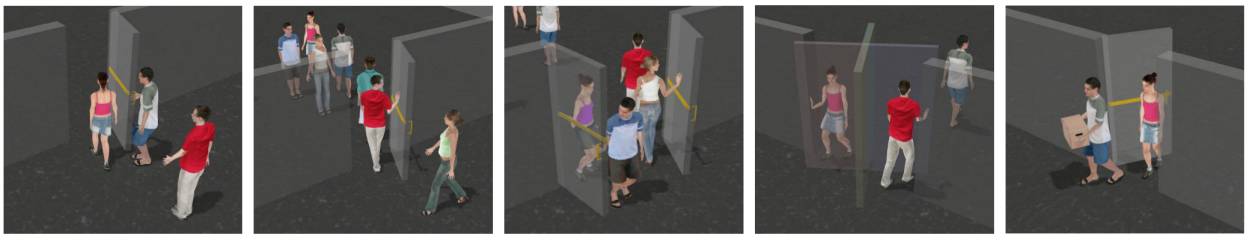


Figure 1.3: Planning body movement when opening a door by following necessary social etiquette (from [HT18]).

stories involving characters with distinct personalities, intents, and relations, even though all they saw was simple motion of some geometric shapes. In contrast, it is not difficult to imagine that state-of-the-art computer vision models that can beat humans on recognizing objects (e.g., ResNet [HZR16]) may only yield the most literal interpretation of this animation (i.e., shapes and their motion) without developing much meaningful social understanding.

Just like our understanding of the physical world, social scene understanding also helps shape our behaviors by providing crucial information such as the mental states of other humans which can not be acquired otherwise. For instance, as a basic social etiquette, we should hold the door if we predict that someone else is also trying to go through the same door. As shown in Figure 1.3, to follow this etiquette, we plan our body motion not only by assessing the physical scene (e.g., the position, size and type of the door), but also by inferring the intents, gender, age, physical strengths, and other nuanced information of other people. An agent without sufficient social understanding

of its surrounding environment will surely fail to behave in a socially appropriate way.

Inspired by humans' remarkable social perception and its critical influence on our behaviors in the real world, this dissertation aims to study how to build computational models that can construct interpretable and structured representations of social scenes and how to incorporate such representations into the decision making process of a robot when interacting with humans. Furthermore, we also take a deep look into how humans integrate the perception of physical and social scenes, and how we can develop a joint modeling of physical laws and social behaviors.

In summary, this dissertation studies three core problems in social scene understanding: group activity parsing, human-robot interactions, and perception of animacy. We summarize the contributions of this dissertation on these three problems as follows.

- **Group Activity Parsing**. Prior work on social scene understanding in videos has mostly focused on group activity recognition, which is essentially defined as video recognition – given a short video clip of human activities, a model predicts the activity label of this clip as a whole [CSS09, LWY12]. This is clearly far from the true understanding of social scenes. So what is social scene understanding? What are the fundamental elements in a social scene that a computational model should reason about? In Chapter 2, we attempt to offer an answer by proposing a new framework for parsing group activities in long videos. Specifically, we argue that a holistic understanding of social scenes should include a joint inference of three key elements: i) social groups, ii) events that people in each social group engage in, and iii) roles of the group members. For this, we propose spatiotemporal AND-OR graph (ST-AOG), a stochastic grammar model, as a hierarchical representation of the three key elements to model long-term spatiotemporal patterns in group activities. For evaluation, we compiled the first aerial event dataset that includes multiple social activities in an open space. The experimental results suggest that the joint inference of groups, events, and roles results in a more complete and longer-term understanding of observed social scenes and also improves the parsing accuracy of each individual element. Chapter 3 further incorporates the joint inference to structured deep neural networks (DNNs) through an energy-based model, which demonstrates a performance boost compared to the standard feed-forward predictions

from DNNs. Such advantage is particularly significant when we only have limited training examples and/or the input is noisy. In fact, this is the first DNN-based model that outperforms conventional approaches that rely on hand-crafted features on the Collective Activity dataset [CSS09].

- **Human-Robot Interactions**. With the recent progress in robotics, robots now have been able to perform many complex tasks for humans. As a result, it is inevitable that the robots will interact with humans in various social situations, such as service robots taking care of elderly people, robot co-workers collaborating with humans in a workplace, or simply a robot navigating through human crowds. Similar to human social interactions, human-robot interactions (HRI) must also follow certain social etiquette or social norms, in order to make humans comfortable. To this end, we propose the first computational framework to learn social affordances (i.e., suitable actions to take when interacting with humans in social activities) from a handful of human interaction videos as demonstrations (Chapter 4). We then represent the learned social affordances in the form of spatiaotemporal stochastic grammar, based on which we may generate robot plans to transfer human-human interactions into human-robot interactions (Chapter 5). To our knowledge, this is the first work on learning transferable knowledge (i.e., social affordances) from observing human interactions for enabling human-robot social interactions.

- **Perception of Animacy**. Heider and Simmel's pioneering work poses many unsolved questions about human's perception of animacy: How humans judge whether an entity in Heider-Simmel animations is an animated agent or a physical object? Do they use ad-hoc visual cues [DL94, ST00, TF00, TF06, GNS09, GMS10] or do they try to reason about the mental states (intents, desires, beliefs) of agents [BST09, UBM10]? What is the connection between physical perception (or intuitive physics) and social perception (or intuitive psychology)? Is there a principled way to jointly model both physical perception and social perception? There have been many research efforts devoted to these questions. However, so far we haven't had clear answers to them. This dissertation attempts to address questions from a computational perspective – designing computational models that account for

humans' perception of animacy (Chapter 6 and Chapter 7). In particular, we are interested in bridging physical perception and social perception in a unified framework. For this, we propose new approaches for i) automatically generating Heider-Simmel animations with rich and realistic behaviors by building a joint physical-social simulation engine, for ii) learning physical and social concepts by a unified paradigm, and for iii) constructing a joint representation of human perception of both physical and social events. Through multiple human experiments, we demonstrate that our unified framework indeed can discover a distribution of human perception of physical and social events in a unified psychological space, which sheds light on how to integrate intuitive physics and intuitive psychology.

CHAPTER 2

Joint Inference of Groups, Events and Human Roles in Aerial Videos

2.1 Introduction

2.1.1 Motivation and Objective

Video surveillance of large spatial areas using unmanned aerial vehicles (UAVs) becomes increasingly important in a wide range of civil, military and homeland security applications. For example, identifying suspicious human activities in aerial videos has the potential of saving human lives and preventing catastrophic events. Yet, there is scant prior work on aerial video analysis [KGS13, IRF13, PG14], which for the most part is focused on tracking people and vehicles (with few exceptions [OMS10]) in relatively sanitized settings.

Towards advancing aerial video understanding, we present a new problem of parsing extremely low-resolution aerial videos of large spatial areas, such as picnic areas rich with co-occurring group events, viewed top-down under camera motion, as illustrated in Figure 2.1 and 2.2. Given an aerial video, our objectives include:

1. Grouping people based on their events;
2. Recognizing events present in each group;
3. Recognizing roles of people involved in these events.

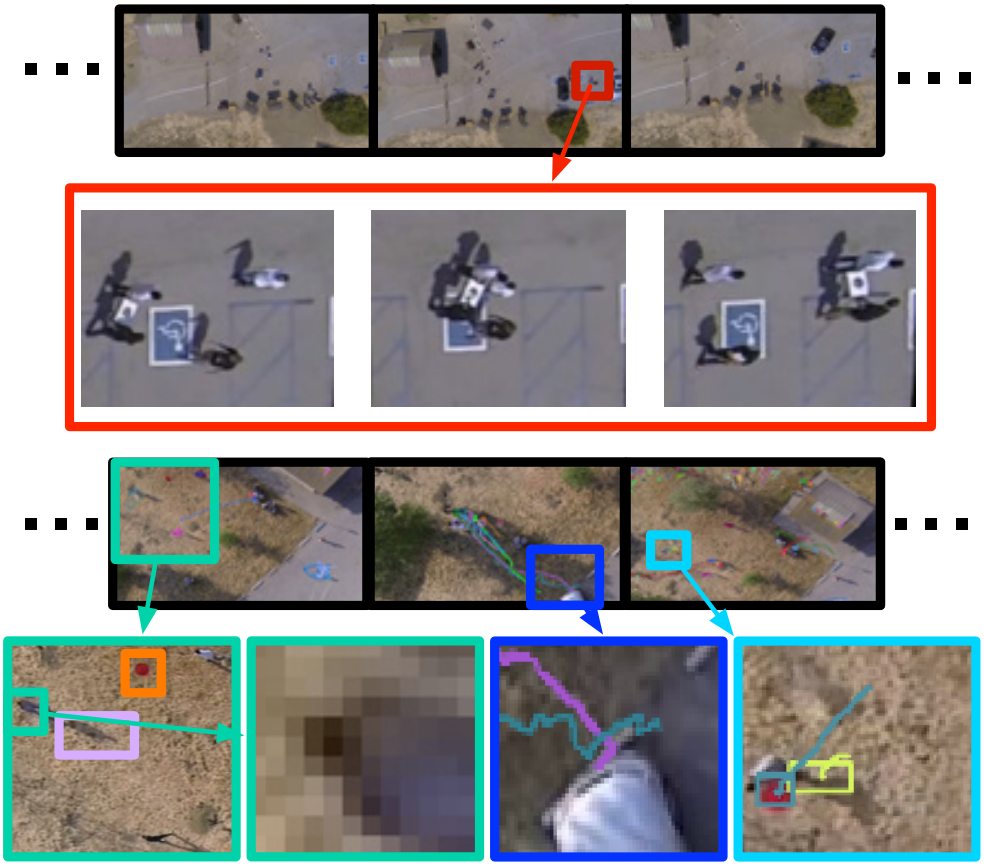


Figure 2.1: Our low-resolution aerial videos show top-down views of people engaged in a number of concurrent events, under camera motion. Different types of challenges are color-coded. The red box marks a zoomed-in video part with varying dynamics among people and their roles *Deliverer* and *Receiver* in *Exchange Box*. The green marks extremely low resolution and shadows. The blue indicates only partially visible *Car*. The cyan marks noisy tracking of person and the small object *Frisbee*.

2.1.2 Scope and Challenges

As illustrated in Figure 2.1, we focus on videos of relatively wide spatial areas (e.g., parks with parking lots) with interesting terrains, taken on-board of a UAV flying at a large altitude (25m) from the ground. People in such videos are formed into groups engaged in different events, involving complex n-ary interactions among themselves (e.g., a *Guide* leading *Tourist*s in *Group Tour*), as well as interactions with objects (e.g., *Play Frisbee*). Also, people play particular roles in each event (e.g., *Deliverer* and *Receiver* roles in *Exchange Box*).

1. **Low resolution.** People and their portable objects are viewed at an extremely low resolution. Typically, the size of a person is only 15×15 pixels in a frame, and small objects critical for distinguishing one event from another may not be even distinguishable by a human eye.

2. **Camera motion** makes important cues for event recognition (e.g., object like *Car*) only partially visible or even out of view, and thus may require seeing longer video footage for their reliable detection.

3. **Shadows in top view** make background subtraction very challenging.

Unfortunately, popular appearance-based approaches to detecting people and objects used to produce input for recognizing group events and interactions [PES09, CS14, RA11, LSM12, RYF13, FHR12] do not handle the above three challenges. Thus we have to depart from the appearance-based event recognition.

In addition, in the face of these challenges, the state of the art methods in people and vehicle tracking frequently miss to track moving foreground, and typically produce short, broken tracklets with a high rate of switched track IDs.

4. **Space-time dynamics.** Our events are characterized by both very large and very small space-time dynamics within a group of people. For example, in the event of a line forming in front of a vending machine, called *Queue for Vending machine*, the participants may be initially scattered across a large spatial area, and may form the line very slowly, while partially occluding one another when closely standing in the line.

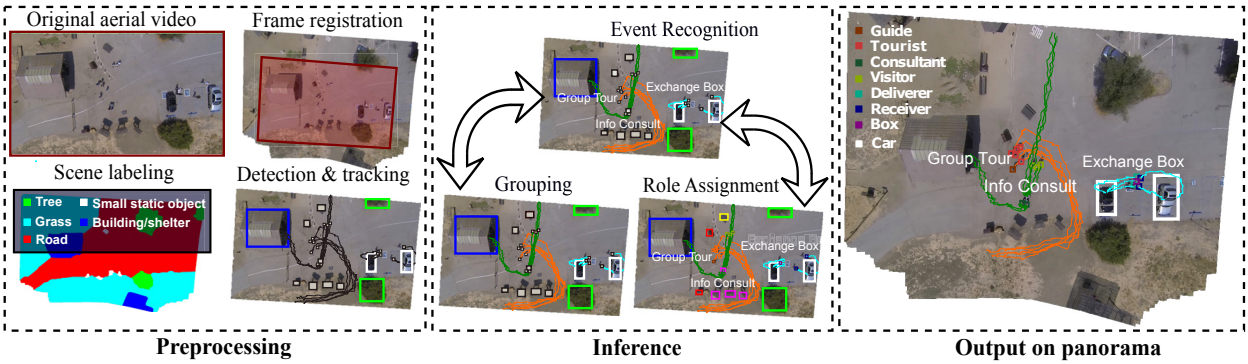


Figure 2.2: The main steps of our approach. Our recognition accounts for the temporal layout of latent sub-events, people's roles within events (e.g., *Guide*, *Visitor*), and small objects that people interact with (e.g., *Box*, *trash bin*). We iteratively optimize groupings of the foreground trajectories, infer their events and human roles (color-coded tracks) within events.

2.1.3 Overview of Our Approach

As Figure 2.2 illustrates, our approach consists of two main steps:

1. **Preprocessing.** We ground our approach onto noisy detections and tracking. Foreground tracking under camera motion is made feasible by registering video frames onto a reference plane. By frame registration, we generate a panorama for scene labeling. Due to the challenges mentioned in Section 2.1.2, tracking of small portable objects and people produces highly unreliable frequently broken tracklets, with a high miss rate. We improve the initial tracking results by agglomeratively clustering tracklets into longer trajectories based on their spatial layout and velocity. We detect large objects (e.g.. buildings, cars) using the approach of [RPZ13], and classify superpixels [ASS12] of the panorama for scene labeling.

2. **Inference.** We seek event occurrences in the space-time patterns of the foreground trajectories and their relations with the detections of objects in the scene. To constrain our recognition hypotheses under uncertainty, we resort to domain knowledge represented by a probabilistic grammar – namely, a spatiotemporal AND-OR graph (ST-AOG). ST-AOG encodes decompositions of events into temporal sequences of sub-events. Sub-events are defined by our new formalism called *latent spatiotemporal templates* of n-ary relations among people and objects. The templates jointly

encode varying spatiotemporal relations of characteristic roles of all people, as well as their interactions with objects, while engaged in the event.

We specify an iterative algorithm based on Markov Chain Monte Carlo (MCMC [KL12]) along with dynamic programming (DP) to jointly infer groups, events and human roles.

2.1.4 Prior Work and Our Contributions

Our work is related to three research streams.

Event Recognition in Aerial Videos. Prior work on aerial image and video understanding typically puts restrictions on their settings for limited tasks. For example, [PA12] requires robust motion segmentation and learning of object shapes for tracking objects; [IRF13] recognizes people based on background subtraction and motion; and [PG14] depends on appearance-based regressor and background subtraction for tracking vehicles. Regarding the objectives, these approaches mainly focus on detecting and tracking people or vehicles [XCS10, OMS10, KGS13]. We advance prior work by relaxing their assumptions about the setting, and by extending their objectives to jointly infer groups, events, human roles.

Group Activity Recognition. Simultaneous tracking of multiple people, discovering groups of people, and recognizing their collective activities have been addressed only in every-day videos, rather than aerial videos [CSS09, RA11, LWY12, GCR12, LPZ13, CS14, CCP14, AO14, SAL14, TML14]. Also, work on recognizing group activities in large spatial scenes requires high-resolution videos for a "digital zoom-in" [AXZ12]. As input, these approaches use person detections along with cues about human appearance, pose, and orientation — i.e., information that cannot be reliably extracted from our aerial videos. There are also some trajectory-based methods for event recognition [NZH03, SHJ14, LXG12], but they focus on simpler events compared to what we discuss in this chapter. Regarding the representation of collective activities, prior work has used a descriptor of human locations and orientations, similar to shape-context [CS14, AO14]. We advance prior work with our new formalism of latent spatiotemporal template of human roles and their interactions with other actors and objects.

Recognition of Human Roles. Existing work on recognizing social roles and social in-

teractions of people typically requires perfect tracking results [RYF13], reliable estimation of face direction and attention in 3D space [FHR12], detection of agent's feet location in the scene [ZHY11], and thus are not applicable to our domain. Our approach is related to recent approaches aimed at jointly recognizing events and social roles by identifying interactions of sub-groups [GCR12, LPZ13, LSM12, KHH13].

Contributions:

1. Addressing a more challenging setting of aerial videos;
2. New formalism of latent spatiotemporal templates of n-ary relations among human roles and objects;
3. Efficient inference using dynamic programming aimed at grouping, recognition and localizing temporal extents of events and human roles
4. New dataset of aerial videos with per-frame annotations of people's trajectories, object labels, roles, events and groups.

2.2 Representation

2.2.1 Representing of Group Events by ST-AOG

Similar with hierarchical representation in [GSS09, LGL09, PSY13, PR14], domain knowledge is formalized as ST-AOG, depicted in Figure 2.3. Its nodes represent the following four sets of concepts: events $\Delta_{\mathrm{E}} = \{E_i\}$; sub-events $\Delta_L = \{L_a\}$; human roles $\Delta_{\mathrm{R}} = \{R_j\}$; small objects that people interact with $\Delta_{\mathrm{O}} = \{O_j\}$; and large objects and scene surfaces $\Delta_{\mathrm{S}} = \{S_j\}$. A particular pattern of foreground trajectories observed in a given time interval gives rise to a sub-event, and a particular sequence of sub-events defines an event.

Edges of the ST-AOG represent decomposition and temporal relations in the domain. In particular, the nodes are hierarchically connected by decomposition edges into three levels, where the root level corresponds to events, middle level encodes sub-events, and leaf level is grounded onto foreground tracklets and object detections in the video. The nodes of sub-events are also laterally

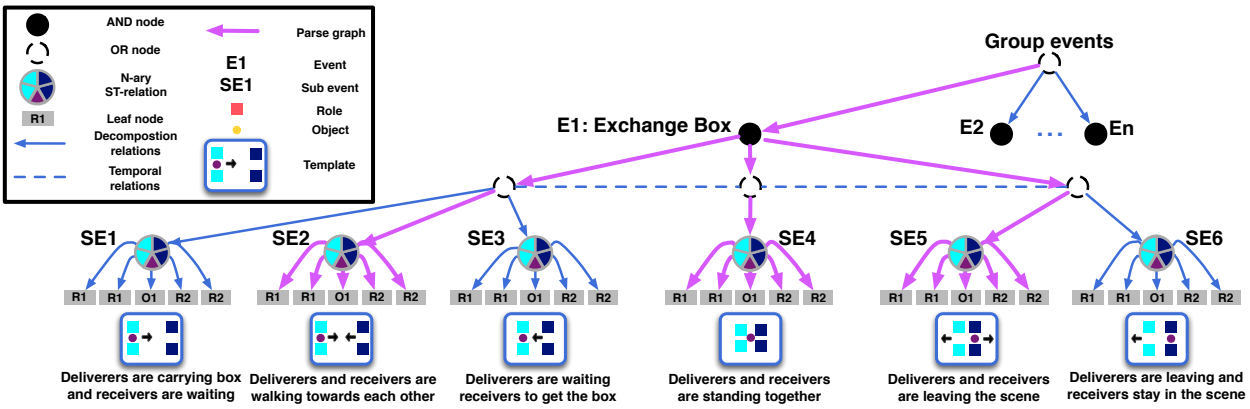


Figure 2.3: A part of ST-AOG for *Exchange Box*. The nodes are hierarchically connected (solid blue) into three levels, where the root level corresponds to events, middle level encodes sub-events, and leaf level is grounded onto foreground tracklets and small static objects in the video. The lateral connections (dashed blue) indicate temporal relations of sub-events. The colored pie-chart nodes represent templates of n-ary spatiotemporal relations among human roles and objects (see Figure 2.4). The magenta edges indicate an inferred parse graph which recognizes and localizes temporal extents of events, sub-events, human roles and objects in the video.

connected for capturing "followed-by" temporal relations of sub-events within the corresponding events.

ST-AOG has special types of nodes. An AND node, $\wedge$, encodes a temporal sequence of latent sub-events required to occur in the video so as to enable the event occurrence (e.g., in order to *Exchange Box*, the *Deliverer*s first need to approach the *Receiver*s, give the *Box* to the *Receiver*s, and then leave). For a given event, an OR node, $\vee$, serves to encode alternative space-time patterns of distinct sub-events.

2.2.2 Sub-events as Latent Spatiotemporal Templates

A temporal segment of foreground trajectories corresponds to a sub-event. ST-AOG represents a sub-event as the *latent* spatiotemporal template of n-ary spatiotemporal relations among foreground trajectories within a time interval, as illustrated in Figure 2.4. In particular, as an event is unfolding in the video, foreground trajectories form characteristic space-time patterns, which may not be semantically meaningful. As they frequently occur in the data, they can be robustly extracted from training videos through unsupervised clustering. Our spatiotemporal templates formalize these patterns within the Bayesian framework using unary, pairwise, and n-ary relations among the foreground trajectories. In addition, our unsupervised learning of spatiotemporal templates address unstructured events in a unified manner. Namely, more structured events need more templates and an unstructured one is represented by a single template.

Unary attributes. A foreground trajectory, $\Gamma = [\Gamma^1, ..., \Gamma^k, ...]$, can be viewed as spanning a number of time intervals, $\tau_k = [t_{k-1}, t_k]$, where $\Gamma^k = \Gamma(\tau_k)$. Each trajectory segment, Γ^k, is associated with unary attributes, $\boldsymbol{\phi} = [\boldsymbol{r}^k, s^k, \boldsymbol{c}^k]$. Elements of the role indicator vector $\boldsymbol{r}^k(l) = 1$ if Γ^k belongs to a person with role $l \in \Delta_{\mathrm{R}}$ or object class $l \in \Delta_{\mathrm{O}}$; otherwise $\boldsymbol{r}^k(l) = 0$. The speed indicator $s^k = 1$ when the normalized speed of Γ^k is greater than a threshold (we use 2 pixels/sec); otherwise, $s^k = 0$. Elements of the closeness indicator vector $\boldsymbol{c}^k(l) = 1$ when Γ^k is close to any of the large objects or types of surfaces detected in the scene indexed by $l \in \Delta_{\mathrm{S}}$, such as *Building*, *Car*, for a threshold (70 pixels); o.w., $\boldsymbol{c}^k(l) = 0$.

Pairwise relations. of a pair of trajectory segments, Γ^k_j and $\Gamma^k_{j'}$, are aimed at capturing spa-

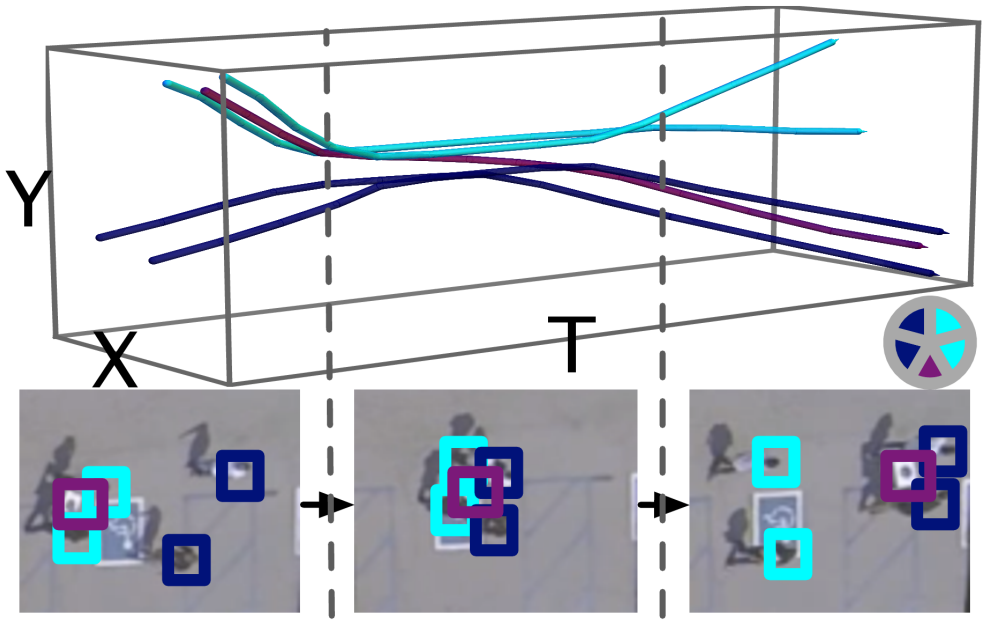


Figure 2.4: Three example templates of n-ary spatiotemporal relations among foreground trajectories extracted from the video (XYT-space) for the event *Exchange Box*. The recognized roles *Deliverers*, *Receivers* and the object *Box* in each template are marked cyan, blue and purple, respectively. Spatiotemporal templates are depicted as colored pie-chart nodes in Figure 2.3.

tiotemporal relations of human roles or objects represented by the two trajectories, as illustrated in Figure 2.4. The pairwise relations are specified as: $\boldsymbol{\phi}_{jj'} = [d^k_{jj'}, \theta^k_{jj'}, \boldsymbol{r}^k_{jj'}, s^k_{jj'}, \boldsymbol{c}^k_{jj'}]$, where $d^k_{jj'}$ is the mean distance between Γ^k_j and $\Gamma^k_{j'}$; $\theta^k_{jj'}$ is the angle subtended between Γ^k_j and $\Gamma^k_{j'}$; and the remaining three pairwise relations check for compatibility between the aforementioned binary relations as: $\boldsymbol{r}^k_{jj'} = \boldsymbol{r}^k_j \oplus \boldsymbol{r}^k_{j'}$, $s^k_{jj'} = s^k_j \oplus s^k_{j'}$, $\boldsymbol{c}^k_{jj'} = \boldsymbol{c}^k_j \oplus \boldsymbol{c}^k_{j'}$, where $\oplus$ denotes the Kronecker product.

n-ary relations. Towards encoding unique spatiotemporal patterns of a set of trajectories, we specify the following n-ary attribute. A set of trajectory segments, $G_i(\tau_k) = G^k_i = \{\Gamma^k_j\}$, can be described by a 18-bin histogram $\boldsymbol{h}^k$ of their velocity vectors. $\boldsymbol{h}^k$ counts orientations of velocities at every point along the trajectories in a polar coordinate system: 6 bins span the orientations in $[0, 2\pi]$, and 3 bins encode the locations of trajectory points relative to a given center. As the polar-coordinate origin, we use the center location of a given event in the scene.

Unsupervised Extraction of Templates. Given training videos with ground-truth partition of all their ground-truth foreground trajectories G into disjoint subsets $G = \{G_i\}$. Every G_i can be further partitioned into equal-length time intervals $G_i = \{G^k_i\}$ ($|\tau^k| = 2$sec). We use K-means clustering to group all $\{\Gamma^k_{i,j}\}$, and then estimate spatiotemporal templates $\{L_a\}$ as representatives of the resulting clusters a. For K-means clustering, we use ground-truth values of the aforementioned unary and pairwise relations of $\{\Gamma^k_{i,j}\}$. In our setting of 11 categories of events occurring in aerial videos, we estimate $|\Delta_L| = 27$ templates.

2.3 Formulation and Learning of Templates

Given the spatiotemporal templates, $\Delta_L = \{L_a\}$, extracted by K-means clustering from training videos (see Section 2.2.2), we will conduct inference by seeking these latent templates in foreground trajectories of the new video. To this end, we define the log-likelihood of a set of foreground trajectories $G = \{\Gamma_j\}$ given $L_a \in \Delta_L$ as

$$\begin{aligned} \log p(G|L_a) &\propto \sum_j \boldsymbol{w}^1_a \cdot \boldsymbol{\phi}_j + \sum_{jj'} \boldsymbol{w}^2_a \cdot \boldsymbol{\phi}_{jj'} + \boldsymbol{w}^3_a \cdot \boldsymbol{h}, \\ &= \boldsymbol{w}_a \cdot [\sum_j \phi_j, \sum_{jj'} \boldsymbol{\phi}_{jj'}, \boldsymbol{h}] = \boldsymbol{w}_a \cdot \boldsymbol{\psi}. \end{aligned} \tag{2.1}$$

where the bottom equation of (2.1) formalizes every template as a set of parameters $\boldsymbol{w}_a = [\boldsymbol{w}_a^1, \boldsymbol{w}_a^2, \boldsymbol{w}_a^3]$ appropriately weighting the unary, pairwise and n-ary relations of G, $\boldsymbol{\psi}$. Recall that our spatiotemporal templates are extracted from unit-time segments of foreground trajectories in training. Thus, the log-likelihood in Eq. (2.1) is defined only for sets G consisting of unit-time trajectory segments.

From Eq. (2.1), the parameters $\boldsymbol{w}_a$ can be learned by maximizing the log-likelihood of $\{\boldsymbol{\psi}_a^k\}$ extracted from the corresponding clusters a of training trajectories.

The log-posterior of assigning template L_a to longer temporal segments of trajectories, falling in $\tau = (t', t)$, $t' < t$, is specified as

$$\log p(L_a(\tau)|G(\tau)) \propto \sum_{k=t'}^{t} \log p(G^k|L_a) + \log p(L_a(\tau)) \tag{2.2}$$

where $p(L_a(\tau))$ is a log-normal prior that L_a can be assigned to a time interval of length $|\tau|$. The hyper-parameters of $p(L_a(\tau))$ are estimated using the MLE on training data.

2.4 Probabilistic Model

A parse graph is an instance of ST-AOG, explaining the event, sequence of sub-events, and human role and object label assignment. The solution of our video parsing is a set of parse graphs, $W = \{pg_i\}$, where every pg_i explains a subset of foreground trajectories, $G_i \subset G$, as

$$pg_i = \{e_i, \tau_i = [t_{i,0}, t_{i,T}], \{L(\tau_{i,u})\}, \{\boldsymbol{r}_{i,j}\}\}, \tag{2.3}$$

where $e_i \in \Delta_{\mathrm{E}}$ is the recognized event conducted by G_i; $\tau_i = [t_{i,0}, t_{i,T}]$ is the temporal extent of e_i in the video starting from frame $t_{i,0}$ and ending at frame $t_{i,T}$; $\{L(\tau_{i,u})\}$ are the templates (i.e., latent sub-events) assigned to non-overlapping, consecutive time intervals $\tau_{i,u} \subset \tau_i$, such that $|\tau_i| = \sum_u |\tau_{i,u}|$; and $\boldsymbol{r}_{i,j}$ is the human role or object class assignment to jth trajectory $\Gamma_{i,j}$ of G_i.

Our objective is to infer W that maximizes the log-posterior $\log p(W|G) \propto -\mathcal{E}(W|G)$, given all foreground trajectories G extracted from the video. The corresponding energy $\mathcal{E}(W|G)$ is specified for a given partitioning of G into N disjoint subsets G_i as

$$\mathcal{E}(W|G) \propto \sum_{i=1}^{N} \Big[\underbrace{-\log p(\wedge_{e_i}|\vee_{\text{root}})}_{\text{select event } e_i} + \sum_u \big[\underbrace{-\log p(\wedge_{L_a}|\vee_{e_i})}_{\text{select template } L_a} \underbrace{-\log p(L_a(\tau_{i,u})|G_i(\tau_{i,u}))}_{\text{assign template}} \big] \Big] \tag{2.4}$$

where $G_i(\tau_{i,u})$ denotes temporal segments of foreground trajectories falling in time intervals $\tau_{i,u}$, $|\tau_i| = \sum_u |\tau_{i,u}|$, and $\log p(L(\tau_{i,u})|G_i(\tau_{i,u}))$ is given by Eq. (2.2). Also, $\log p(\wedge_{e_i}|\vee_{\text{root}})$ and $\log p(\wedge_{L_a}|\vee_{e_i})$ are the log-probabilities of the corresponding switching OR nodes in ST-AOG for selecting particular events $e_i \in \Delta_E$ and spatiotemporal templates $L_a \in \Delta_L$. These two switching probabilities are simply estimated as the frequency of corresponding selections observed in training data.

2.5 Inference

Given an aerial video, we first build a video panorama and extract foreground trajectories G. Then, the goal of inference is to: (1) partition G into disjoint groups of trajectories $\{G_i\}$ and assign label event $e_i \in \Delta_{\text{E}}$ to every G_i; (2) assign human roles and object labels $\boldsymbol{r}_{i,j}$ to trajectories $\Gamma_{i,j}$ within each group G_i; and 3) assign latent spatiotemporal templates $L(\tau_{i,u}) \in \Delta_L$ to temporal segments $\tau_{i,u}$ of foreground trajectories within every G_i. For steps (1) and (2) we use two distinct MCMC processes. Given groups G_i, event labels e_i and role assignment $r_{i,j}$ proposed in (1) and (2), step (3) uses dynamic programming for efficient estimation of sub-events $L(\tau)$ and their temporal extents τ. Steps (1)–(3) are iterated until convergence, i.e., when $\mathcal{E}(W|G)$, given by Eq. (2.4), stops decreasing after a sufficiently large number of iterations.

2.5.1 Grouping

Given G, we first use [GCR12] to perform initial clustering of foreground trajectories into atomic groups. Then, we apply the first MCMC to iteratively propose either to merge two smaller groups into a merger, with probability $p(1) = 0.7$, or to split a merger into two smaller groups, with probability $p(2) = 0.3$. Given the proposal, each resulting group G_i is labeled with an event $e_i \in \Delta_{\text{E}}$ (we enumerate all possible labels). In each proposal, the MCMC jumps from current solution W to a new solution W' generated by one of the dynamics. The acceptance rate is $\alpha = \min\left\{1, \frac{Q(W \to W')p(W'|G)}{Q(W' \to W)p(W|G)}\right\}$, where the proposal distribution $Q(W \to W')$ is one of $p(1)$ or $p(2)$ depending on the proposal, and $p\left(W|G\right)$ is given by Eq. (2.4).

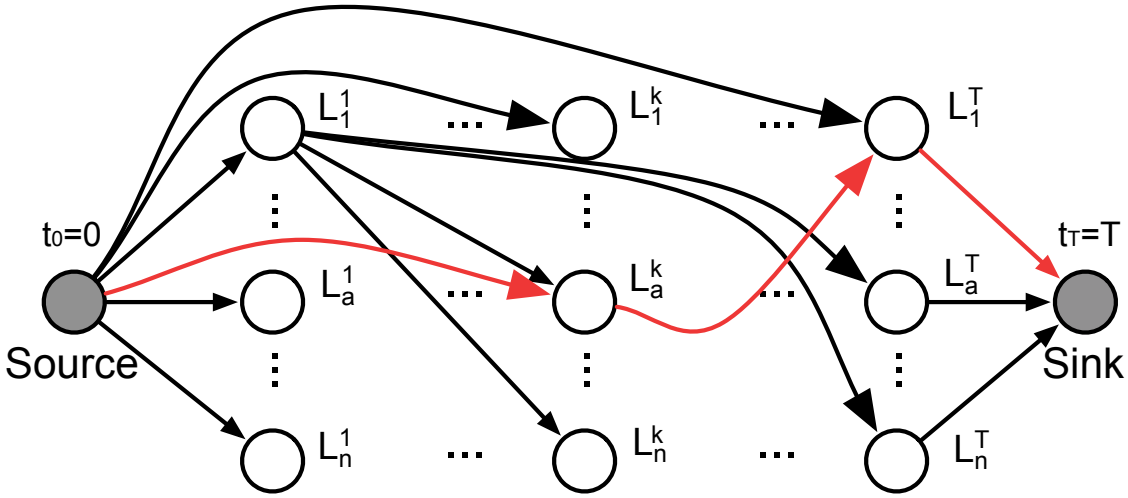


Figure 2.5: Our DP process can be illustrated by this DAG (directed acyclic graph). An edge between $L_{a'}^{k'}$ and L_a^k means the transition $L_{a'} \rightarrow L_a$ follows the rule defined in ST-AOG and the time interval $[t_{a'}, t_a]$ is assigned with template L_a. In this sense, with the transition rules and the prior defined in Eq. (2.2) (we do not consider the assignment with low prior probability), we can define the edges of such DAG. So the goal of DP is equivalent to finding a shortest path between source and sink. The red edges highlight a possible path. Suppose we find a path $source \rightarrow L_3^8 \rightarrow L_1^{20} \rightarrow sink$. This means that we decompose $[0, T]$ into 2 time intervals: $[0, 8\delta t]$, $[8\delta t, T]$, and they are assigned with template L_3 and L_1 respectively.

2.5.2 Human Role Assignment

Given a partitioning of G into groups $\{G_i\}$ and their event labels $\{e_i\}$, we use the second MCMC process within every G_i to assign human roles and object labels to trajectories. Each trajectory $\Gamma_{i,j}$ in G_i is randomly assigned with an initial human-role/object label $\boldsymbol{r}_{i,j}$ for solution pg_i. In each iteration, we randomly select $\Gamma_{i,j}$ and change it's role label to generate a new proposal pg_i'. The acceptance rate is $\alpha = \min\left\{1, \frac{Q(pg_i \rightarrow pg_i')p(pg_i'|G_i)}{Q(pg_i' \rightarrow pg_i)p(pg_i|G_i)}\right\}$, where $\frac{Q(pg_i \rightarrow pg_i')}{Q(pg_i' \rightarrow pg_i)} = 1$ and $p\left(pg_i'|G_i\right)$ is maximized by dynamic programming specified in the next section 2.5.3.

2.5.3 Detection of Latent Sub-events with DP

From steps (1) and (2), we have obtained the trajectory groups $\{G_i\}$, and their event $\{e_i\}$ and role labels $\{\boldsymbol{r}_{i,j}\}$. Every G_i can be viewed as occupying time interval of $\tau_i = [t_{i,0}, t_{i,T}]$. The results of steps (1) and (2) are jointly used with detections of large objects $\{S_i\}$ to estimate all unary, pairwise, and n-ary relations $\boldsymbol{\psi}_i$ of every G_i. Then, we apply dynamic programming for every G_i in order to find latent templates $L(\tau_{i,u}) \in \Delta_L$ and their optimal durations $\tau_{i,u} \subset [t_{i,0}, t_{i,T}]$. In the sequel, we drop notion i for the group, for simplicity.

The optimal assignment of sub-events can be formulated using a graph, shown in Figure 2.5. To this end, we partition $[t_0, t_T]$ into equal-length time intervals $\{[t_{k-1}, t_k]\}$, where $t_k - t_{k-1} = \delta t$, $\delta t = 2$sec. Nodes L_a^k in the graph represent the assignment of templates $L_a \in \Delta_L$ to the intervals $[t_{k-1}, t_k]$. The graph also has the source and sink nodes.

Directed edges in the graph are established only between nodes $L_a^{k'}$ and L_a^k, $1 \leq k' < k$, to denote a possible assignment of the very same template L_a to the temporal sequence $[t_{k'}, t_k]$. The directed edges are assigned weights (a.k.a. belief messages), $m(L_a^{k'}, L_a^k)$, defined as

$$m(L_a^{k'}, L_a^k) = \log p(L_a(t_{k'}, t_k) | G_i(t_{k'}, t_k)), \tag{2.5}$$

where $\log p(L_a(t_{k'}, t_k)|G_i(t_{k'}, t_k))$ is given by Eq. (2.2). Consequently, the belief of node L_a^k is defined as

$$b(L_a^k) = \max_{k', a'} \; b(L_{a'}^{k'}) + m(L_a^{k'}, L_a^k). \quad \text{[Forward pass]} \tag{2.6}$$

Here $b(L_a^0) = 0$. We compute the optimal assignment of latent sub-events using the above graph in two passes. In the *forward pass*, we compute the beliefs of all nodes in the graph using Eq. (2.6). Then, in the *backward pass*, we backtrace the optimal path between the sink and source nodes, in the following steps:

0: Let $t_k \leftarrow t_T$;

1: Find the optimal sub-event assignment at time t_k as $L_{a^*}^k = \arg\max_a \; b(L_a^k)$; let $a \leftarrow a^*$;

2: Find the best time moment in the past t_{k^*}, $k^* < k$, and its best sub-event assignment as $L_{a^*}^{k^*} = \max_{a', k'} b(L_{a'}^{k'}) + m(L_a^{k'}, L_a^k)$; Let $a \leftarrow a^*$ and $k \leftarrow k^*$.

3: If $t_k > t_0$, go to Step 2.

2.6 Experiment

Existing Datasets. Existing datasets on aerial videos, group events or human roles are inappropriate for our evaluation. These aerial videos or images indeed show some group events, but the events are not annotated ([ARS07, OMS10, Oh11]). Most aerial datasets are compiled for tracking evaluation only [KGS13, IRF13, PG14]. Existing group-activity videos [CSS09, RA11, AXZ12, LPZ13] or social role videos [ZHY11, FHR12, LSM12, RYF13, KHH13] are captured on or near the ground surface, and have sufficiently high resolution for robust people detection. Thus, we have prepared and released a new aerial video dataset [1] with the new challenges listed in Section 2.1.2.

Aerial Events Dataset. A hex-rotor with a GoPro camera was used to shoot aerial videos at altitude of 25 meters from the ground. The videos show two different scenes, viewed top-down from the flying hex-rotor. The dataset contains 27 videos, 86 minutes, 60 fps, resolution of 1920×1080, with about 15 actors in each video. All video frames are registered onto a reference plane of the video panorama. Annotations are provided ([VPR13]) as: bounding boxes around groupings of people, events, human roles, and small and large objects. The objects include: 1. *Building*, 2. *Vending Machine*, 3. *Table & Seat*, 4. *BBQ Oven*, 5. *Trash Bin*, 6. *Shelter*, 7. *Info Booth*, 8. *Box*, 9. *Frisbee*, 10. *Car*, 11. *Desk*, 12. *Blanket*. The events include: 1. *Play Frisbee*, 2. *Serve Table*, 3. *Sell BBQ*, 4. *Info Consult*, 5. *Exchange Box*, 6. *Pick Up*, 7. *Queue for Vending Machine*, 8. *Group Tour*, 9. *Throw Trash*, 10. *Sit on Table*, 11. *Picnic*. The human roles include: 1. *Player*, 2. *Waiter*, 3. *Customer*, 4. *Chef*, 5. *Buyer*, 6. *Consultant*, 7. *Visitor*, 8. *Deliverer*, 9. *Receiver*, 10. *Driver*, 11. *Queuing Person*, 13. *Guide*, 14. *Tourist*, 15. *Trash Thrower*, 16. *Picnic Person*.

Evaluation Metrics. We split the 27 videos into 3 sets, such that different event categories are evenly distributed, and use a three-fold cross validation for our evaluation. Although our training and test videos show the same two scenes, we make the assumption that the layout of ground surfaces and large objects is unknown. Also, different videos in our dataset cover different parts of these large scenes, which are also assumed unknown. We evaluate accuracy of: i) grouping people, ii) event recognition, iii) role assignment. While our approach also estimates sub-events, note that

[1] Dataset can be downloaded from https://tshu.io/AerialVideo/AerialVideo.html.

they are latent and not annotated. The results are all time-averaged with the lengths of trajectories in each video. For specifying evaluation metrics we use the following notation. $G = \{G_i\}$ and $G' = \{G'_i\}$ are the sets of groups in ground-truth and inference results respectively. Γ_{ij} is the jth trajectory in ith group in ground-truth data, with duration of $|\tau_{ij}|$, group label g_{ij}, event type e_{ij} and human role r_{ij} in ground-truth. So is Γ'_{ij} in our inference. For group G_i, we call the best matched (i.e. overlapped) group in G' as M_i. For group G'_i, we call the best match group in G as M'_i. Then, precision and recall of grouping are

$$Pr_g = \sum_{G_i \in G} \Big(\sum_{\Gamma_{ij} \in G_i} \mathbb{1}\left(M_i = g'_{ij}\right) \cdot |\tau_{ij}| / \sum_{\Gamma_{ij} \in G_i} |\tau_{ij}| \Big) \tag{2.7}$$

$$Rc_g = \sum_{G'_i \in G'} \Big(\sum_{\Gamma'_{ij} \in G'_i} \mathbb{1}\left(M'_i = g_{ij}\right) \cdot |\tau'_{ij}| / \sum_{\Gamma'_{ij} \in G'_i} |\tau'_{ij}| \Big) \tag{2.8}$$

Accuracy of grouping is $F_g = 2 \Big/ (1/Pr_g + 1/Rc_g)$.

Event recognition accuracy E_e and role assignment accuracy E_r are defined as

$$E_e = \sum_{G'_i \in G'} \Big(\sum_{\Gamma'_{ij} \in G'_i} \mathbb{1}\left(e_{ij} = e'_{ij}\right) \cdot |\tau_{ij}| \Big) / \sum_{G'_i \in G'} \sum_{\Gamma'_{ij} \in G'_i} |\tau_{ij}| \tag{2.9}$$

$$E_r = \sum_{G'_i \in G'} \Big(\sum_{\Gamma'_{ij} \in G'_i} \mathbb{1}\left(r_{ij} = r'_{ij}\right) \cdot |\tau_{ij}| \Big) / \sum_{G'_i \in G'} \sum_{\Gamma'_{ij} \in G'_i} |\tau_{ij}|. \tag{2.10}$$

Baselines. To evaluate effectiveness of each module of our approach, we compare with baselines and variants of our method defined in Table 2.1. For the baselines we extract the following low-level features on trajectories: shape-context like feature [CSS09], average velocity, aligned orientation, distance from each type of large objects. All elements of feature vectors are normalized to fall in [0, 1].

Results. We register raw videos by RANSAC over Harris Corner feature points, then apply method of [IRF13] for tracking, which is based on background subtraction [YO07, Sob13]. We also use the detector of [RPZ13] to detect buildings and cars, while other static objects are inferred in scene labeling. We do not detect portable objects, e.g., *Frisbee* and *Box*.

We evaluate our approach on both annotated bounding boxes and real tracking results. Example qualitative results are presented in Figure 2.6. As can be seen, the results are reasonably good.

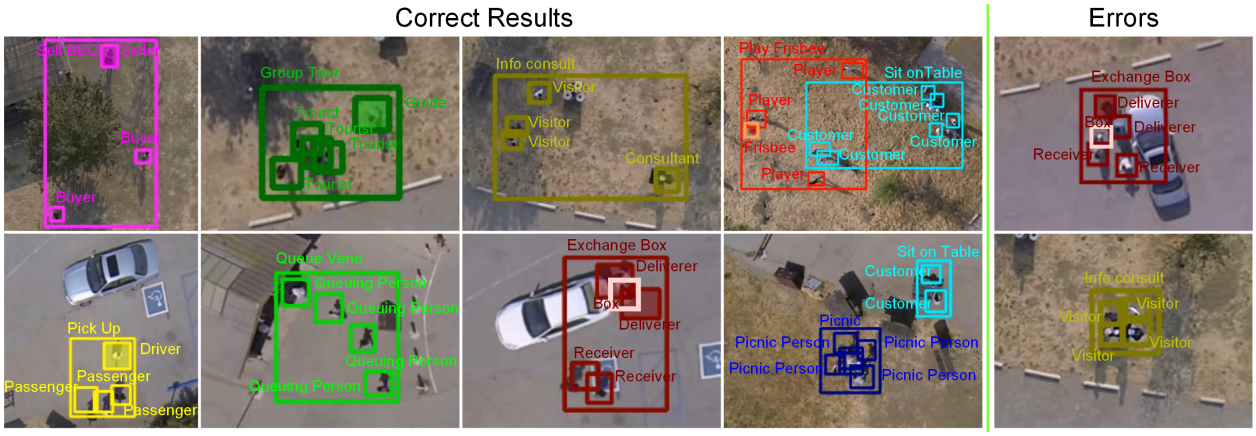
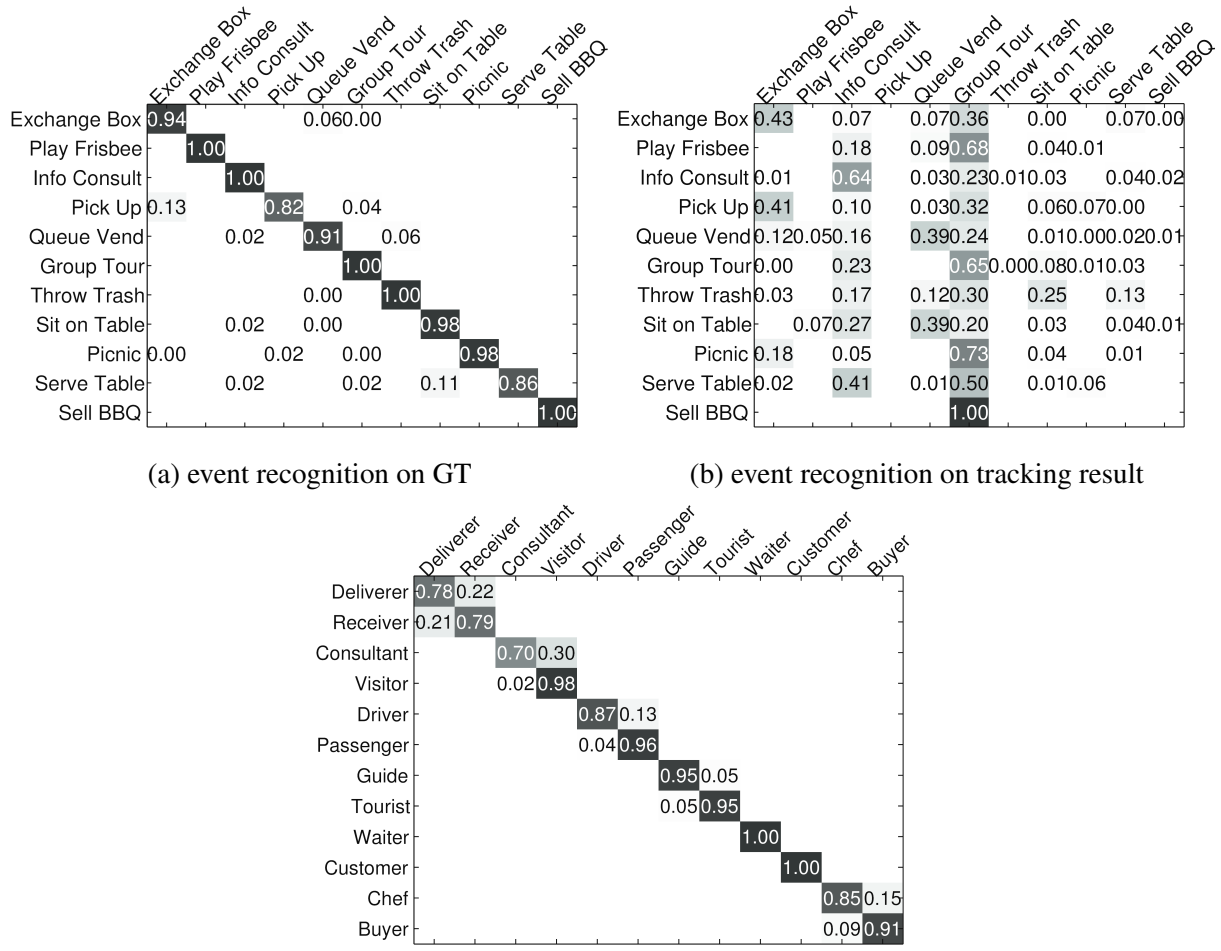


Figure 2.6: Visualization of results including groups (large bounding boxes), events (text) and human roles (small bounding boxes with text). In events with more than one role, we use the shaded bounding box to represent the second role; small portable objects are labeled with lighter color. From event and human role recognition, we can group people even when they are far from each other (e.g., *Play Frisbee* and *Sell BBQ*). In the top-rightmost failure example, true event *Pick Up* is wrongly recognized as *Exchange Box* because one person's trajectory is inferred as *Box*. In bottom-rightmost failure example, our event recognition is correct, but true *Consultant* role is wrongly inferred as *Visitor* role.



(a) event recognition on GT

(b) event recognition on tracking result

(c) role assignment on GT

Figure 2.7: Confusion matrices of event recognition and role assignment result. (a) is event recognition result based on ground-truth (GT) bounding boxes and object labels; (b) is result based on real tracking and detections. From (a) and (b) we can see that *Info Consult*, *Sit on Table*, *Serve Table* cannot be easily distinguished from each other solely based on noisy tracklets. Some events (e.g. *Group Tour*) tend to be wrongly favored by our approach, especially when we do not observe some distinguishing objects. (c) is role assignment result confusion matrix within event class based on ground-truth bounding boxes and object labels. Each 2×2 block is a confusion matrix of role assignment within that event.

Table 2.1: Comparison of our method with baseline methods and variants of our approach. Our method yields best accuracy based on ground-truth bounding boxes and object labels compared to the baseline methods. Using noisy tracking and object detection results, the accuracy is limited, yet better than the baseline methods under the same condition. This demonstrates the advantages of our joint inference. When given access to the ground-truth of objects or people grouping, our results improve. Without reasoning about latent sub-events, accuracy drops significantly, which justifies our model's ability to capture the structural variations of group events.

	Method	Input setting	Group	Event	Role
Baseline Var	[GCR12] for grouping, [CS14] for event and role	GT tracks + object annotation	77.71%	17.22%	13.98%
Baseline	Baseline method as above.	Tracking result	39.64%	16.94%	5.53%
Ours Var1	Our full model	GT tracks + object annotation	95.48%	96.38%	89.94%
Ours Var2	Our full model	Tracking result + object annotation	87.55%	54.75%	28.86%
Ours Var3	Our full model	Tracking result + group labeling	N/A	39.92%	18.71%
Ours Var4	Ours w/o temporal event grammar	Tracking result	40.41%	18.51%	8.69%
Ours	Our full model	Tracking result	49.47%	32.84%	18.92%

The quantitative results are shown in Table 2.1. Confusion matrices of event recognition and role assignment are shown in Figure 2.7. Additional results are presented in the supplementary material[2].

2.7 Conclusion

We collected a new aerial video dataset with detailed annotations, which presents new challenges to computer vision and complements existing benchmarks. We specified a framework for joint inference of events, human roles and people groupings using noisy input. Our experiments showed that addressing each of these inference tasks in isolation is very difficult in aerial videos, and thus provided justification for our holistic framework. Our results demonstrated significant performance improvements over baselines when we constrained uncertainty in input features with domain knowledge.

[2]The supplementary material is available at https://tshu.io/AerialVideo/Aerial_Video_Supp.pdf

Our model is limited and can be extended in two directions. First, we infer the function of the objects implicitly based on the group events currently. In the future, we wish to explicitly infer the functional map for a given site, in the sense that certain area corresponds to specific human activities, e.g., dinning area, parking lot, etc. Unlike appearance-based aerial image parsing [PWZ10], the spatial segmentation will be guided by the spatiotemporal characteristics of human activities. Second, similar to what [XTZ13] did for the prediction of individual intention, we would like to reason the intention of a group as another extension of our work.

CHAPTER 3

CERN: Confidence-Energy Recurrent Network for Group Activity Recognition

3.1 Introduction

In Chapter 2, we have developed a joint inference approach based on a stochastic grammar model, which requires hand-crafted spatiotemporal features. In this chapter, we extend the joint inference to deep neural networks, which can learn these features automatically and potentially achieve a better performance. In particular, the goal of the joint inference is to recognize the overall event of a group of people as well as their individual actions and/or interactions. We leave the grouping inference for future work.

Recent deep architectures [IMD16, RHA16], representing a multi-level cascade of Long Short-Term Memory (LSTM) networks [HS97], have shown great promise in recognizing video events. In these approaches, the LSTMs at the bottom layer are grounded onto individual human trajectories, initially obtained from tracking. These LSTMs are aimed at extracting deep visual representations and predicting individual actions of the respective human trajectories. Outputs of the bottom LSTMs are forwarded to a higher-level LSTM for predicting events. All predictions are made in a feed-forward way using the softmax layer at each LSTM. Such a hierarchy of LSTMs is trained end-to-end using backpropagation-through-time of the cross-entropy loss.

Motivated by the success of these approaches, we start off with a similar two-level hierarchy of LSTMs for recognizing individual actions, interactions, and events. We extend this hierarchy for producing more reliable and accurate predictions in the face of the uncertainty of the visual input.

Ideally, the aforementioned cascade should be learned to overcome uncertainty in a given do-

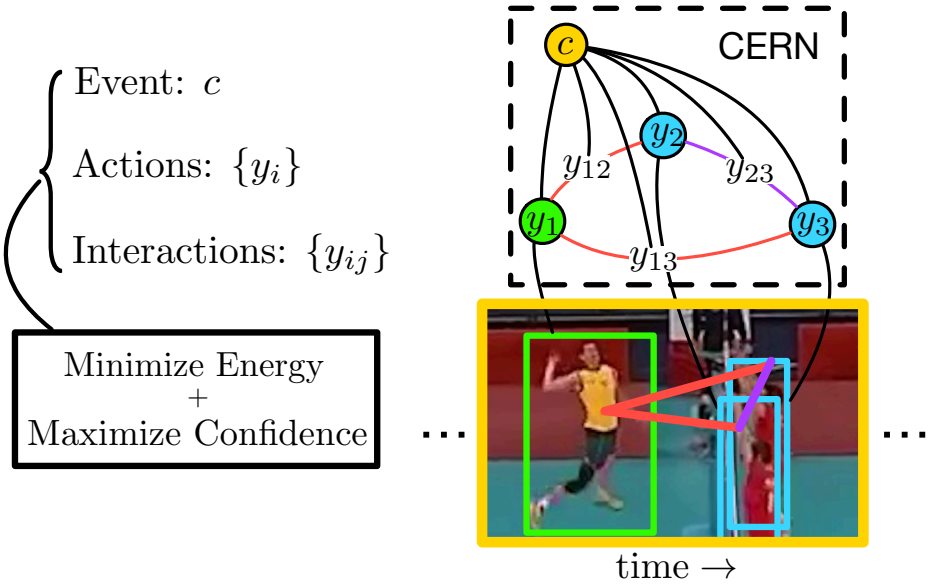


Figure 3.1: Our CERN represents a two-level hierarchy of LSTMs grounded onto human trajectories, where the LSTMs predict individual actions $\{y_i\}$, human interactions $\{y_{ij}\}$, or the event class c in a given video. CERN outputs an optimal configuration of LSTM predictions which jointly minimizes the energy of the predictions and maximizes their confidence, for addressing the brittleness of cascaded predictions under uncertainty. This is realized by extending the two-level hierarchy with an additional energy layer, which can be trained in an end-to-end fashion.

main (e.g., occlusion, dynamic background clutter). However, our empirical evaluation suggests that existing benchmark datasets (e.g., the Collective Activity dataset [CSS09] and the Volleyball dataset [IMD16]) are relatively too small for a robust training of all LSTMs in the cascade. Hence, in cases that have not been seen in the training data, we observe that the feed-forwarding of predictions is typically too brittle, as errors made at the bottom level are directly propagated to the higher level. One way to address this challenge is to augment the training set. But it may not be practical as collecting and annotating group activities is usually difficult.

As shown in Figure 3.1, we take another two-pronged strategy toward more robust activity recognition that includes:

1. Minimizing **energy** of all our predictions at the different semantic levels considered, and

2. Maximizing **confidence** (reliability) of the predictions.

Hence the name of our approach – Confidence-Energy Recurrent Network (CERN).

Our first contribution is aimed at mitigating the brittleness of the direct cascading of predictions in previous work. We specify an energy function for capturing dependencies between all LSTM predictions within CERN, and in this way enable recognition by energy minimization. Specifically, we extend the aforementioned two-layer hierarchy of LSTMs with an additional energy layer (EL) for estimating the energy of our predictions. The EL replaces the common softmax layer at the output of LSTMs. Importantly, this extension allows for a robust, energy-based, and end-to-end training of the EL layer on top of all LSTMs in CERN.

Our second contribution is aimed at improving the numerical stability of CERN's predictions under perturbations in the input, and resolving ambiguous cases with multiple similar-valued local minima. Instead of directly minimizing the energy, we consider more reliable solutions, as illustrated in Figure 3.2. The reliability or confidence of solutions is formalized using the classical tool of a statistical hypothesis test [Fis50] – namely, p-values of the corresponding LSTM's hypotheses (i.e., class predictions). Thus, we seek more confident solutions by regularizing energy minimization with constraints on the p-values. This effectively amounts to a joint maximization of confidence and minimization of energy of CERN outputs. Therefore, we specify the EL to estimate the minimum energy with certain confidence constraints, rather than just the energy.

We also use the energy regularized by p-values for robust deep learning. Specifically, we formulate an energy-based loss which not only accounts for the energy but also the p-values of CERN predictions on the training data.

Our evaluation on the Collective Activity [CSS09] and Volleyball [IMD16] datasets demonstrates: (i) advantages of the above contributions compared with the common softmax and energy-based formulations and (ii) a superior performance relative to the state-of-the-art methods.

In the following, Section 3.2 reviews prior work, Section 3.3 specifies CERN, Section 3.4 and 3.5 formulate the energy and confidence, Section 3.6 describes the energy layer, Section 3.7 specifies our learning, and finally Section 3.8 presents our results.

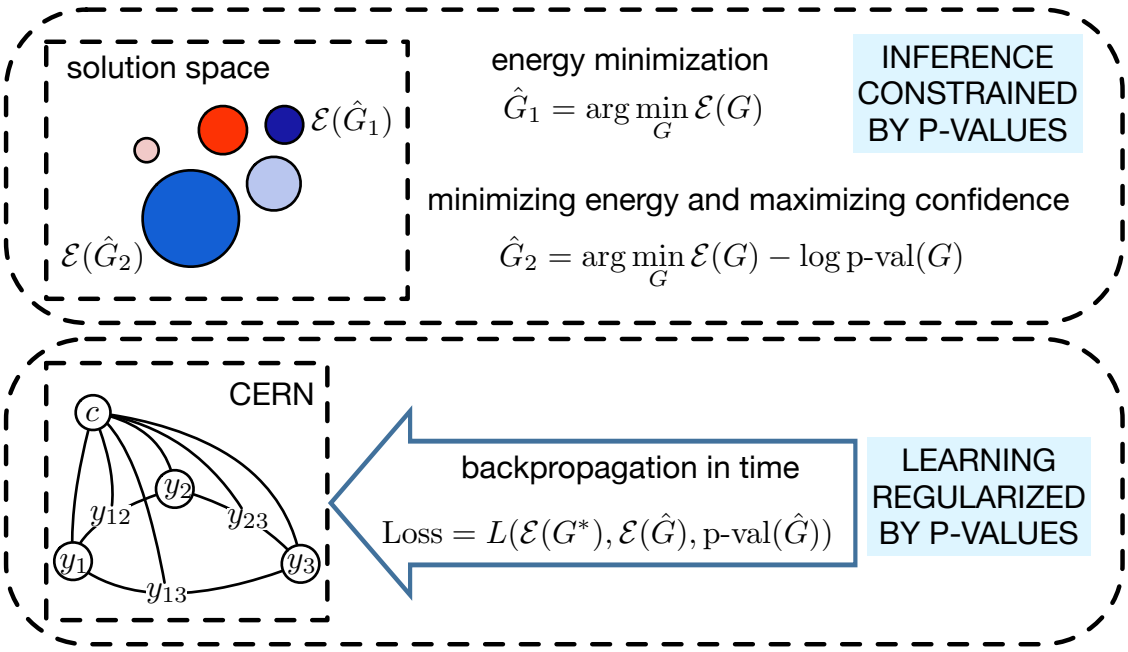


Figure 3.2: (top) An imaginary illustration of the solution space where each circle represents a candidate solution. The colors and sizes of the circles indicate the energy (red:high, blue:low) and confidence (the larger the radius the higher confidence) computed by the energy layer in CERN. A candidate solution $\hat{G}_1$ has the minimum energy, but seems numerically unstable for small perturbations in input. A joint maximization of confidence and minimization of energy gives a different, more confident solution $\hat{G}_2$. Confidence is specified in terms of p-values of the energy potentials. (bottom) We formulate an energy-based loss for end-to-end learning of CERN. The loss accounts for the energy and p-values.

3.2 Related Work

Group activity recognition using DNNs. Previous work typically used graphical models [LSM12, LWY12, RYF13, ALT14, CCP14] or AND-OR grammar models [AXZ12, SXR15] to learn the structures grounded on hand-crafted features. Recent methods learn a graphical model, typically MRF [CSY15, WLT16] or CRF [ZJR15, JZS16, LSF16], using recurrent neural networks (RNNs). Also, work on group activity recognition [IMD16, DVH16] has demonstrated many advantages of using deep architectures of RNNs over the mentioned non-deep approaches. Our approach extends this work by replacing the RNN's softmax layer with a new energy layer, and by specifying a new

energy-based model that takes into account p-values of the network's predictions.

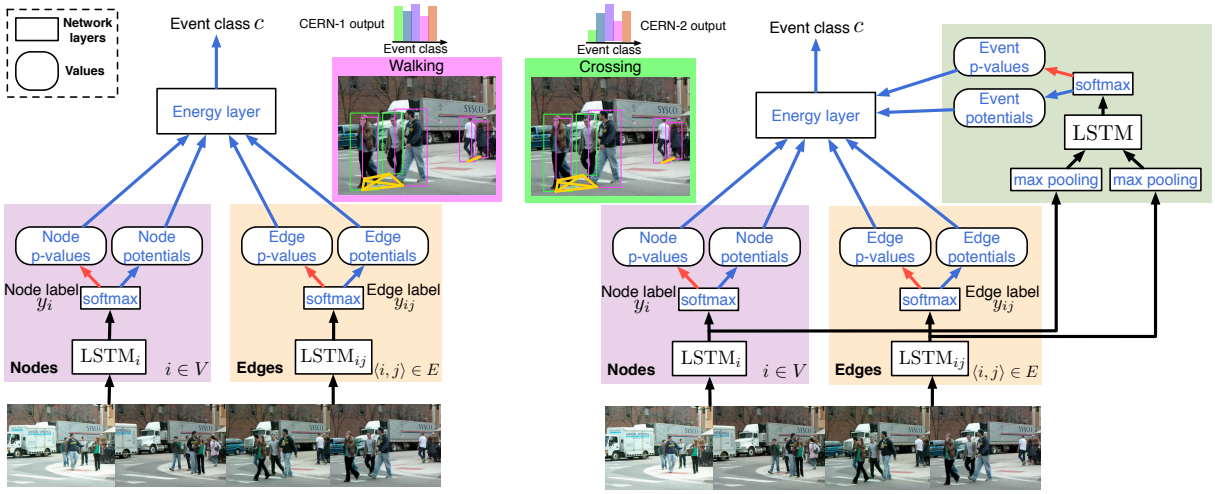
Energy-based learning. While energy-based formulations of inference and learning are common in non-deep group activity recognition [RYF13, ALT14, CCP14, SXR15], they are seldom used for deep architectures. Recently, a few approaches have tried to learn an energy-based model [LH05, LCH06] using deep neural networks [BM16, ZML16]. They have demonstrated that energy-based objectives have great potential in improving the performance of structured predictions, especially when training data are limited. Our approach extends this work by regularizing the energy-based objective such that it additionally accounts for the confidence of predictions.

Reliability of Recognition. Most energy-based models in computer vision have only focused on the energy minimization for various recognition problems. Our approach additionally estimates and regularizes inference with p-values. The p-values are specified within the framework of conformal prediction [SV08]. This allows the selection of more reliable and numerically stable predictions.

3.3 Components of the CERN Architecture

For recognizing events, interactions, and individual actions, we use a deep architecture of LSTMs, called CERN, shown in Figure 3.3. CERN is similar to the deep networks presented in [IMD16, JZS16], and can be viewed as a graph $G = \langle V, E, c, Y \rangle$, where $V = \{i\}$ is the set of nodes corresponding to individual human trajectories, and $E = \{(i, j)\}$ is the set of edges corresponding to pairs of human trajectories. These human trajectories are extracted using an off-the-shelf tracker [DHK14]. Also, $c \in \{1, \cdots, C\}$ denotes an event class (or group activity), and $Y = Y^V \cup Y^E$ is the union set of individual action classes $Y^V = \{y_i : y_i \in \mathcal{Y}^V\}$ and human interaction classes $Y^E = \{y_{ij} : y_{ij} \in \mathcal{Y}^E\}$ associated with nodes and edges.

In CERN, we assign an LSTM to every node and edge in G. All the node LSTMs share the same weights and all the edge LSTMs also have the same weights. These LSTMs use convolutional neural networks (CNNs) to compute deep features of the corresponding human trajectories, and output softmax distributions of individual action classes, $\psi^V(x_i, y_i)$, or softmax distributions of human interaction classes, $\psi^E(x_{ij}, y_{ij})$. The LSTM outputs are then forwarded to an energy layer



(a) CERN-1 (b) CERN-2

Figure 3.3: We specify and evaluate two versions of CERN. CERN is a deep architecture of LSTMs, which are grounded via CNNs to video frames at the bottom. The LSTMs forward their class predictions to the energy layer (EL) at the top. CERN-1 has LSTMs only at the bottom level which compute distributions of individual action classes (colored boxes) or distributions of interaction classes (colored links between green boxes). CERN-2 has an additional LSTM for computing the distribution of event (or group activity) classes. The EL takes the LSTM outputs, and infers an energy minimum with the maximum confidence. The figure shows that CERN-1 and CERN-2 give different results for the group activity *crossing*. CERN-1 wrongly predicts *walking*. CERN-2 typically yields better results for group activities that can not be defined only by individual actions.

(EL) in CERN for computing the energy $\mathcal{E}(G)$. Finally, CERN outputs a structured prediction $\hat{G}$ whose energy has a high confidence:

$$\hat{G} = \arg\min_{G} \mathcal{E}(G) - \log \text{p-val}(G). \tag{3.1}$$

As shown in Figure 3.3, we specify and evaluate two versions of CERN. CERN-1 uses LSTMs for predicting individual actions and interactions, whereas the event class is predicted by the EL as in Eq. (3.1). CERN-2 has an additional event LSTM which takes features maxpooled from the outputs of the node and edge LSTMs, and then computes the distribution of event classes, $\psi(c)$. The EL in CERN-2 takes all three types of class distributions as input – specifically, $\{\psi^V(x_i, y_i)\}_{i \in V}$, $\{\psi^E(x_{ij}, y_{ij})\}_{(i,j) \in E}$, and $\psi(c)$ – and predicts an optimal class assignment as in Eq. (3.1).

In the following, we specify $\mathcal{E}(G)$ and p-val(G).

3.4 Formulation of Energy

For CERN-1, the energy of G is defined as

$$\begin{aligned} \mathcal{E}(G) \;\propto\; & \sum_{i \in V} w^V_{c,y_i} \psi^V(x_i, y_i) && \text{node potential} \\ + \; & \sum_{(i,j) \in E} w^E_{c,y_{ij}} \psi^E(x_{ij}, y_{ij}) && \text{edge potential,} \end{aligned} \tag{3.2}$$

where w^V_{c,y_i} and $w^E_{c,y_{ij}}$ are parameters, $\psi^V(x_i, y_i)$ denotes the softmax output of the corresponding node LSTM, and $\psi^E(x_{ij}, y_{ij})$ denotes the softmax output of the corresponding edge LSTM (see Section 3.3), and x_i and x_{ij} denote visual cues extracted from respective human trajectories by a CNN as in [DVH16, IMD16].

For CERN-2, the energy in Eq. (3.2) is augmented by the softmax output of the event LSTM, i.e.,

$$\begin{aligned} \mathcal{E}(G) \;\propto\; & \sum_{i \in V} w^V_{c,y_i} \psi^V(x_i, y_i) && \text{node potential} \\ + \; & \sum_{(i,j) \in E} w^E_{c,y_{ij}} \, \psi^E(x_{ij}, y_{ij}) && \text{edge potential} \\ + \; & w_c \psi(x, c) && \text{event potential,} \end{aligned} \tag{3.3}$$

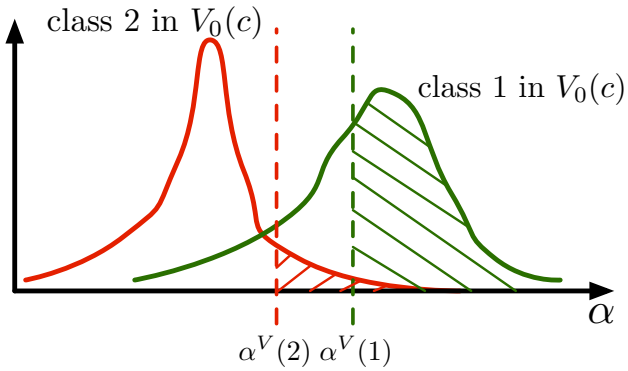


Figure 3.4: A simple illustration of the relationship between the nonconformity measure α of individual actions and the p-value, where the ratio of the dashed region to the whole area under the curve indicates the p-value. Clearly, for the given instance, action class 2 has a larger softmax output but action class 1 has a higher confidence. $V_0(c)$ is the training set of videos showing event c.

where x in $\psi(x, c)$ is the visual representation of all actions and interactions maxpooled from the outputs of the node LSTMs and edge LSTMs.

3.5 Formulation of Confidence

There are several well-studied ways to define the p-values [Fis50]. In this chapter, we follow the framework of conformal prediction [SV08]. Conformal prediction uses a nonconformity (dissimilarity) measure to estimate the extent to which a new prediction is different from the system's predictions made during training. Hence, it provides a formalism to estimate the confidence of new predictions based on the past experience on the training data. Below, we define the nonconformity measure, which is used to compute the p-values for LSTMs' predictions of individual actions, interactions, and events.

3.5.1 Nonconformity Measure and P-values

Given the node potential $\psi^V(x_i, y_i)$, we define a nonconformity measure for action predictions:

$$\alpha^V(y_i) = 1 - \frac{\psi^V(x_i, y_i)}{\sum_{y \in \mathcal{Y}^V} \psi^V(x_i, y)} = 1 - \psi^V(x_i, y_i), \tag{3.4}$$

where the above derivation step holds because $\psi^V(x_i, y_i)$ is the softmax output normalized over action classes. $\alpha^V(y_i)$ is used to estimate the p-value of predicting action class y_i under the context of event class c as

$$p_i^V(c, y_i) = \frac{\sum_{i' \in V_0(c)} \mathbb{1}(y_{i'} = y_i)\mathbb{1}(\alpha^V(y_{i'}) \geq \alpha^V(y_i))}{\sum_{i' \in V_0(c)} \mathbb{1}(y_{i'} = y_i)}. \tag{3.5}$$

where $\mathbb{1}(\cdot)$ is the indicator, and $V_0(c)$ denotes the set of all human trajectories in training videos with ground truth labels $y_{i'}$ and belonging to the ground truth event class c. From Eq. (3.5), the LSTM prediction $\psi^V(x_i, y_i)$ is reliable – i.e., has a high p-value – when many training examples i' of the same class have larger nonconformity measures.

To better understand the relationship between the nonconformity measure and the p-value, let us consider a simple case illustrated in Figure 3.4. The figure plots the two distributions of nonconformity measures of two action classes in the training examples (green: class 1, red: class 2). Suppose that we observe a new instance whose softmax output indicates that action class 2 has a higher probability to be the true label, i.e., $\psi^V(x_i, 1) < \psi^V(x_i, 2)$, and $\alpha^V(1) > \alpha^V(2)$. From the two curves, however, we see that this softmax output is very likely to be wrong. This is because from Figure 3.4 we have the p-values $p_i^V(c, 1) > p_i^V(c, 2)$, since a majority of training examples with the class 1 label have larger nonconformity measures than $\alpha^V(1)$, and hence class 1 is a more confident solution.

Similarly, given the softmax output of the edge LSTM, $\psi^E(x_{ij}, y_{ij})$, we specify a nonconformity measure of predicting interaction classes:

$$\alpha_{ij}^E(y_{ij}) = 1 - \frac{\psi^E(x_{ij}, y_{ij})}{\sum_{y \in \mathcal{Y}^E} \psi^E(x_{ij}, y)} = 1 - \psi^E(x_{ij}, y_{ij}), \tag{3.6}$$

which is then used to estimate the p-value of predicting interaction class y_{ij} under the context of

event class c as

$$p_{ij}^{E}(c, y_{ij}) = \frac{\sum_{(i',j')\in E_0(c)} \mathbb{1}(y_{i'j'} = y_{ij})\mathbb{1}(\alpha_{i'j'}^{E}(y_{i'j'}) \geq \alpha_{ij}^{E}(y_{ij}))}{\sum_{(i',j')\in E_0(c)} \mathbb{1}(y_{i'j'} = y_{ij})}, \tag{3.7}$$

where $E_0(c)$ denotes the set of all pairs of human trajectories in training videos with ground truth labels $y_{i'j'}$ and belonging to the ground truth event class c. From Eq. (3.7), the LSTM prediction $\psi^{E}(x_{ij}, y_{ij})$ has a high p-value when many training examples (i', j') in $E_0(c)$ have larger nonconformity measures.

Finally, in CERN-2, we also have the LSTM softmax output $\psi(x, c)$, which is used to define a nonconformity measure for event predictions:

$$\alpha(c) = 1 - \frac{\psi(x, c)}{\sum_{c\in\mathcal{C}} \psi(x, c)} = 1 - \psi(x, c), \tag{3.8}$$

and the p-value of predicting event class c as

$$p(c) = \frac{\sum_{v\in V_0} \mathbb{1}(c_v = c)\mathbb{1}(\alpha(c_v) \geq \alpha(c))}{\sum_{v\in V_0} \mathbb{1}(c_v = c)}. \tag{3.9}$$

where V_0 denotes the set of all training videos.

3.5.2 Confidence of the Structured Prediction G

To define the statistical significance of the hypothesis G among other hypotheses (i.e., possible solutions), we need to combine the p-values of predictions assigned to nodes, edges and the event of G. More rigorously, for specifying the p-value of a compound statistical test, p-val(G), consisting of multiple hypotheses, we follow the Fisher's combined hypothesis test [Fis50]. The Fisher's theory states that N independent hypothesis tests, whose p-values are $p_1, \cdots p_N$, can be characterized by a test statistic χ_{2N}^2 as

$$\chi_{2N}^2 = -2\sum_{n=1}^{N} \log p_n, \tag{3.10}$$

where the statistic χ_{2N}^2 is proved to follow the χ^2 probability distribution with $2N$ degrees of freedom. From Eq. (3.10), it follows that minimization of the statistic χ_{2N}^2 will yield the maximum p-value characterizing the Fisher's combined hypothesis test.

In the following section, we will use this theoretical result to specify the energy layer of our CERN.

3.6 The Energy Layer of CERN

We extend the deep architecture of LSTMs with an additional energy layer (EL) aimed at jointly minimizing the energy, given by Eq. (3.3), and maximizing a p-value of the Fisher's combined hypothesis test, given by Eq. (3.10). For CERN-2, this optimization problem can be expressed as

$$\begin{array}{ll} \min\limits_{c,Y} & \mathcal{E}(G) \\ \text{s.t.} & -\sum_{i\in V'} \log p_i^V(c, y_i) \leq \tau^V, \\ & -\sum_{(i,j)\in E'} \log p_{ij}^E(c, y_{ij}) \leq \tau^E, \\ & -\log p(c) < \tau^c, \end{array} \tag{3.11}$$

where τ^V, τ^E, and τ^c are parameters that impose lower-bound constraints on the p-values. Recall that according to the Fisher's theory on a combined hypothesis test, decreasing the constraint parameters τ^V, τ^E, and τ^c will enforce higher p-values of the solution.

From Eq. (3.3) and Eq. (3.11), we derive the following Lagrangian, also referred to as regularized energy $\tilde{\mathcal{E}}(X, Y, c)$, which can then be readily implemented as the EL:

$$\begin{array}{rcl} \tilde{\mathcal{E}}(X, Y, c) & = & \sum\limits_{i\in V} w_{c,y_i}^V \psi^V(x_i, y_i) - \lambda^V \sum\limits_{i\in V} \log p_i^V(c, y_i) \\ & & + \sum_{(i,j)\in E} w_{c,y_{ij}}^E \psi^E(x_{ij}, y_{ij}) - \lambda^E \sum_{(i,j)\in E} \log p_{ij}^E(c, y_{ij}) \\ & & + w_c \psi(x, c) - \lambda \log p(c), \end{array} \tag{3.12}$$

Note that for CERN-1, we drop the last two terms in Eq. (3.12), $w_c\psi_c$ and $\lambda \log p(c)$. $\tilde{\mathcal{E}}(X, Y, c)$ can be expressed in a more compact form as

$$\begin{array}{rcl} \tilde{\mathcal{E}}(X, Y, c) & = & \mathbf{w}_c^{V\top} \boldsymbol{\psi}^V - \boldsymbol{\lambda}^{V\top} \log \mathbf{p}_c^V \\ & & + \mathbf{w}_c^{E\top} \boldsymbol{\psi}^E - \boldsymbol{\lambda}_c^{E\top} \log \mathbf{p}_c^E \\ & & + w_c \psi_c - \lambda \log p_c, \end{array} \tag{3.13}$$

where all parameters, potentials, and p-values are grouped into corresponding vectors. Specifically, parameters of the EL are grouped into $\{w_c\}_{c=1,\cdots,C}$, λ, and the following parameter vectors:

$$\begin{array}{ll} \mathbf{w}_c^V = \left[w_{c,1}^V, \cdots, w_{c,|\mathcal{Y}^V|}\right]^\top, & \mathbf{w}_c^E = \left[w_{c,1}^E, \cdots, w_{c,|\mathcal{Y}^E|}\right]^\top, \\ \boldsymbol{\lambda}^V = \left[\lambda^V, \cdots, \lambda^V\right]^\top, & \boldsymbol{\lambda}^E = \left[\lambda^E, \cdots, \lambda^E\right]^\top, \end{array} \tag{3.14}$$

and the input to the EL is specified in terms of the LSTM softmax outputs and p-values:

$$
\begin{aligned}
\boldsymbol{\psi}^V &= \Big[\sum_{i:y_i=1} \psi^V(x_i, y_i), \cdots, \sum_{i:y_i=|\mathcal{Y}^V|} \psi^V(x_i, y_i) \Big]^\top, \\
\boldsymbol{\psi}^E &= \Big[\sum_{\substack{(i,j):\\ y_{ij}=1}} \psi^E(x_{ij}, y_{ij}), \cdots, \sum_{\substack{(i,j):\\ y_{ij}=|\mathcal{Y}^E|}} \psi^E(x_{ij}, y_{ij}) \Big]^\top, \\
\psi_c &= \psi(x, c), \\
\mathbf{p}_c^V &= \Big[\sum_{i:y_i=1} p_i^V(c, y_i), \cdots, \sum_{i:y_i=|\mathcal{Y}^V|} p_i^V(c, y_i) \Big]^\top, \\
\mathbf{p}_c^E &= \Big[\sum_{(i,j):y_{ij}=1} p_{ij}^V(c, y_i), \cdots, \sum_{(i,j):y_{ij}=|\mathcal{Y}^E|} p_{ij}^V(c, y_i) \Big]^\top, \\
p_c &= p(c).
\end{aligned}
\tag{3.15}
$$

Figure 3.5a shows a unit in the EL which computes Eq. (3.13). After stacking these units, as shown in Figure 3.5b, we select the solution $\hat{G}$ with the minimum $\tilde{\mathcal{E}}(\hat{G})$.

In the following, we explain our energy-based end-to-end training of the EL.

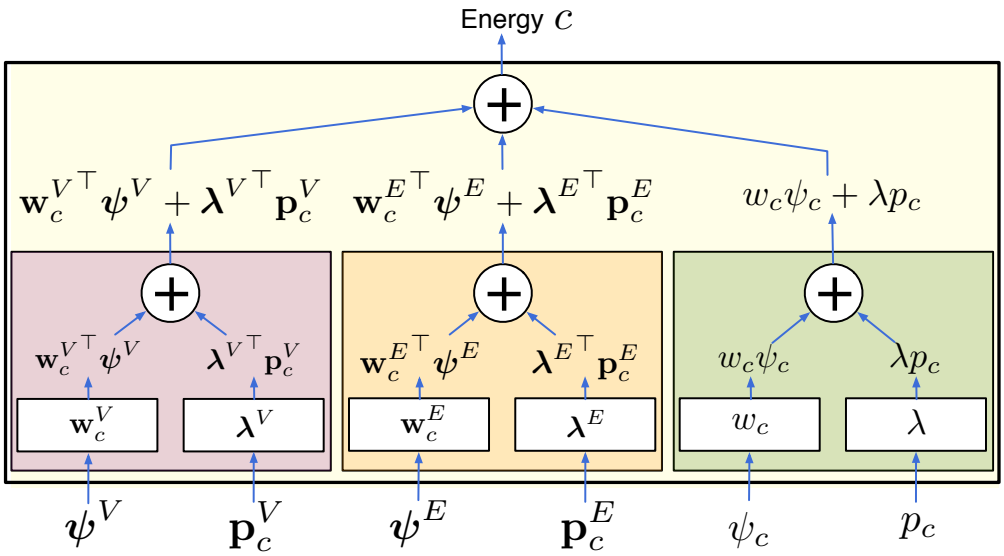
3.7 Learning Regularized By Confidence

Following [LCH06, BM16], we use an energy-based loss for a training instance X^i and its ground truth labels (Y^i, c^i) to learn parameters of the EL, i.e., the regularized energy, specified in Eq. (3.12):

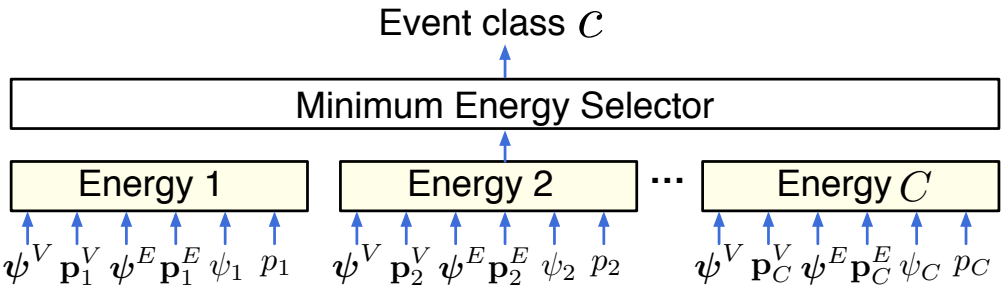
$$
L(X^i, Y^i, c^i) = \max \Big(0, \tilde{\mathcal{E}}(X^i, Y^i, c^i) - \tilde{\mathcal{E}}(X^i, \bar{Y}, \bar{c}) + \mathbb{1}(c^i \neq \bar{c}) \Big), \tag{3.16}
$$

where $\bar{Y}, \bar{c} = \operatorname{argmin}_{Y, c \neq c^i} \tilde{\mathcal{E}}(X^i, Y, c) - \mathbb{1}(c^i \neq c)$ is the most violated case. Alternatively, this loss can be replaced by other energy-based loss functions also considered in [LCH06]. Here we treat Y as latent variables for simplicity and thus only consider accuracy of c. However, one can include a comparison between Y and its corresponding ground truth label Y^i into the loss function. It is usually difficult to find the most violated case. However, as [LH05] points out, the inference of the most violated case does not require a global minimum solution since the normalization term is not modeled in our energy-based model, so we can simply set $\bar{Y}$ to be the output of the node and edge LSTMs.

In practice, one can first train a network using common losses such as cross-entropy to learn the



(a) The unit for computing the regularized energy of category c, given by Eq. (3.13).



(b) Diagram of all units in the energy layer.

Figure 3.5: The EL takes the softmax outputs of all LSTMs along with estimated p-values as input, and outputs a solution that jointly minimizes the energy and maximizes a p-value of the Fisher's combined hypothesis test.

representation excluding the EL, namely from the input layer to softmax layers. Then the p-value of a training instance can be computed by removing itself from the training sets V_0 and E_0. Finally we train the weights in Eq. (3.12) by minimizing the loss.

3.8 Results

Implementation details. We stack the node LSTMs and edge LSTMs on top of a VGG-16 model [SZ14] without the FC-1000 layer. The VGG-16 is pre-trained on ImageNet [DDS09], and fine-

tuned with LSTMs jointly. We train the top layer of CERN by fixing the weights of the CNNs and the bottom layer LSTMs. The batch size for the joint training of the bottom LSTMs and VGG-16 is 6. The training converges within 20000 iterations. The event LSTM and the EL are trained using 10000 iterations with a batch size of 2000. For the mini-batch gradient descent, we use RMSprop [TH] with a learning rate ranging from 0.000001 to 0.001. We use Keras [Cho15] with Theano [The16] as the backend to implement CERN, and run training and testing with a single NVIDIA Titan X (Pascal) GPU. For a fair comparison with [IMD16], we use the same tracker and its implementation as in [IMD16]. Specifically, we use the tracker of [DHK14] from the Dlib library [Kin09]. The cropped image sequences of persons and pairs of persons are used as the inputs to node LSTMs and edge LSTMs, respectively.

We compare our approach with the state-of-the-art methods [HYV15, IMD16]. In addition, we evaluate the following reasonable baselines.

Baselines:

- 2-layer LSTMs (B1). We test a network of 2-layer LSTMs similar to [IMD16]. All other baselines below and our full models use B1 to compute their potentials and p-values. B1 does not have the energy layer, but only a feed-forward network. The event class is predicted by the softmax output of the event LSTM.
- CERN-1 w/o p-values (B2). This baseline represents the CERN-1 network with the EL, however, the p-values are not computed and not used for regularizing energy minimization. Hence, the event class prediction of B2 comes from the standard energy minimization.
- CERN-2 w/o p-values (B3). Similar to B2, in this B3, we do not estimate and do not use the p-values in the EL of CERN-2.

Datasets. We evaluate our method in two domains: collective activities and sport events using the Collective Activity dataset [CSS09] and the Volleyball dataset [IMD16] respectively.

3.8.1 Collective Activity Dataset

The Collective Activity dataset consists of 44 videos, annotated with 5 activity categories (*crossing*, *walking*, *waiting*, *talking*, and queueing), 6 individual action labels (*NA*, *crossing*, *walking*, *waiting*,

Table 3.1: Comparison of different methods for group activity recognition on the Collective Activity dataset.

Method	MCA	MPCA
Cardinality kernel [HYV15]	83.4	81.9
2-layer LSTMs [IMD16]	81.5	80.9
B1: 2-layer LSTMs	79.7	80.3
B2: CERN-1 w/o p-values	83.8	84.3
B3: CERN-2 w/o p-values	83.8	83.7
CERN-1	84.8	85.5
CERN-2	**87.2**	**88.3**

talking, and queueing), and 8 pairwise interaction labels (*NA*, *approaching*, *leaving*, *passing-by*, *facing-each-other*, *walking-side-by-side*, *standing-in-a-row*, *standing-side-by-side*). The interaction labels are provided by the extended annotation in [CCP14].

For this dataset, we first train the node LSTMs and edge LSTMs with 10 time steps and 3000 nodes. Then, we concatenate the outputs of these two types of LSTMs at the bottom layer of CERN, along with their VGG-16 features, and pass the concatenation to the bidirectional event LSTM with 500 nodes and 10 time steps at the top layer of CERN. The concatenation is passed through a max pooling layer and a fully-connected layer with a output dimension of 4500.

For comparison with [HYV15, IMD16] and baselines B1-B3, we use the following performance metrics: (i) multi-class classification accuracy (MCA), and (ii) mean per-class accuracy (MPCA). Our split of training and testing sets is the same as in [HYV15, IMD16]. Table 3.1 summarizes the performance of all methods on recognizing group activities. Note that in Table 3.1 only [HYV15] does not use deep neural nets. As can be seen, our energy layer significantly boosts the accuracy, outperforming the state-of-the-art by a large margin. Even when we only have the bottom layer of LSTMs, CERN-1 still outperforms the 2-layer LSTMs in [IMD16] thanks to the EL. Without the EL, the baseline B1 yields lower accuracy than [IMD16] even with additional LSTMs for the interactions.

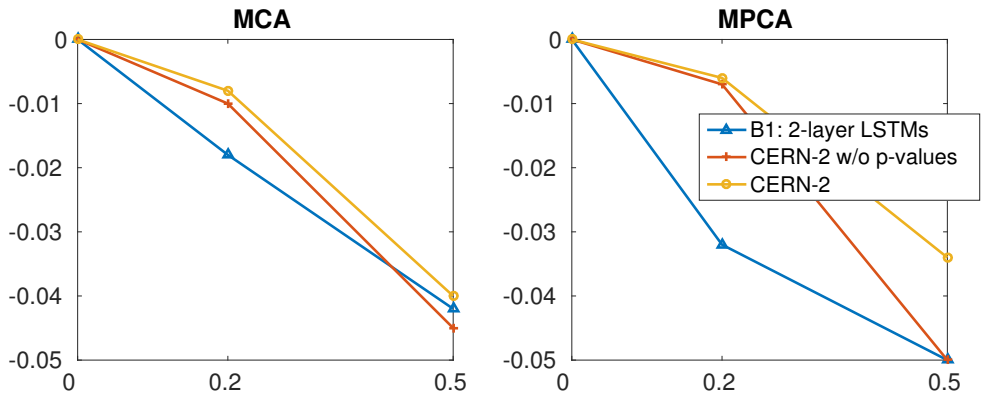


Figure 3.6: Performance decrease of group activity recognition for a varying percentage of corruption of human trajectories in the Collective Activity dataset. We compare 2-layer LSTMs (B1), CERN-2 w/o p-values (B3) and CERN-2 using the same corrupted trajectories as input.

Our accuracies of recognizing individual actions and interactions on the Collective Activity dataset are 72.7% and 59.9%, using the node LSTMs and edge LSTMs respectively. Note that B1, CERN-1 and CERN-2 share the same node and edge LSTMs.

For evaluating numerical stability of predicting group activity classes by CERN-2, we corrupt all human trajectories in the testing data, and control the amount of corruption with the corruption probability. For instance, for the corruption probability of 0.5, we corrupt one bounding box of a person in every video frame with a 0.5 chance. When the bounding box is selected, we randomly shift it with a horizontal and a vertical displacement ranging from 20% to 80% of the original bounding box's width and height respectively. As Figure 3.6 shows, CERN-2 consistently experiences a lower degradation in performance compared to the baselines without p-values. This indicates that incorporating the p-values into the energy model indeed benefits the inference stability. Such benefit becomes more significant as the amount of corruption in input data increases.

Figure 3.7 shows an example of the *crossing* activity. As can be seen, although B1 and CERN-2 share the same individual action labels, where a majority of the people are assigned incorrect action labels, CERN-2 can still correctly recognize the activity.

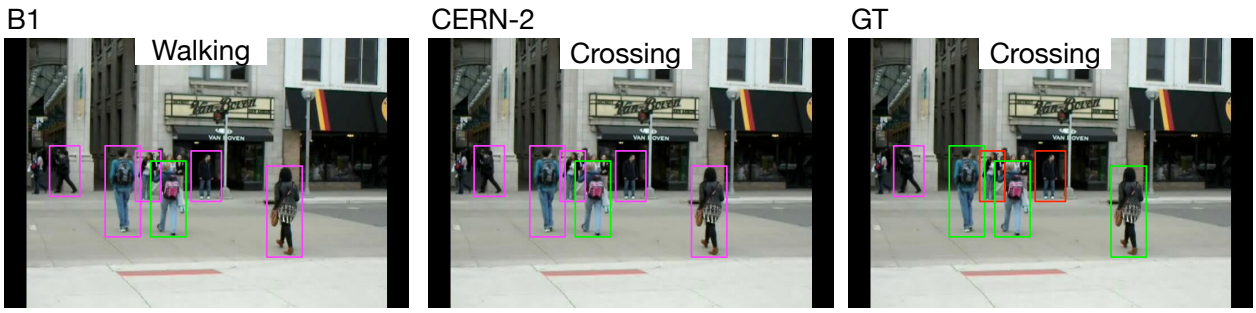


Figure 3.7: The qualitative results on the Collective Activity dataset. From left to right, we show the inference results from B1, CERN-2 and the ground truth (GT) labels respectively. The colors of the bounding boxes indicate the individual action labels (green: *crossing*, red: *waiting*, magenta: *walking*). The interaction labels are not shown here for simplicity.

3.8.2 Volleyball Dataset

The Volleyball dataset consists of 55 videos with 4830 annotated frames. The actions labels are *waiting*, *setting*, *digging*, *failing*, *spiking*, *blocking*, *jumping*, *moving*, and *standing*; and the group activity classes include *right set*, *right spike*, *right pass*, *right winpoint*, *left winpoint*, *left pass*, *left spike*, and *left set*. Interactions are not annotated in this dataset, so we do not recognize interactions and remove the edge LSTMs.

The node LSTMs have 3000 nodes and 10 time steps (including 5 preceding and 4 succeeding frames). The event LSTM in CERN-2 is a bidirectional LSTM with 1000 nodes and 10 time steps. In [IMD16], the max pooling has two types: 1) pooling over the output of all node LSTMs, or 2) dividing the players into two groups (the left team and the right team) first and pooling over each group separately. We test both types of max pooling for our approach to rule out the effect of pooling type in the comparison. CERN-1 does not have the pooling layer, thus is categorized as 1 group style.

Recognition accuracy of individual actions is 69.1% using node LSTMs, and the accuracies of recognizing group activities are summarized in Table 3.2. Cleary, the regularized energy minimization increases the accuracy compared to the conventional energy minimization (B2 and B3), and CERN-2 outperforms the state-of-the-art when using either of the pooling types. CERN-1 does not achieve accuracy that is comparable to that of CERN-2 on the Volleyball dataset. This

Table 3.2: Comparison of different methods for group activity recognition on the Volleyball dataset. The first block is for the methods with 1 group and the second one is for those with 2 groups.

Method	MCA	MPCA
2-layer LSTMs [IMD16] (1 group)	70.3	65.9
B1: 2-layer LSTMs (1 group)	71.3	69.5
B2: CERN-1 w/o p-values (1 group)	33.3	34.3
B3: CERN-2 w/o p-values (1 group)	71.7	69.8
CERN-1 (1 group)	34.4	34.9
CERN-2 (1 group)	73.5	72.2
2-layer LSTMs [IMD16] (2 groups)	81.9	82.9
B1: 2-layer LSTMs (2 group)	80.3	80.5
B3: CERN-2 w/o p-values (2 groups)	82.2	82.3
CERN-2 (2 groups)	**83.3**	**83.6**

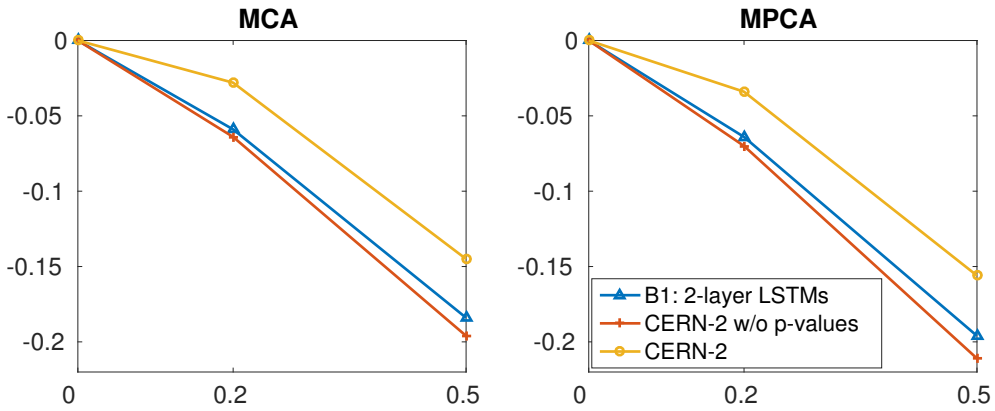


Figure 3.8: The decrease of group activity recognition accuracy over different input distortion percentages on the Volleyball dataset (all use the 2 groups style). CERN-2 is compared with 2-layer LSTMs (B1) and CERN-2 w/o p-values (B3).

is mainly because CERN-1 reasons the group activity based on individual actions, which may not provide sufficient information for recognizing complex group activities in sports videos. CERN-2 overcomes this problem by adding the event LSTM.

We also evaluate the stability of recognizing group activities by CERN-2 under corruption of input human trajectories. As Figure 3.8 indicates, the p-values in the EL indeed increase the inference reliability on the Volleyball dataset.

The qualitative results (2 groups) of a *right pass* activity is depicted in Figure 3.9, which demonstrates the advantage of the inference based on our regularized energy compared to the softmax output of the deep recurrent networks when the action predictions are not accurate.

3.9 Conclusion

We have addressed the problem of recognizing group activities, human interactions, and individual actions with a novel deep architecture, called Confidence-Energy Recurrent Network (CERN). CERN extends an existing two-level hierarchy of LSTMs by additionally incorporating a confidence measure and an energy-based model toward improving reliability and numerical stability of inference. Inference is formulated as a joint minimization of the energy and maximization of the confidence measure of predictions made by the LSTMs. This is realized through a new dif-

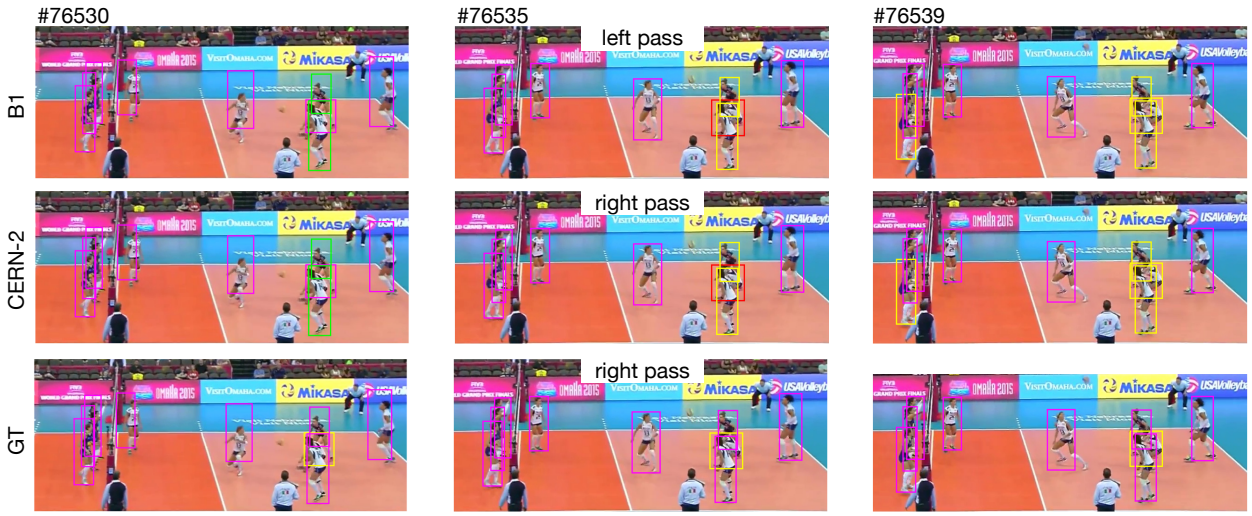


Figure 3.9: The qualitative results on the Volleyball dataset: results of B1 (top), results of CERN-2 (middle) and the ground truth (GT) labels (bottom). The colors of the bounding boxes indicate the individual action labels (green: *waiting*, yellow: *digging*, red: *falling*, magenta: *standing*), and the numbers are the frame IDs.

ferentiable energy layer (EL) that computes the energy regularized by a p-value of the Fisher's combined statistical test. We have defined an energy-based loss in terms of the regularized energy for learning the EL end-to-end. CERN has been evaluated on the Collective Activity dataset and Volleyball dataset. In comparison with previous approaches that predict group activities in a feed-forward manner using deep recurrent networks, CERN gives a superior performance, and also gives more numerically stable solutions under uncertainty. For collective activities, our simpler variant CERN-1 gives more accurate predictions than a strong baseline representing a two-level hierarchy of LSTMs with softmax outputs taken as predictions. Our variant CERN-2 increases complexity but yields better accuracy on challenging group activities which are not merely a sum of individual actions but a complex whole.

CHAPTER 4

Learning Social Affordances

4.1 Introduction

The concept of "affordance learning" is receiving an increasing amount of attention from robotics, computer vision, and human-robot interaction (HRI) researchers. The term *affordance* was originally defined as "action possibilities" of things (e.g., objects, environments, and other agents) by [Gib79], and it has attracted researchers to study computational modeling of such concept [MLB08, GSE11, KRK11, MMO12, JKS13, ZFF14, KS14, PEK14, PEK15, SDC15, ZZZ15]. The idea behind modern affordance learning research is to enable robot learning of "what activities are possible" (i.e., semantic-level affordances) and "where/how it can execute such activities" (i.e., spatial-level and motion-level affordances) from human examples. Such ability not only enables robot planning of possible actions, but also allows robots to replicate complicated human activities. Based on training videos of humans performing activities, the robot will infer when particular sub-events can be executed and how it should move its own body-parts in order to do so.

So far, most previous works on robot affordance learning have only focused on the scenario of a single robot (or a single human) manipulating an object (e.g., [KS14]). These systems assumed that affordance solely depends on the spatial location of the object, its trajectory, and the intended action of the robot. Consequently affordance was defined as a unary function in the sense that there is only one agent (i.e., the robot) involved.

However, in order for a robot to perform collaborative tasks and *interact* with humans, computing single-robot object manipulation affordances based on object recognition is insufficient. In these human-robot interaction scenarios, there are multiple agents (humans and robots) in the scene and they interact and react. Thus, the robot must (1) represent its affordance as "interactions" be-

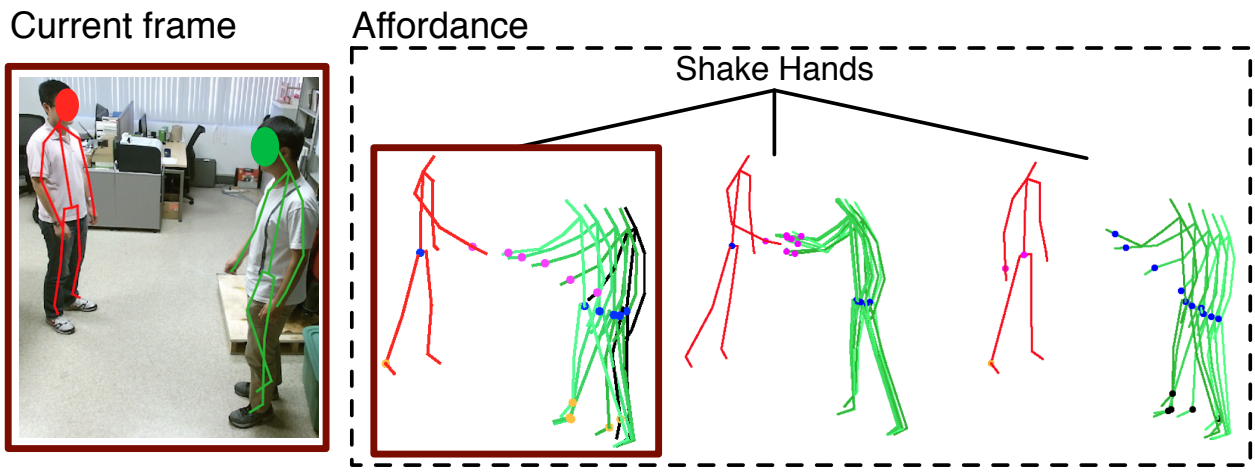


Figure 4.1: Visualization of our social affordance. The green (right) person is considered as our agent (e.g., a robot), and we illustrate (1) what sub-event the agent needs to do given the current status and (2) how it should move in reaction to the red (left) person's body-parts to execute such sub-event. The black skeleton indicates the current frame estimation, and greens are for future estimates. The right figure shows a hierarchical activity affordance representation, where affordance of each sub-event is described as the motion of body joints. We also visualize the learned affordable joints with circles, and their grouping is denoted by the colors. Note that the grouping varies in different sub-events.

tween body joints of multiple agents, and (2) learn to compute such hierarchical affordances based on its status. Its affordance should become activated only when the action makes sense in the social context. For instance, the fact that human's hand is a location of affordance doesn't mean that the robot can grab it whenever it feels like. The robot should consider grabbing the human hand only when the person is interested in performing hand-shake activity with it.

Therefore, in this chapter, we introduce the new concept of *social affordances*, and present an approach to learn them from human videos. We formulate the problem as the learning of structural representations of social activities describing how the agents and their body-parts move. Such representation must contain a sufficient amount of information to execute the activity (e.g., how should it be decomposed? what body-parts are important? how should the body-parts move?), allowing its social affordance at each time frame to be computed by inferring the status of the activity and by computing the most appropriate motion to make the overall activity successful (Figure 4.1).

Since we consider the problem particularly in the context of human-robot interaction, activity representation involving multiple agents with multiple affordable body-parts must be learned, and the inference on a robot's affordance should be made by treating it as one of the agents.

Our problem is challenging for the following reasons: (i) human skeletons estimated from RGB-D videos are noisy due to occlusion, making the learning difficult; (ii) human interactions have much more complex temporal dynamics than simple actions; and (iii) our affordance learning is based on a small training set with only weak supervision.

For the learning, we propose a Markov Chain Monte Carlo (MCMC) based algorithm to iteratively discover latent sub-events, important joints, and their functional grouping from noisy and limited training data. In particular, we design two loops in the learning algorithm, where the outer loop uses a Metropolis-Hasting algorithm to propose temporal parsing of sub-events for each interaction instance (i.e., sub-event learning), and the inner loop selects and groups joints within each type of sub-event through a modified Chinese Restaurant Process (CRP). Based on the discovered latent sub-events and affordable joints, we learn both spatial and motion potentials for grouped affordable joints in each sub-event. For the motion synthesis, we apply the learned social affordance to unseen scenarios, where one agent is assumed to be an observed human, and the other agent is assumed to be the robot that we control to interact with the observed agent (an object will be treated as part of the observation if it is also involved). To evaluate our approach, we collected a new RGB-D video dataset including 3 human-human interactions and 2 human-object-human interactions. Note that there are no human-object-human interactions in the existing RGB-D video datasets.

To our knowledge, this is the first work to study robot learning of affordances for social activities. Our work differs from the previous robot affordance learning works in the aspect that it (1) considers activities of multiple agents, (2) decomposes activities into multiple sub-events/sub-goals and learns their affordances (i.e., hierarchical affordance) that are grounded to the skeleton sequences, and (3) learns both spatial and motion affordances of multiple body-parts involved in interactions.

4.2 Related work

Although there are previous studies on vision-based hierarchical activity recognition [GSS09, RA11, LWY12, AXZ12, PSY13, CCP14, SXR15] and human-human interaction recognition [RA11, LCS14, HK14], research on affordances of high-level activities has been very limited. For the robotic motion planning and object manipulation, [LSK13, YLC15, WZS15] presented symbolic representation learning methods for single agent activities, but low-level joint trajectories were not explicitly modeled in those works. In computer graphics, some motion synthesis approaches have been proposed [LWS02, THR06, WFH08, FLP15], but they only learn single agent motion based on highly accurate skeleton inputs from motion capture systems.

In contrast, in this chapter, we are studying affordances of dynamic agents with multiple body parts, including human-human interactions (e.g., shaking hands) as well as human-object-human interactions (e.g., object throw-catch). Its importance was also pointed out in [Gib79] as "the richest and most elaborate affordances", and we are exploring such concept for the first time for robots. We specifically denote such affordances as *social affordances*, and present an approach to learn them from human activity videos.

4.3 Representation and Formulation

We propose a graphical model to represent the social affordance in a hierarchical structure, which is grounded to skeleton sequences (Figure 4.2a). Our representation not only describes what human skeletons (i.e., body-joint locations) are likely to be observed when two persons are performing interactions, but also indicates how each interaction need to be decomposed in terms of sub-events/sub-goals and how agents should perform such sub-events in terms of joint motion.

Skeleton sequences. An interaction instance is represented by the skeleton sequences of the two agents. We use $J^t = \{J^t_{1i}\} \cup \{J^t_{2i}\}$ to denote the positions of the two agents' joints at time $t = 1, \cdots, T$. If an interaction involves an object, then $J^t = \{J^t_{1i}\} \cup \{J^t_{2i}\} \cup O^t$, where O^t is the position of the object at t. In practice, we select 5 most important joints – base joint, left/right writs, and left/right ankles for the social affordance, whose indexes are denoted as a set $\mathcal{I}$. This

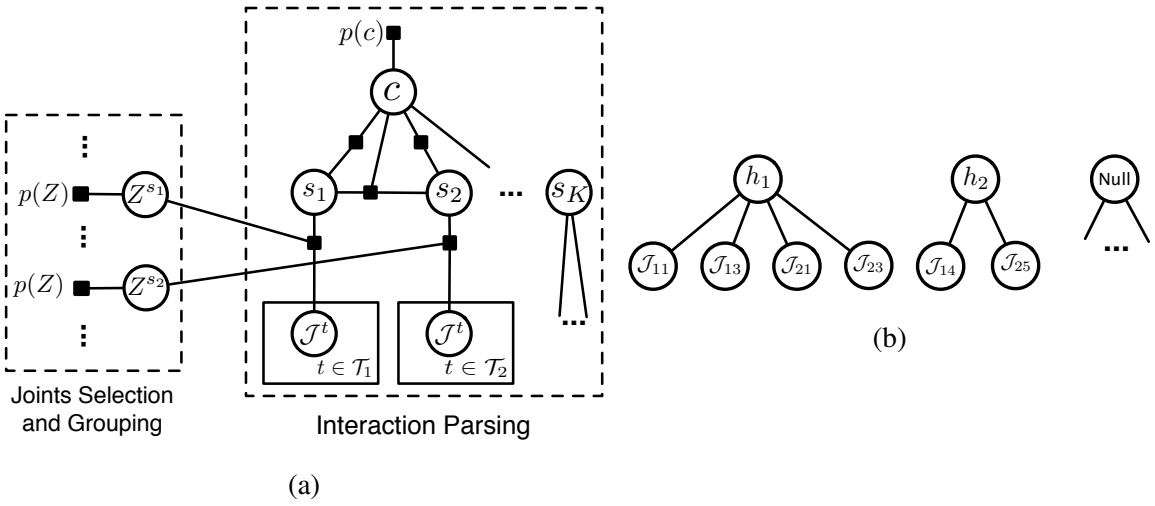


Figure 4.2: Our model. (a) Factor graph of an interaction. (b) Selection and grouping of joints for a sub-event.

reasonable simplification helps us eliminate the noise introduced by skeleton extraction from RGB-D videos while maintaining the overall characteristics of each interaction.

Interaction label. A label $c \in \mathcal{C}$ is given to an interaction to define its category, where $\mathcal{C}$ is a predefined dictionary.

Latent sub-events. One of our key intuitions is that a complex interaction usually consists of several steps. In order to enable the robots to mimic the human behavior, it is necessary to discover these underlying steps as latent sub-events. Here, a sub-event is defined as a sub-interval within a complete interaction. There are two crucial components in a sub-event: 1) the sub-goal to achieve at the end of the sub-event, and 2) the motion patterns to follow in this sub-event. Since it is difficult for humans to manually define and annotate the sub-events, we only specify the number of latent sub-events, i.e., $|\mathcal{S}|$, and our learning method automatically searches the optimal latent sub-event parsing for each training instance. Here, a latent sub-event parsing of an interaction instance whose length is T is represented by non-overlapping intervals $\{\mathcal{T}_k\}_{k=1,\cdots,K}$ such that $\sum_k |\mathcal{T}_k| = T$, where $\mathcal{T}_k = \{t : t = \tau_k^1, \cdots, \tau_k^2\}$, and the sub-event labels of the corresponding intervals, i.e., $\{s_k\}_{k=1,\cdots,K}$. Note that K, the number of sub-events, may vary in different instances.

Joint selection and grouping. Another key intuition of ours is to discover the affordable joints and their functional groups in each latent sub-event. This means that 1) some joints do

not contribute much to accomplishing the latent sub-event due to the lack of clear motion and/or specific spatial relations among them, 2) and the rest joints are regarded as affordable joints and are further clustered together to form several functional groups, each of which has rigid spatial relations among the grouped joints in the sub-events. Figure 4.2b illustrates the selection and grouping of joints in a sub-event: we first select affordable joints with a Bernoulli distribution prior and remain the rest joints in a $Null$ group; then we assign each affordable joint to a functional group from a infinity number of latent functional classes $\mathcal{H} = \{h_1, \cdots, h_\infty\}$. The grouping can be addressed by a Chinese Restaurant Process (CRP), where a functional class is a table, and each affordable joint can be perceived as a customer to be seated at a table. We introduce auxiliary variables $Z^s = \{z^s_{ai} : z^s_{ai} \in \mathcal{H}, a \in \{1, 2\}, i = 1, \cdots, N_J\}$ to indicate the joint selection and grouping in a sub-event $s \in \mathcal{S}$ of interaction $c \in \mathcal{C}$. J_{ai} is assigned to $h_{z^s_{ai}}$ if $z^s_{ai} > 0$; otherwise, J_{ai} is assigned to the *Null* group. Together $Z_c = \{Z^s\}_{s \in \mathcal{S}}$ represents the joint selection and grouping in a type of interaction, c.

Sub-goals and motion patterns. After grouping joints, the sub-goal of a sub-event is defined by the spatial relations (i.e., spatial potentials Ψ_g) among joints within the functional groups, and movements of affordable joints are described with the motion pattens (i.e., motion potentials Ψ_m). These allow us to infer "how" each agent should move.

Parse graph. As shown in Figure 4.2a, an interaction instance is represented by a parse graph $G = \langle c, S, \{J^t\}_{t=1,\cdots,T} \rangle$. With the corresponding joint selection and grouping Z_c, we formalize the social affordance of an interaction as $\langle G, Z_c \rangle$. Note that Z_c is fixed as common knowledge while G depends on the observed instance.

4.3.1 Probabilistic Modeling

In this subsection, we provide how our approach models the joint probability of each parse graph G and the joint selection and grouping Z, allowing us to use it for both (i) learning the structure and parameters of our representation based on observed human skeletons (Section 4.4) and (ii) inferring/synthesizing new skeleton sequences for the robot using the learned model (Section 4.5).

For each interaction c, our social affordance representation has two major parts: 1) optimal

body-joint selection and grouping Z_c, and 2) parse graph G for each observed interaction instance of c. Given Z_c, the probability of G for an instance is defined as

$$\begin{aligned} p(G|Z_c) \quad \propto \quad & \underbrace{\prod_k p(\{J^t\}_{t\in\mathcal{T}_k}|Z^{s_k}, s_k, c)}_{\text{likelihood}} \cdot \underbrace{p(c)}_{\text{interaction prior}} \\ & \cdot \underbrace{\prod_{k=2}^{K} p(s_k|s_{k-1}, c)}_{\text{sub-event transition}} \cdot \underbrace{\prod_{k=1}^{K} p(s_k|c)}_{\text{sub-event prior}}, \end{aligned} \tag{4.1}$$

and the prior for joint selection and grouping is

$$p(Z_c) = \prod_{s\in\mathcal{S}} p(Z^s|c). \tag{4.2}$$

Hence the joint probability is

$$p(G, Z_c) = p(G|Z_c)p(Z_c). \tag{4.3}$$

Likelihood. The likelihood term in Eq. (4.1) consists of i) spatial potential $\Psi_g(\{J^t\}_{t\in\mathcal{T}}, Z^s, s)$ for the sub-goal in sub-event s, and ii) motion potential $\Psi_m(\{J^t\}_{t\in\mathcal{T}}, Z^s, s)$ for motion patterns of the affordable joints in s:

$$p(\{J^t\}_{t\in\mathcal{T}}|Z^s, s, c) = \Psi_g(\{J^t\}_{t\in\mathcal{T}}, Z^s, s)\Psi_m(\{J^t\}_{t\in\mathcal{T}}, Z^s, s). \tag{4.4}$$

Spatial potential. We shift the affordable joints at the end of each sub-event (i.e., τ^2) in an interaction w.r.t. the mass center of the assigned functional group. The shifted joint locations at t are denoted as $\widetilde{J}^t_{ai}$. If there is only one joint in a group, the reference point will be the base joint location of the other agent at the moment instead. Then for each joint, we have

$$\psi_g(\widetilde{J}^t_{ai}) = \psi_{\text{xy}}(\widetilde{J}^t_{ai})\psi_{\text{z}}(\widetilde{J}^t_{ai})\psi_{\text{o}}(\widetilde{J}^t_{ai}), \tag{4.5}$$

where $\psi_{\text{xy}}(\widetilde{J}^t_{ai})$ and $\psi_{\text{z}}(\widetilde{J}^t_{ai})$ are Weibull distributions of the horizontal and vertical distance between the joint and the reference point, and $\psi_{\text{o}}(\widetilde{J}^t_{ai})$ is a von Mises distribution for the angle between the two points. Note that the spatial potential only accounts for affordable joints (i.e., $z^s_{ai} > 0$). Thus

$$\Psi_g(\{J^t\}_{t\in\mathcal{T}}, Z^s, s) = \prod_{a,i} \psi_g(\widetilde{J}^{\tau^2}_{ai})^{\mathbb{1}(z^s_{ai}>0)}. \tag{4.6}$$

Motion potential. In a sub-event s of an interaction, we compute the movement of a joint J_{ai} by $\boldsymbol{d}_{ai} = J_{ai}^{\tau^2} - J_{ai}^{\tau^1}$. Similar to the spatial potential, this joint's motion potential is

$$\psi_m(\{J_{ai}^t\}_{t\in\mathcal{T}}) = \psi_m(\boldsymbol{d}_{ai}) = \psi_{\text{xy}}(\boldsymbol{d}_{ai})\psi_{\text{z}}(\boldsymbol{d}_{ai})\psi_{\text{o}}(\boldsymbol{d}_{ai}). \tag{4.7}$$

For an affordable joint, we use Weibull distributions for both horizontal and vertical distances and a von Mises distribution for the orientation. To encourage static joints to be assigned to the $Null$ group, we fit exponential distributions for the distances while keeping $\psi_{\text{o}}(\boldsymbol{d}_{ai})$ the same if $z_{ai}^s = 0$. Hence,

$$\Psi_m(\{J^t\}_{t\in\mathcal{T}}, Z^s, s) = \prod_{a,i} \psi_m(\{J_{ai}^t\}_{t\in\mathcal{T}_k}). \tag{4.8}$$

Prior for interaction category and sub-event transition. We assume uniform distribution for $p(c)$ and compute the transition frequency from training data for $p(s_k|s_{k-1}, c)$.

Sub-event prior. The duration of a sub-event s_k in interaction c is regularized by a log-normal distribution $p(s_k|c)$:

$$p(s_k|c) = \exp\{-(\ln|\mathcal{T}_k| - \mu)^2/(2\sigma^2)\}/(|\mathcal{T}_k|\sigma\sqrt{2\pi}). \tag{4.9}$$

Joint selection and grouping prior. Combined with Bernoulli distribution and the prior of CRP, the joint selection and grouping prior for Z^s in sub-event type s of interaction c is defined as

$$p(Z^s|c) = \underbrace{\frac{\prod_h (M_h - 1)!}{M!}}_{\text{CRP prior}} \prod_{ai} \underbrace{\beta^{\mathbb{1}(z_{ai}^s>0)}(1-\beta)^{(1-\mathbb{1}(z_{ai}^s>0))}}_{\text{Bernoulli prior for a joint}}. \tag{4.10}$$

where M_h is the number of joints assigned to latent function group h, and M is the total number of affordable joints, i.e., $M = \sum_{a,i} \mathbb{1}(z_{ai}^s > 0)$.

4.4 Learning

Given the skeleton sequences and their interaction labels, we learn the model for each interaction category in isolation. Assume that we have N training instances for interaction c, then will have N parse graphs $\mathcal{G} = \{G^n\}_{n=1,\ldots,N}$, and a common Z_c for this type of interaction. The objective of our

Algorithm 1 Learning Algorithm

Input: $\{J^t\}_{t=1,\cdots,T}$ of each instance with the same interaction category $c \in \mathcal{C}$
Obtain the atomic time intervals by K-means clustering
Initialize S of each instance, and Z_c
repeat
    Propose S'
    repeat
        Sample new Z_c through Gibbs sampling
    until Convergence
    $\alpha = \min\{\frac{Q(S' \to S)P^*(\mathcal{G}', Z_c')}{Q(S \to S')P^*(\mathcal{G}, Z_c)}, 1\}$
    $u \sim Unif[0, 1]$
    If $u \leq \alpha$, accept the proposal S'
until Convergence

learning algorithm is to find the optimal $\mathcal{G}$ and Z_c that maximize the following joint probability:

$$p(\mathcal{G}, Z_c) = p(Z_c) \prod_n^N p(G^n | Z_c). \tag{4.11}$$

Note that the size of latent sub-event dictionary, $|\mathcal{S}|$, is specified for each interaction.

We propose a MCMC learning algorithm as Alg. 1, which includes two optimization loops:

1 Metropolis-Hasting algorithm for sub-event parsing.

2 Given sub-event parsing, apply Gibbs sampling for the optimization $Z_c^* = \text{argmax}_{Z_c} p(\mathcal{G}, Z_c) = \text{argmax}_{Z_c} p(\mathcal{G}|Z_c)p(Z_c)$.

The details of two loops are introduced as follows.

4.4.1 Outer Loop for Sub-Event Parsing

In the outer loop, we optimize the sub-event parsing by a Metropolis-Hasting algorithm. We first parse each interaction sequence into atomic time intervals using K-means clustering of agents' skeletons (we use 50 clusters). Then the sub-events are formed by merging some of the atomic

time intervals together. At each iteration, we propose a new sub-event parsing S' through one of the following dynamics:

Merging. In this dynamics, we merge two sub-events with similar skeletons together and uniformly sample a new sub-event label for it, which forms a new sub-event parsing S'. For this, we first define the distance between two consecutive sub-events by the mean joint distance between the average skeletons in these two sub-events, which is denoted by d. Then the proposal distribution is $Q(S \rightarrow S'|d) \propto e^{-\lambda d}/N_L$, where λ is a constant number, and N_L is number of possible label assignments for the new sub-event. In practice, we set $\lambda = 1$.

Splitting. We can also split a sub-event with multiple atomic time intervals into two non-overlapping sub-events with two new labels. Note that an atomic time interval is not splittable. Similarly, we can compute the distance d between the average skeletons of these two new sub-events and assume uniform distributions for the new labels. To encourage the split of two sub-events with very different skeletons, we define the proposal distribution to be $Q(S \rightarrow S'|d) \propto (1 - e^{-\lambda d})/N_L$, where N_L is number of possible new labels.

Re-labeling. We relabel a uniformly sampled sub-event for this dynamics, which gives the proposal distribution $Q(S \rightarrow S'|d) = 1/(N_L \cdot N_S)$, where N_L and N_S are the numbers of possible labels and current sub-events respectively.

In addition, the type of dynamics at each iteration is sampled w.r.t. these three probabilities, $q_1 = 0.4$, $q_2 = 0.4$, $q_3 = 0.2$, for the above three types respectively.

The acceptance rate α is then defined as $\alpha = \min\{\frac{Q(S' \rightarrow S)P^*(\mathcal{G}', Z_c')}{Q(S \rightarrow S')P^*(\mathcal{G}, Z_c)}, 1\}$, where $P^*(\mathcal{G}, Z_c)$ is the highest joint probability given current sub-event parsing S, i.e., $P^*(\mathcal{G}, Z_c) = \max_{Z_c} p(\mathcal{G}, Z_c)$. Similarly, $P^*(\mathcal{G}', Z_c') = \max_{Z_c'} p(\mathcal{G}', Z_c')$.

4.4.2 Inner Loop for Joint Selection and Grouping

To obtain $P^*(\mathcal{G}', Z_c')$ in the acceptance rate defined for the outer loop given the proposed S', we use Gibbs sampling to iteratively update Z_c'. At each iteration, we assign a joint from $\mathcal{I}$ to a new

group in each type of sub-event by

$$z_{ai}^s \sim p(\mathcal{G}|Z_c')p(z_{ai}^s|Z_{-ai}^s). \tag{4.12}$$

Based on Eq. (4.10), we have

$$p(z_{ai}^s|Z_{-ai}^s) = \begin{cases} \beta\dfrac{\gamma}{M-1+\gamma} & \text{if } z_{ai}^s > 0, M_{z_{ai}^s} = 0 \\ \beta\dfrac{M_{z_{ai}^s}}{M-1+\gamma} & \text{if } z_{ai}^s > 0, M_{z_{ai}^s} > 0 \\ 1-\beta & \text{if } z_{ai}^s = 0 \end{cases} \tag{4.13}$$

where the variables have the same meaning as in Eq. (4.10) and $\beta = 0.3$ and $\gamma = 1.0$ are the parameters for our CRP.

4.5 Motion Synthesis

Our purpose for learning social affordance is to teach a robot how to interact with a human. Hence, we design an online simulation method to "synthesize" a skeleton sequence (i.e., $\{J_{2i}\}_{t=1,\cdots,T}^t$) as a robot's action sequence to interact with a human (i.e., the first agent) and an object given the observed skeleton sequence (i.e., $\{J_{1i}\}_{t=1,\cdots,T}^t$), where T is the length of the interaction. The idea is to make our approach automatically "generate" an agent's body joint motion based on the learned social affordance and the other agents' motion. Note that the human skeleton sequence has not been seen in the training data and we assume that the interaction category c is given. The estimated object trajectory $\{O^t\}_{t=1,\cdots,T}$ will also be used if an object is involved. Since we define the social affordance for a interaction instance as $\langle G, Z_c \rangle$, the synthesis is essentially to infer the joint locations for the second agent (i.e., $\{J_{2i}\}^t$) by maximizing the joint probability defined in Eq. (4.3).

The main steps of our motion synthesis are summarized in Alg. 2. At any time t, we first use a dynamic programming (DP) algorithm to estimate current sub-event type based on our observations of the human agent (and the object if it exists) as well as the skeletons that we have synthesized so far. Then we sample the new joint locations by maximizing the spatial and motion potentials under current sub-event.

Algorithm 2 Motion Synthesis Algorithm

Give the interaction label c and the total length T; set unit time interval for simulation to be $\Delta T = 5$; input the skeletons in the first $T_0 = 10$ frames, i.e., $\{J^t\}_{t=1,\cdots,T_0}$; set $\tau \leftarrow T_0$

repeat

  Input $\{J_{1i}\}^t_{t=\tau+1,\cdots,\tau+\Delta T}$

  Extend $\{J_{2i}\}^t$ to $\tau + \Delta T$ by copying $\{J_{2i}\}^\tau$ temporarily

  Infer S of $\{J^t\}_{t=1,\cdots,\tau+\Delta T}$ by DP; we assume that the last sub-event, s_K, is the current on-going sub-event type

  Predict the ending time τ^2_K of s_K by sampling the complete duration $|\mathcal{T}|$ w.r.t. the prior defined in Eq. (4.9), and generate $N = 100$ possible samples for the locations of the modeled five joints in $\mathcal{I}$, i.e., $\{\hat{J}^n_{2i'}\}_{i'\in\mathcal{I},n=1,\cdots,N}$; note that the joints in the *Null* group are assumed to be static in the current sub-event

  Obtain the N corresponding joint locations at current time $\tau + \Delta T$, $\{J^n_{2i'}\}_{i'\in\mathcal{I},n=1,\cdots,N}$, by interpolation based on $\{\hat{J}^n_{2i'}\}$

  We choose the one that maximizes the likelihood, i.e., $\{J^*_{2i'}\}_{i'\in\mathcal{I}}$, by computing motion and spatial potentials

  Fit clustered full body skeletons from K-means to $\{J^*_{2i'}\}_{i'\in\mathcal{I}}$ by rotating limbs, and obtain the closest one $\{J^*_{2i}\}$

  $J^{\tau+\Delta T}_{2i} \leftarrow J^*_{1i}$

  Interpolate the skeletons from $\tau + 1$ to $\tau + \Delta T$, and update $\{J_{2i}\}^t_{t=\tau+1,\cdots,\tau+\Delta T}$

  $\tau \leftarrow \tau + \Delta T$

until $\tau \geq T$

4.5.1 Dynamic Programming

We use the following DP algorithm to efficiently infer the latent sub-events given the skeletons of two agents (and the object trajectory if present) by maximizing the probability of the parse graph defined in Eq. (4.1). For a sequence of interaction c, we first define $m(s', t', s, t)$ as the log probability of assigning sub-event type s to the time interval $[t', t]$ when the preceding sub-event

type is s', which can be computed as

$$\begin{aligned} m(s', t', s, t) &= \log p(\{J^t\}_{t\in[t',t]}|Z^s, s, c) \\ &\quad + \log p(t - t' + 1|s, c) + \log p(s|s', c) \end{aligned} \tag{4.14}$$

Then we define the highest log posterior probability for assigning type s to the last sub-event of $\{J^t\}_{t=1,\cdots,t}$ as $b(s, t)$:

$$b(s, \tau) = \max_{s'\neq s, t'<t} \{b(s', t') + m(s', t', s, t)\} \tag{4.15}$$

where $b(s, 0) = 0$. By recording all pairs of s' and t' that maximize $b(s, t)$ in Eq. (4.15), we can easily backtrace the optimal latent sub-event parsing including labels $s_1, \cdots, s_K$ and corresponding intervals $\mathcal{T}_1, \cdots, \mathcal{T}_K$, starting from the last frame until the first frame in a reverse process.

4.6 Experiment

We collected a new RGB-D video dataset, i.e., UCLA Human-Human-Object Interaction (HHOI) dataset, which includes 3 types of human-human interactions, i.e., *shake hands*, *high-five*, *pull up*, and 2 types of human-object-human interactions, i.e., *throw and catch*, and *hand over a cup*. On average, there are 23.6 instances per interaction performed by totally 8 actors recorded from various views. Each interaction lasts 2-7 seconds presented at 10-15 fps. We used the MS Kinect v2 sensor for the collection, and also took advantage of its skeleton estimation. The objects are detected by background subtraction on both RGB and depth images. The dataset is available at: http://tsho.io/SocialAffordance.

We split the instances by four folds for the training and testing where the actor combinations in the testing set are different from the ones in the training set. For each interaction, our training algorithm converges within 100 outer loop iterations, which takes 3-5 hours to run on a PC with an 8-core 3.6 GHz CPU. Our motion synthesis can be ran at the average speed of 5 fps with our unoptimized Matlab code.

Experiment 1: Our approach learns affordance representations from the training set, and uses the testing set to "synthesize" the agent (i.e., robot) skeletons in reaction to the interacting human skeletons (and an object). We first measured the average joint distance between synthesized skeletons and the ground truth (GT) skeletons since good synthesis should not be very different from

Table 4.1: Average joint distance (in meters) between synthesized skeletons and GT skeletons for each interaction.

Method	Shake Hands	Pull Up	High-Five	Throw & Catch	Hand Over	Average
HMM	0.362	0.344	0.284	0.189	0.229	0.2816
V1	0.061	0.144	0.079	0.091	0.074	0.0899
V2	0.066	0.231	0.090	0.109	0.070	0.1132
Ours	**0.054**	**0.109**	**0.058**	**0.076**	**0.068**	**0.0730**

GT. A multi-level hidden Markov model (HMM) is implemented as the baseline method, where the four levels from top to bottom are: 1) the quantized distance between agents, 2) the quantized relative orientation between agents, 3) the clustered status of the human skeleton and the object, and 4) the clustered status of the synthesized skeleton. In addition, we also compare our full model with a few variants: ours without joint selection and grouping (V1), and ours without the latent sub-events (V2). Notice that this social affordance based skeleton synthesis is a new problem and we are unaware of any exact prior state-of-the-art approach.

The average joint distance for different methods are compared in Table. 4.1. Our full model outperforms all other approaches by a large margin, which proves the advantage of our hierarchical generative model with latent sub-events and joint grouping. Note that the tracking error of Kinect 2 for a joint ranges from 50 mm and 100 mm [WKO15]. Figure 4.3 demonstrates a few joint selection and grouping results for some automatically discovered latent sub-events in different interactions. We also visualize several synthesized interactions in Figure 4.4, where the synthesized skeletons from ours and the HMM baseline are compared with GT skeletons.

Experiment 2: In addition, we also conducted a user study experiment of comparing the naturalness of our synthesized skeleton vs. ground truths. Similar to [MSI09], we asked 14 human subjects (undergraduate/graduate students at UCLA) to rate the synthesized and GT interactions. For this, we predefined 4 sets of videos, where there were 5 videos for each interaction in a set, and all these 5 videos were either from GT or ours. Thus each set had a mixture of videos of GT and ours, but GT and ours did not co-exist for any interaction. Then we randomly assigned these 4 sets

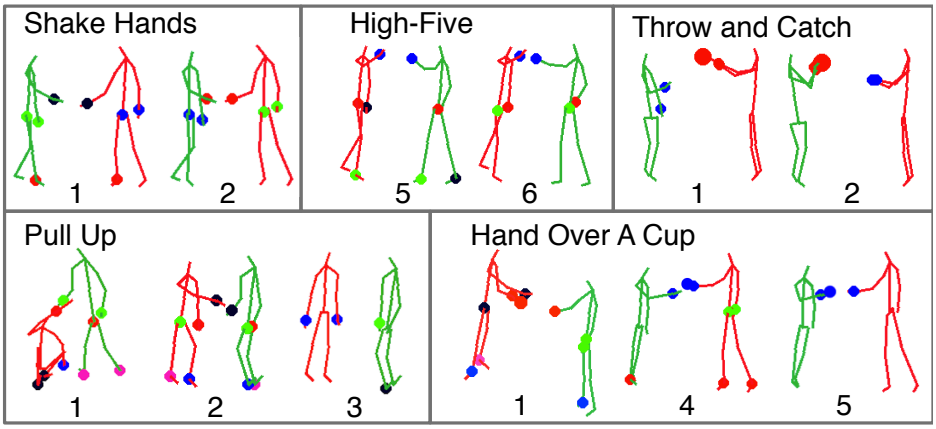


Figure 4.3: Visualization of some discovered sub-events and their joint grouping in the five interactions, where the number denotes the sub-event label and the joint colors show the groups. For *throw and catch* and *hand over a cup*, an object is also displayed as an additional affordable joint. The shown frames are the last moments of the corresponding sub-events, which depict the learned sub-goals.

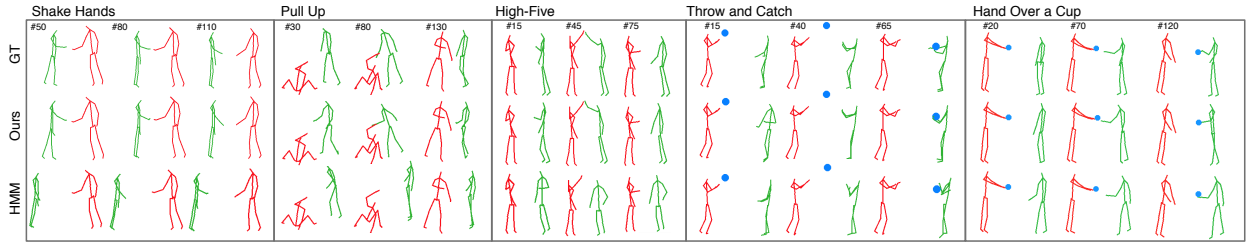


Figure 4.4: Comparison between synthesized and GT skeletons. The red agent and the blue object are observed; the green agents are either GT skeletons, synthesized skeletons by ours, or those by HMM respectively. The numbers are the frame indexes.

to the subjects who were asked to watch each video in the given set only once and rate it from 1 (worst) to 5 for three different questions: "Is the purpose of the interaction *successfully* achieved?" (Q1), "Is the synthesized agent behaving *naturally*?" (Q2), and "Does the synthesized agent look like a *human* rather than a robot?" (Q3). The subjects were instructed that the red skeleton was a real human and the green skeleton was synthesized in all videos. They were not aware of the fact that GT and our synthesized sequences were mixed in the stimuli.

Table 4.2 compares the mean and standard deviation of human ratings per interaction per ques-

Table 4.2: The means and standard deviations of human ratings for the three questions. The highlighted ratings indicate that the sequences synthesized by ours have higher mean ratings than GT sequences.

	Source	Shake Hands	Pull Up	High-Five	Throw & Catch	Hand Over
Q1	Ours	$\mathbf{4.60} \pm 0.69$	3.90 ± 0.70	4.53 ± 0.30	$\mathbf{4.31} \pm 0.89$	4.40 ± 0.37
	GT	4.50 ± 0.82	4.29 ± 0.58	4.64 ± 0.33	4.20 ± 0.76	4.64 ± 0.30
Q2	Ours	$\mathbf{4.23} \pm 0.34$	2.80 ± 0.75	3.70 ± 0.47	$\mathbf{4.06} \pm 0.83$	3.89 ± 0.38
	GT	4.20 ± 0.47	4.23 ± 0.48	4.64 ± 0.17	3.86 ± 0.53	4.24 ± 0.46
Q3	Ours	4.23 ± 0.50	2.63 ± 0.60	3.57 ± 0.73	$\mathbf{4.03} \pm 0.88$	3.69 ± 0.64
	GT	4.30 ± 0.60	3.71 ± 1.15	4.40 ± 0.63	3.97 ± 0.74	4.40 ± 0.24

tion. Following [WN11], we test the equivalence between the ratings of ours and GT for each question using 90% confidence interval. When the equivalence margin is 0.5, *shake hands* and *throw and catch* pass the test for all three questions while the rest interactions only pass the test for Q1. When we consider the equivalence margin to be 1, only *pull up* does not pass the equivalence test for Q2 and Q3. Overall, our motion synthesis is comparable to Kinect-based skeleton estimation, especially for Q1, suggesting that we are able to learn an appropriate social affordance representation. The lower ratings for *pull up* mainly results from much noisier training sequences. Interestingly, the synthesized sequences of *shake hands* and *throw and catch* have sightly higher ratings than GT for Q1 and Q2. This is because our model learns affordances from multiple training sequences, whereas GT is based on a single and noisy Kinect measure. One distinguishable effect is hand touching, which is a critical pattern for the human subjects to rate the videos according to their feedback after the experiment. In GT videos, especially *shake hands* and *throw and catch*, the hand touching (either with another agent's hand or the ball) is not captured due to occlusion, whereas our synthesized skeletons have notably better performances since our method automatically groups the corresponding wrist joints (and the ball) together to learn their spatial relations, as shown in Figure 4.4. This shows that our approach is learning sub-goals of the interactions correctly even with noisy Kinect skeletons.

For Q3, we also counted the frequencies of the high scores (4 or 5) given to the five interactions: 0.87, 0.17, 0.53, 0.77, 0.63 for ours, and 0.88, 0.69, 0.84, 0.66, 0.84 for GT respectively (ordered as in Table 4.2). This is similar to the Turing test: we are measuring whether the subjects perceived the agent as more human-like or more robot-like.

After synthesizing the skeleton sequence, applying the social affordances learned from human activities to the robot replication is straightforward. Since we explicitly represent the spatial and motion patterns of the base joint and the end points of the limbs, we can match them to the corresponding base position and end positions of limbs on a robot. Consequently movement control of these key positions of a robot can be achieved by moving them based on the synthesized trajectories of their human joint counterparts to reach the desired sub-goals. We will implement this on a real robotic system in the future work.

4.7 Conclusion

In this chapter, we discussed the new concept of *social affordance*. We were able to confirm that our approach learns affordance on human body-parts from human interactions, finding important body joints involved in the interactions, discovering latent sub-events, and learning their spatial and motion patterns. We also confirmed that we are able to synthesize future skeletons of agents by taking advantage of the learned affordance representation, and that it obtains results comparable to RGBD-based ground truth skeletons estimated from Kinect.

One future work is to transfer our learned human motion model to a robot motion model. In this chapter, we focused on the affordance "learning" part, and we took advantage of it to synthesize skeleton motion sequences by assuming that humans and robots share their body configurations and motion (i.e., a humanoid robot). However, in practice, robots have different configurations and mechanical constraints than humans. In order for the learned social affordance to be useful for robots in general (e.g., non-humanoid robots), motion transfer is needed as a future research challenge.

CHAPTER 5

Motion Transfer for Human-Robot Interactions Using Social Affordance Grammar

5.1 Introduction

As discussed in Chapter 4, human-robot interactions must follow certain social etiquette or social norms, in order to make humans comfortable, just like human social interactions. When interacting with humans in various social situations, a robot should reason the intentions and feelings of humans who are near it and only perform socially appropriate actions while trying to achieve its own goal.

We have proposed a statistical approach for learning social affordances as hierarchical representations of human interactions in Chapter 4. In this chapter, we aim at developing a real-time motion inference to enable natural human-robot interactions by i) first learning social affordances (i.e., action possibilities following basic social norms) from human interaction videos, and ii) then generating robot plans based on the learned social affordances. More specifically, we are interested in the following three general types of human-robot interactions that we believe are most dominant interactions for robots: i) social etiquette, e.g., greeting, ii) collaboration, e.g., handing over objects, and iii) helping, e.g., pulling up a person who falls down.

To this end, we propose a new representation for social affordances, i.e., social affordance grammar as a spatiotemporal AND-OR graph (ST-AOG), which encodes both important latent sub-goals for a complex interaction and the fine grained motion grounding such as human body gestures and facing directions. We learn the grammar from RGB-D videos of human interactions as Figure 5.1 depicts. Our grammar model also enables short-term motion generation (e.g., raising

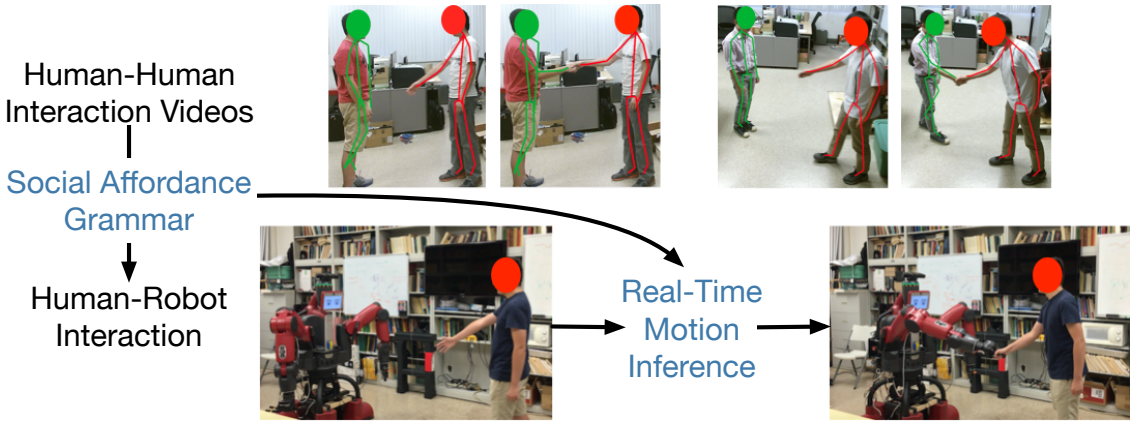


Figure 5.1: The framework of our approach.

an arm) for each agent independently while providing long-term spatiotemporal relations between two agents as sub-goals to achieve for both of them (e.g., holding the right hand of each other), which simultaneously maximizes the flexibly of our motion inference (single agent action) and grasps the most important aspects of the intended human-robot interactions (sub-goals in joint tasks).

Contributions:

1. A general framework for weakly supervised learning of social affordance grammar as a ST-AOG from videos;

2. A real-time motion inference based on the ST-AOG for transferring human interactions to HRI.

5.2 Related Work

Affordances. In the existing affordance research, the domain is usually limited to object affordances [MLB08, KRK11, MMO12, ZFF14, KS14, PEK14, SDC15, ZZZ15], e.g., possible manipulations of objects, and indoor scene affordances [GSE11, JKS13], e.g., walkable or standable surface, where social interactions are not considered. [SRZ16] is the first to propose a social affordance representation for HRI. However, it could only synthesize human skeletons rather than control a real robot, and did not have the ability to generalize the interactions to unseen scenarios.

We are also interested in learning social affordance knowledge, but emphasize on transferring such knowledge to a humanoid in a more flexible setting.

Structural representation of human activities. In recent years, several structural representations of human activities for the recognition purposes have been proposed for human action recognition [GSS09, BT11, PSY13, LZZ15] and for group activity recognition [RA11, LWY12, AXZ12, CCP14, LCS14, SXR15, DVH16]. There also have been studies of robot learning of grammar models [LSK13, YLC15, XSX16], but they were not aimed for HRI.

Social norms learning for robots. Although there are previous works on learning social norms from human demonstrations aimed for robot planning, they mostly focused on relatively simple social scenarios, such as navigation [LSS12, OA16]. On the contrary, we are learning social affordances as a type of social norm knowledge for much more complex interactions, which involve the whole body movements.

5.3 Framework Overview

The framework of our approach illustrated in Figure 5.1 can be outlined as follows:

Human videos. We collect RGB-D videos of human interactions, where human skeletons were extracted by Kinect. We use the noisy skeletons of these interactions as the input for the affordance learning.

Social affordance grammar learning. Based on the skeletons from human interaction videos, we design a Gibbs sampling based weakly supervised learning method to construct a ST-AOG grammar as the representation of social affordances for each interaction category.

Real-Time motion inference. For transferring human interactions to human-robot interactions, we propose a real-time motion inference algorithm by sampling parse graphs as hierarchical plans from the learned ST-AOG and generate human-like motion accordingly for a humanoid to interact with a human agent.

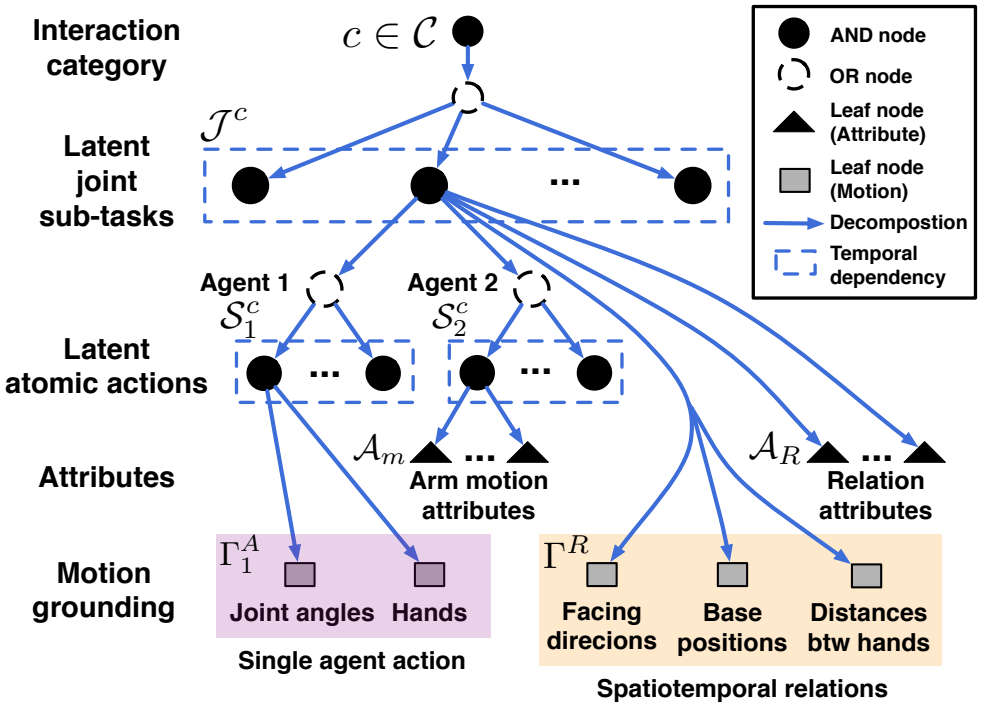


Figure 5.2: Social affordance grammar as a ST-AOG.

5.4 Representation

We represent the social affordance knowledge as stochastic context sensitive grammar using a spatiotemporal AND-OR graph (ST-AOG), as shown in Figure 5.2. The key idea is to model the joint planning of two agents on top of independent action modeling of individual agents. Following the Theory of Mind (ToM) framework, a ST-AOG defines the grammar of possible robotic actions (agent 2) at a specific moment given the observation of agent 1's actions as the belief, the joint sub-tasks as sub-goals, and the interaction category as the overall goal.

We first define a few dictionaries for the grammar model encoding the key elements in the social affordances. We constrain the human-robot interactions in a set of categories $\mathcal{C}$. Dictionaries of arm motion attributes $\mathcal{A}_M$ and relation attributes $\mathcal{A}_R$ are specified and shared across all types of interactions. Also, for each category c, there are dictionaries of latent joint sub-tasks $\mathcal{J}^c$, latent atomic actions of agent i, $\mathcal{S}_i^c$, where $\mathcal{S}_i^c$ are shared by different joint sub-tasks within c. Note that joint sub-tasks and atomic actions are not predefined labels but rather latent symbolic concepts mined from human activity videos, which boosts the flexibility of our model and requires much

less human annotation efforts.

There are several types of nodes in our ST-AOG: An AND node defines a production rule that forms a composition of nodes; an OR node indicates stochastic switching among lower-level nodes; the motion leaf nodes show the observation of agents' motion and their spatiotemporal relations; attribute leaf nodes provide semantics for the agent motion and spatiotemporal relations, which can greatly improve the robot's behavior. In our model, we consider four arm motion attributes, i.e., *moving left/right arm*, *static left/right arm*. Inspired by prior work on social perception which has revealed that touching is one of the most critical cues that signal social interactions [STC16], we specify the relation attributes as *approaching* and *holding* between two agents' hands (possibly an object).

The edges $\mathcal{E}$ in the graph represent decomposition relations between nodes. At the top level, a given interaction category leads to a selection of joint sub-tasks as the sub-goal to achieve for the given moment. A joint sub-task further leads to the atomic action selection of two agents and can also be bundled with relation attributes. An atomic action encodes a consistent arm motion pattern, which may imply some arm motion attributes of agent 2 for the purpose of motion inference. Some of the nodes in the dashed box are connected representing the "followed by" relations between joint sub-tasks or atomic actions with certain transition probabilities.

The motion grounding is designed for motion transfer from a human to a humanoid, which entails social etiquette such as proper standing distances and body gestures. As shown in Figure 5.3, the pose of a human arm at time t can be conveniently mapped to a robot arm by four degrees: $\theta^t = \langle s_0, s_1, e_0, e_1 \rangle$. The wrist angles are not considered due to the unreliable hand gesture estimation from Kinect. Thus, in an interaction whose length is T, there is a sequence of joint angles, i.e., $\Theta_{il} = \{\theta_{il}^t\}_{t=1,\cdots,T}$ for agent i's limb l, where $l = 1$ stands for left arm and $l = 2$ indicates right arm. Similarly the hand trajectories $H_{il} = \{\boldsymbol{h}_{il}^t\}$ are also considered in order to have a precise control of the robot's hands. We model the spatiotemporal relations with agent 2's the relative facing directions, $O = \{o^t\}_{t=1,\cdots,T}$, and relative base positions (in the top-down view), $X = \{\boldsymbol{x}^t\}_{t=1,\cdots,T}$, by setting the facing directions and base joint positions of agent 1 as references respectively. We also consider the distances between two agents' hands, $D_{ll'} = \{\boldsymbol{d}_{ll'}^t\}_{t=1,\cdots,T}$ (l is the limb of agent 1 and l' is the limb of agent 2) for the relations. The distances between agent 2's

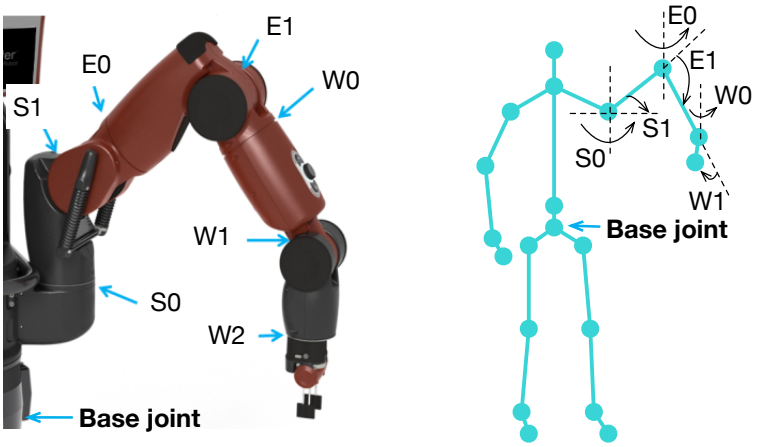


Figure 5.3: (a) The joint angles of the arm of a Baxter robot (from http://sdk.rethinkrobotics.com/wiki/Arms), which are directly mapped to a human's arm (b). The additional angles (e.g., w_2) can be either computed by inverse kinematics or set to a constant value.

hands and an object can be included if an object is involved. For an interaction instance, we then define the action grounding of agent i to be $\Gamma_i^A = \langle \Theta \rangle$, and the relation grounding of both agents to be $\Gamma^R = \langle O, X, D \rangle$, where $\Theta = \{\Theta_{il}\}_{l=1,2}$, $H = \{H_{il}\}_{l=1,2}$, and $D = \{D_{ll'}\}_{l,l' \in \{1,2\}}$. Hence, the overall motion grounding is $\Gamma = \langle \{\Gamma_i^A\}_{i=1,2}, \Gamma^R \rangle$.

Finally, the ST-AOG of interactions $\mathcal{C}$ is denoted by $\mathcal{G} = \langle \mathcal{C}, \{\mathcal{J}^c\}_{c \in \mathcal{C}}, \{\mathcal{S}_i^c\}_{c \in \mathcal{C}, i=1,2}, \mathcal{A}_M, \mathcal{A}_R, \Gamma, \mathcal{E} \rangle$. At any time t, we use a sub-graph of the ST-AOG, i.e., a parse graph $pg^t = \langle c, j^t, s_1^t, s_2^t \rangle$, to represent the actions of individual agents (s_1^t, s_2^t) as well as their joint sub-tasks (j^t) in an interaction c. Note that the attributes are implicitly included in the parse graphs since they are bundled with labels of j^t and s_2^t.

For an interaction in $[1, T]$, we may construct a sequence of parse graphs $PG = \{pg^t\}_{t=1,\cdots,T}$ to explain it, which gives us three label sequences: $J = \{j^t\}_{t=1,\cdots,T}$, $S_1 = \{s_1^t\}$ and $S_2 = \{s_2^t\}$. By merging the consecutive moments with the same label of joint sub-tasks or atomic actions, we obtain three types of temporal parsing, i.e., $\mathcal{T}^J = \{\tau_k^J\}_{k=1,\cdots,K^J}$, $\mathcal{T}_1^S = \{\tau_{1k}^S\}_{k=1,\cdots,K_1^S}$, and $\mathcal{T}_2^S = \{\tau_{2k}^S\}_{k=1,\cdots,K_2^S}$ for the joint sub-tasks and the atomic actions of two agents respectively, each of which specifies a series of consecutive time intervals where the joint sub-task or the atomic

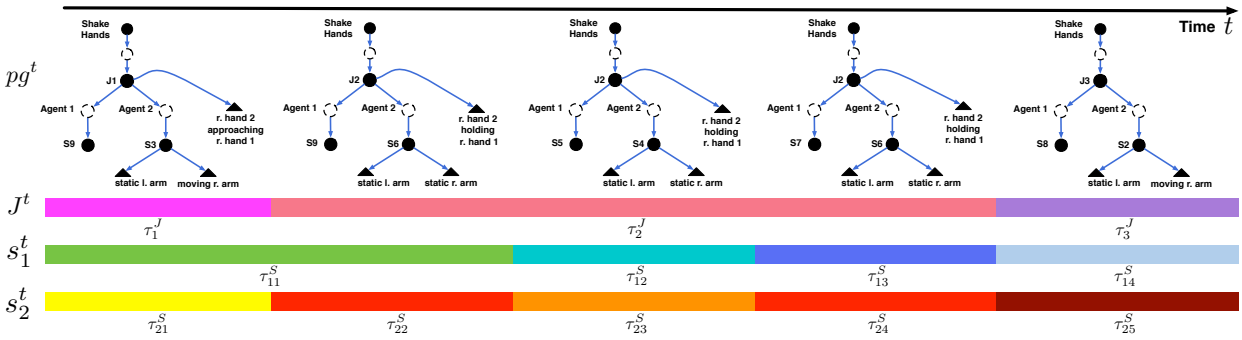


Figure 5.4: A sequence of parse graphs in a shaking hands interaction, which yields the temporal parsing of joint sub-tasks and atomic actions depicted by the colored bars (colors indicate the labels of joint sub-tasks or atomic actions).

action remains the same in each interval. Hence, in $\tau_k^J = [t_k^1, t_k^2]$, $j^t = j(\tau_k^J)$, $\forall t \in \tau_k^J$, and for agent i, $s_i^t = s_i(\tau_{ik}^S)$, $\forall t \in \tau_{ik}^S$ in $\tau_{1k}^S = [t_{ik}^1, t_{ik}^2]$. Figure 5.4 shows an example of the temporal parsing from the parse graph sequence. Note the numbers of time intervals of these three types of temporal parsing, i.e., K^J, K_1^S, and K_2^S, may be different. Such flexible temporal parsing allows us to model long-term temporal dependencies among atomic actions and joint sub-tasks.

5.5 Probabilistic Model

We propose a probabilistic model for our social affordance grammar model.

Given the motion grounding, Γ, the posterior probability of a parse graph sequence PG is defined as

$$p(PG|\Gamma) \propto \underbrace{p(\{\Gamma_i^A\}_{i=1,2}|PG)}_{\text{arm motion likelihood}} \underbrace{p(\Gamma^R|PG)}_{\text{relation likelihood}} \underbrace{p(PG)}_{\text{parsing prior}}. \tag{5.1}$$

Conditioned on the temporal parsing of atomic actions and joint sub-tasks, the likelihood terms model the arm motion and the relations respectively, whereas the parsing prior models the temporal dependencies and the concurrency among joint sub-tasks and atomic actions. We introduce these three terms in the following subsections.

5.5.1 Arm Motion Likelihood

First, we define three types of basic potentials that are repeatedly used in the likelihood terms:

1) **Orientation potential** $\psi_o(\theta)$. This potential is a von Mises distribution of the orientation variable θ. If θ has multiple angular variables, e.g., the four joint angles $\theta = \langle s_0, s_1, e_0, e_1 \rangle$, then the potential is the product of the von Mises distributions of these individual angular variables.

2) **Three-dimensional motion potential** $\psi_{3v}(\boldsymbol{x})$. Assuming that spherical coordinate of $\boldsymbol{x}$ is (r, θ, ϕ), the potential is characterized by three distributions, i.e., $\psi_{3v}(\boldsymbol{x}) = p(r)p(\theta)p(\phi)$, where the first one is a Weibull distribution and the remaining are von Mises distributions.

3) **Two-dimensional position potential** $\psi_{2v}(\boldsymbol{x})$. We fit a bivariate Gaussian distribution for $\boldsymbol{x}$ in this potential.

For joint angles and hand positions in an atomic action, we are interested in their final statuses and change during the atomic action. Thus, for the limb l of agent i in the interval τ_{ik}^S assigned with atomic action $s_i(\tau_{ik}^S) \in \mathcal{S}^\rfloor$ such that $s_i^t = s_i(\tau_{ik}^S), \forall t \in \tau_{ik}^S$, the arm motion likelihood

$$p(\Theta_{il}, H_{il}|\tau_{ik}^S, s_i(\tau_{ik}^S)) \propto \underbrace{\psi_o(\theta_{il}^{t'} - \theta_{il}^t)}_{\text{joint angles's change}} \underbrace{\psi_o(\theta_{il}^{t'})}_{\text{final joint angles}} \underbrace{\psi_{3v}(\boldsymbol{h}_{il}^{t'} - \boldsymbol{h}_{il}^t)}_{\text{hand movement}} \underbrace{\psi_{3v}(\boldsymbol{h}_{il}^{t'})}_{\text{final hand position}}, \tag{5.2}$$

where $t = t_{ik}^1$ and $t' = t_{ik}^2$ are the starting and ending moments of τ_{ik}^S. Assuming independence between the arms, the arm motion likelihood for agent i in τ_{ik}^S is

$$p(\Gamma_i^A|\tau_{ik}^S, s(\tau_{ik}^S)) = \prod_l p(\Theta_{il}, H_{il}|\tau_{ik}^S, s_i(\tau_{ik}^S)), \tag{5.3}$$

and the arm motion likelihood for the entire interaction is

$$p(\Gamma_i^A|PG) = \prod_k p(\Gamma_i^A|\tau_{ik}^S, s(\tau_{ik}^S)). \tag{5.4}$$

Finally, the overall arm motion likelihood is the product of two agents' arm motion likelihood, i.e.,

$$p(\{\Gamma_i^A\}_{i=1,2}|PG) = \prod_i p(\Gamma_i^A|PG). \tag{5.5}$$

5.5.2 Relation Likelihood

Relation likelihood models the spatiotemporal patterns hidden in facing directions O, base positions X, and the distances between two agents' hands during a joint sub-task. In a interval τ_k^J with the same joint sub-task label $j(\tau_k^J)$ such that $j^t = j(\tau_k^J), \forall t \in \tau_k^J$, the relation likelihood is

$$p(\Gamma^R|\tau_k^J, j(\tau_k^J)) \propto \underbrace{\psi_o(o^{t'})}_{\text{facing direction}} \underbrace{\psi_{2v}(\boldsymbol{x}^{t'})}_{\text{base position}} \cdot \prod_{l,l'} \underbrace{\psi_{3v}(d_{ll'}^{t'})}_{\text{final hand distance}} \underbrace{\psi_{3v}(d_{ll'}^{t'} - d_{ll'}^{t})}_{\text{distance change}}, \tag{5.6}$$

where τ_k^J starts at $t = t_k^1$ and ends at $t' = t_k^2$.

Hence, the overall relation likelihood can be written as

$$p(\Gamma^R|PG) = \prod_k p(\Gamma^R|\tau_k^J, j(\tau_k^J)). \tag{5.7}$$

5.5.3 Parsing Prior

The prior of a sequence of parse graphs is defined by the following terms:

$$\begin{aligned} p(PG) = & \underbrace{\prod_k p\left(|\tau_k^J| \mid j(\tau_k^J)\right)}_{\text{duration prior of joint sub-tasks}} \\ & \underbrace{\cdot \prod_k p\left(|\tau_{1k}^S| \mid s_1(\tau_{1k}^S)\right) \prod_k p\left(|\tau_{2k}^S| \mid s_2(\tau_{2k}^S)\right)}_{\text{duration prior of atomic actions}} \\ & \underbrace{\prod_{k>1} p\left(s_1(\tau_{1k}^S)|s(\tau_{1k-1}^S)\right)}_{\text{action transition for agent 1}} \underbrace{\prod_{k>1} p\left(s_2(\tau_{2k}^S)|s(\tau_{2k-1}^S)\right)}_{\text{action transition for agent 2}} \\ & \underbrace{\cdot \prod_t p(s_1^t|j^t)p(s_2^j|j^t)}_{\text{concurrency}} \underbrace{\prod_{k>1} p\left(j(\tau_k^J)|j(\tau_{k-1}^J)\right)}_{\text{joint sub-task transition}}, \end{aligned} \tag{5.8}$$

where the duration priors follow log-normal distributions and the remaining priors follow multinomial distributions.

5.6 Learning

The proposed ST-AOG can be learned in a weakly supervised manner, where we only specify the generic dictionaries of attributes and the sizes of the dictionaries of joint sub-tasks and atomic

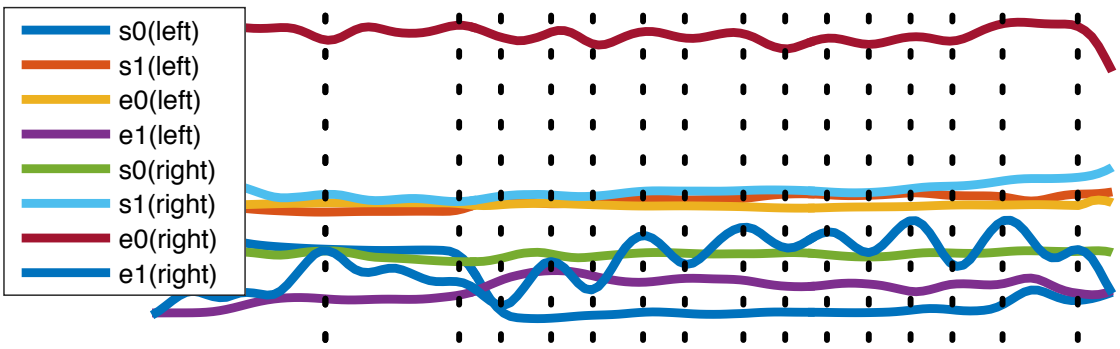


Figure 5.5: The curves show how the joint angles of agent 2's two arms change in an shaking hands interaction. The black dashed indicate the interval proposals from the detected turning points.

actions for each interaction. Given N training instances, $\Gamma = \{\Gamma_n\}_{n=1,\cdots,N}$, of an interaction category, where $\Gamma_n = \langle \{\Gamma_i^A\}_{i=1,2}, \Gamma_i^R \rangle$ is the motion grounding of instance n, the goal of learning is to find the optimal parsing graph sequence, PG_i, for each instance by maximizing the posterior probability defined in Eq. (5.1); then the ST-AOG is easily constructed based on the parse graphs.

It is intractable to search for the optimal parsing of atomic actions and joint sub-tasks simultaneously, which will take an exponential amount of time. Instead, we first 1) parse atomic actions for each agent independently and then 2) parse joint sub-tasks. Based on the likelihood distributions from the parsing results, we may 3) further obtain the implied attributes for each type of joint sub-tasks and atomic actions. We introduce the details in the rest of this section.

5.6.1 Atomic Action Parsing

We expect the motion in an atomic action to be consistent. Since the arm motion is characterized by joint angles and hand positions, the velocities of joints and hand movements should remain the same in an atomic action. Following this intuition, we propose the time intervals for the atomic actions of an agent by detecting the turning points of the sequences of joint angles (see Figure 5.5), which will naturally yields time intervals of atomic actions. To make the angles directly comparable, they are all normalized to the range of $[0, 1]$.

To detect such turning points, we introduce a entropy function for a sequence $\{x^t\}$, i.e., $\mathcal{E}(t, w)$, where t is the location of interest and w is the window size. To compute $\mathcal{E}(t, w)$, we first count

the histogram of the changes between consecutive elements, i.e., $x^t - x^{t-1}$ in the sub-sequence $\{x^{t'}\}_{t'=t-w,,t+w}$, and then $\mathcal{E}(t,w)$ is set to be the entropy of the histogram. By sliding windows with different sizes ($w = 2, 5, 10, 15$), we may detect multiple locations with entropy that is higher than a given threshold. By non-maximum suppression, the turning points are robustly detected.

After obtaining the time intervals, we assign optimal atomic action labels to each interval by Gibbs sampling. At each iteration, we choose an interval τ and sample a new label s for it based on the following probability:

$$s \sim p(\Gamma_i^A \mid \tau, s) p(\tau, s \mid \mathcal{T}_i^S \backslash \tau, S_i \backslash \{s_i^t\}_{t \in \tau}). \tag{5.9}$$

Here, $p(\Gamma_i^A \mid \tau, s)$ is the likelihood in Eq. (5.3), and based on the parsing prior in Eq. (5.8), the labeling prior is computed as

$$p(\tau, s \mid \mathcal{T}_i^S \backslash \tau, S_i \backslash \{s_i^t\}_{t \in \tau}) = p(s \mid s') p(s'' \mid s) p(|\tau| \mid s), \tag{5.10}$$

where s' and s'' are the preceding and following atomic action labels in the adjacent intervals of τ. If either of them is absent, the corresponding probability is then set to be 1. For each new label assignment, the parameters of the related likelihood and prior distributions should be re-estimated. To ensure the distinctness between adjacent intervals, s can not be the same labels of the adjacent intervals.

Therefore, after randomly assigning labels for the intervals as initialization, we conduct multiple sweeps, where in each sweep, we enumerate each interval and sample a new label for it based on Eq. (5.9). The sampling stops when the labeling does not change after the last sweep (convergence). In practice, the sampling can converge within 100 sweeps coupled with a simulated annealing.

5.6.2 Joint Sub-Task Parsing

The joint sub-task parsing is achieved using a similar approach as atomic action parsing. We first propose the time intervals by detecting turning points based on the normalized sequences of O, X, and D. Then the labeling can also be optimized by a Gibbs sampling, where at each iteration, we

sample a new joint sub-task label j for an interval τ by

$$j \sim p(\Gamma^R \mid \tau, j)p(\tau, j \mid \mathcal{T}^J \backslash \tau, S^J \backslash \{j^t\}_{t\in\tau}), \tag{5.11}$$

where $p(\Gamma^R \mid \tau, j)$ is defined in Eq. (5.6) and the prior probability is derived from Eq. (5.8) as

$$p(\tau, j \mid \mathcal{T}^J \backslash \tau, S^J \backslash \{j^t\}_{t\in\tau}) = p(j \mid j')p(j'' \mid j)p(|\tau| \mid j) \prod_{t\in\tau} p(s_1^t \mid j)p(s_2^t \mid j). \tag{5.12}$$

Similar to Eq. (5.9), j' and j'' in the above prior probability are the preceding and following intervals' joint sub-task labels. The corresponding transition probability is assumed to be 1 if either of the adjacent interval does not exist. We also constrain j to be different from the j' and j'' if they exist.

5.6.3 Constructing ST-AOG

After the previous two Gibbs sampling processes, the parameters of our probabilistic model are all estimated based on the parse graph sequences $\{PG_n\}_{n=1,\cdots,N}$. The ST-AOG of category c is then constructed by the following three steps:

Initialization. We start form a "complete" graph, where each non-leaf node is connected to all related lower level nodes (e.g., all joint sub-tasks, all atomic actions of the corresponding agent, etc.), except attribute leaf nodes.

Edge removal. Any edge between two joint sub-task nodes or two atomic action nodes is removed if it has a transition probability lower than a threshold (0.05). For each joint sub-task node, remove the edges connecting the OR node of agent i to the atomic actions whose concurrency priors under the joint sub-task are lower than 0.1. Note that we use these thresholds for all interactions.

Attributes bundling. Motion attributes: For each type of atomic action s of agent i, a *moving* attribute is bundled to a limb if the mean of the corresponding hand movement distribution specified in Eq. (5.2) is lower than a threshold (we use 0.2 m in practice); otherwise, a *static* attribute is bundled to the limb instead. Relation attributes: A type of joint sub-task will be associated with a *holding* attribute between a pair of hands (or a hand and an object) if the mean final hand distance

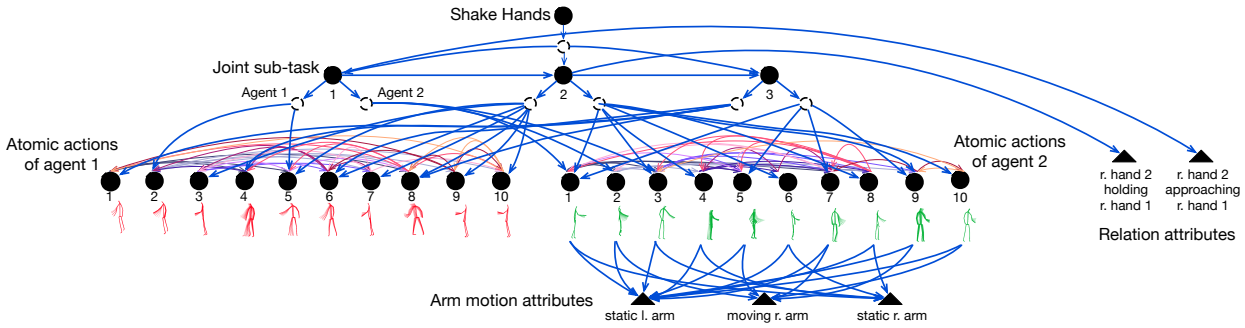


Figure 5.6: The learned ST-AOG for the *Shake Hands* interaction (the motion grounding is not drawn in this figure due to the space limit). The numbers under AND nodes are the labels of joint sub-tasks or atomic actions. The edges between the atomic actions show the "followed by" temporal relations and their colors indicate which atomic actions are the edges' starting point. Similarly, the joint sub-tasks are also connected by edges representing the temporal dependencies between them. There is an example of each atomic actions from our training data, where the skeletons are overlaid with colors from light to dark to reflect the temporal order. The attributes that are not bundled to any atomic action or joint sub-task are not shown here.

is lower than 0.15 m and the mean hand distance's change is lower than 0.1 m according to the corresponding distributions in Eq. (5.6). If only the mean final hand distance meets the standard, an *approaching* will be attached. For the case of multiple qualifying pairs for a hand, the one with the shortest mean distance is selected.

Figure 5.6 is a learned ST-AOG for *Shake Hands* interactions. It can be seen that our learning algorithm indeed mines the critical elements of the interactions and clearly represents their relations through the structure of the ST-AOG.

5.7 Real-time Motion Inference

If we replace agent 2 with a humanoid, we can therefore design a real-time motion inference enabling human-robot interaction based on the learned ST-AOG by sampling parse graphs and controlling the robot's motion accordingly.

For this, we propose two levels of inference procedures: 1) robot motion generation given

the parse graphs, which is essentially transferring the socially appropriate motion from agent 2 in the grammar model to a humanoid; 2) parse graph sampling given the observation of the human agent's actions and the relation between the human agent and the robot according to the learned social affordance grammar.

5.7.1 Robot Motion Generation

As shown in Figure 5.3, we may use the motion grounding of agent 2 for the robot by joint mapping. The robot motion can be generated by sampling agent 2's base position $\boldsymbol{x}^t$, facing direction (i.e., base orientation of the robot) o^t, joint angles $\{\theta_{2l}\}_{l=1,2}$, and hand positions (i.e., end effector positions) $\{\boldsymbol{h}^t_{2l}\}_{l=1,2}$ at each time t based on the motion history of agent 2, $\Gamma^A_2(t-1)$, and the spatiotemporal relations, $\Gamma^R(t-1)$, upon $t-1$ as well as the agent 1's motion, $\Gamma^A_1(t)$, and parse graphs, $PG(t) = \{pg^\tau\}_{\tau=1,\cdots,t}$, upon t.

Since the arm motion is relative to the base position in our motion grounding, we first sample $\boldsymbol{x}^t$ and o^t w.r.t. the relative position and facing direction likelihood in Eq. (5.6), the likelihood probabilities of which must be higher than a threshold (0.05 for $\boldsymbol{x}^t$ and 0.3 for o^t). To avoid jitter, we remain the previous base position and rotation if they still meet the criteria at t.

Then we update the joint angles for each robot arm. Without the loss of generality, let us consider a single arm $l \in \{1, 2\}$. According to the atomic action s^t_2, we may sample desired joint angles $\hat{\theta}^t_{2l}$ and hand position $\hat{\boldsymbol{h}}^t_{2l}$ w.r.t the corresponding likelihood terms in Eq. (5.2). Since we do not model the wrist orientations, the desired $\hat{w}_0$, $\hat{w}_1$, $\hat{w}_2$ are always set to be 0 if the robot arm has these degrees of freedom (Figure 5.3a). If current joint sub-task entails an "approaching" or "holding" attribute for this limb, the desired hand position is set to the position of the target hand or object indicated by the attribute instead. To enforce the mechanical limits and collision avoidance, we minimize a loss function to compute the final joint angels θ^t_{il} for the robot arm:

$$\min_{\theta\in\Omega_\theta} \underbrace{\omega_h||\boldsymbol{f}_l(\theta) - \hat{\boldsymbol{h}}^t_{2l}||^2_2}_{\text{hand position loss}} + \underbrace{\omega_a||\theta - \hat{\theta}^t_{2l}||^2_2}_{\text{joint angle loss}} + \underbrace{\omega_s||\theta - \theta^{t-1}_{il}||^2_2}_{\text{smoothness loss}}, \tag{5.13}$$

where $\boldsymbol{f}_l(\theta)$ is the end effector position of θ based on the forward kinematics of the robot arm l; Ω_θ is the joint angle space that follows the mechanical design (angle ranges and speed limits of

Algorithm 3 Parse Graph Sampling

Input: The initial motion of two agents in $[1, T_0]$, i.e., $\Gamma(T_0)$

Infer $PG(T_0)$ by maximizing the posterior probability in Eq. (5.1)

Let $t \leftarrow T_0 + 1$

repeat

  $\Gamma' = \Gamma(t-1) \cup \{\theta_{1l}^t\}_{l=1,2} \cup \{h_{1l}^t\}_{l=1,2}$

  Infer current atomic action of agent 1 by

  $s_1^t = \text{argmax}_s\, p(PG(t-1) \cup \{s\} \mid \Gamma')$

  for all $j^t \in \mathcal{J}, s_2^t \in \mathcal{S}_2^c$ that are compatible with s_1^t **do**

    $pg^t \leftarrow \langle j^t, s_1^t, s_2^t \rangle$

    $PG(t) \leftarrow PG(t-1) \cup \{pg^t\}$

    Sample a new robot status at t, i.e., $\boldsymbol{x}^t$, o^t, $\{\theta_{2l}^t\}$ and $\{\boldsymbol{h}_{2l}^t\}$, as introduced in Section 5.7.1

    $\Gamma(t) \leftarrow \Gamma(t-1) \cup \{\theta_{il}^t, \boldsymbol{h}_{il}^t\}_{i,l=1,2} \cup \{\boldsymbol{x}^t\} \cup \{o^t\}$

    Compute the posterior probability $p(PG(t) \mid \Gamma(t))$

  end for

  Choose the pg^t and the corresponding new robot status that yield highest posterior probability to execute and update $PG(t)$ and $\Gamma(t)$ accordingly

  $t \leftarrow t + 1$

until $t > T$

arm joints) and the collision avoidance constraints, and ω_h, ω_a, ω_s are weights for the three types of loss respectively. By assigning different weights, we can design three control modes that are directly related to the attributes in ST-AOG:

1) **Hand moving mode**: if "approaching" or "holding" attributes are present in the current joint sub-task, we may use a larger ω_h to ensure an accurate hand position;

2) **Static mode**: if the first case does not hold and the atomic action has a "static" attribute for the limb, then ω_s should be much larger than ω_h and ω_a;

3) **Motion mimicking mode**: if none of the above two cases hold, we emphasize on joint angle loss (i.e., a large ω_a) to mimic the human arm motion.

In practice, we set the large weight to be 1 and the other two may range from 0 to 0.1.

Table 5.1: A summary of our new dataset (numbers of instances).

Category	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Total
Shake Hands	19	10	0	0	29
High Five	18	7	0	23	48
Pull Up	21	16	9	0	46
Wave Hands	0	28	0	18	46
Hand Over	34	6	8	7	55

5.7.2 Parse Graph Sampling

The Parse graph sampling algorithm is sketched in Alg. 3. The basic idea is to first recognize the action of agent 1. Then following the ST-AOG, we may enumerate all possible joint sub-tasks and atomic actions of agent 2 that are compatible with agent 1's atomic action, and sample a new robot status for each of them. Finally, we choose the one with the highest posterior probability to execute. Note that the facing direction of an agent is approximated by his or her moving direction (if not static) or the pointing direction of feet (if static).

5.8 Experiments

Dataset. There are two existing RGB-D video datasets for human-human interactions [YHC12, SRZ16], where the instances within the same category are very similar. To enrich the activities, we collected and compiled a new RGB-D video dataset, UCLA Human-Human-Object Interaction (HHOI) dataset V2, on top of [SRZ16] using Kinect v2 as summarized in Table 5.1, where *Wave Hands* is a new category and the instances in scenario 1 of the other categories are from [SRZ16]. For *Pull Up*, the first 3 scenarios are: A2 (agent 2) stands while A1 (agent 1) is sitting 1) on the floor or 2) in a chair; 3) A1 sits in a chair and A2 approaches. For the other categories, the four scenarios stand for: 1) both stand; 2) A1 stands and A2 approaches; 3) A1 sits and A2 stands nearby; 4) A1 sits and A2 approaches. In the experiments, we only use three fourths of the videos

in scenario 1 (for *Wave Hands*, it is scenario 2) as training data, and the remaining instances are used for testing. The dataset has been released at https://tshu.io/SocialAffordanceGrammar.

Baselines. We compare our approach with two baselines adopted from related methods, extending these method further to handle our problem. The first one (B1) uses the method proposed in [SRZ16] to synthesize human skeletons to interact with the given human agent, from which we compute the desired base positions, joint angles and hand positions for the optimization method defined in Eq. (5.13). Since [SRZ16] only models the end positions of the limbs explicitly and do not specify multiple modes as ours do, we use it with the weights of hand moving mode. The second baseline (B2) uses our base positions and orientations but solve the inverse kinematics for the two arms using an off-the-shelf planner, i.e., RRT-connect [KL00] in MoveIt! based on the desired hand positions from our approach.

5.8.1 Experiment 1: Baxter Simulation

We first implement a Baxter simulation and compare the simulated robot behaviors generated from ours and the two baselines. For each testing instance, we give the first two frames of skeletons of two agents as the initialization; we then update the human skeleton and infer the new robot status accordingly at a rate of 5 fps in real-time. For *Hand Over*, we assume that the cup will stay in the human agent's hand unless the robot hand is close to the center of the cup (< 10 cm) for at least 0.4 s. Note that the planner in B2 is extremely slow (it may take more than 10 s to obtain a new plan), so we compute B2's simulations in an offline fashion and visualize them at 5 fps. Ours and B1 can be run in real-time.

Figure 5.7 shows a simulation example for each interaction. More results are included in the video attachment. From the simulation results, we can see that the robot behaviors (standing positions, facing directions and arm gestures) generated by ours are more realistic than the ones from baselines. Also, thanks to the learned social grammar, the robot can adapt itself to unseen situations. E.g., human agents are standing in the training data for "High Five", but the robot can still perform the interaction well when the human agent is sitting.

We also compare the mean joint angle difference between the robot and the ground truth (GT)

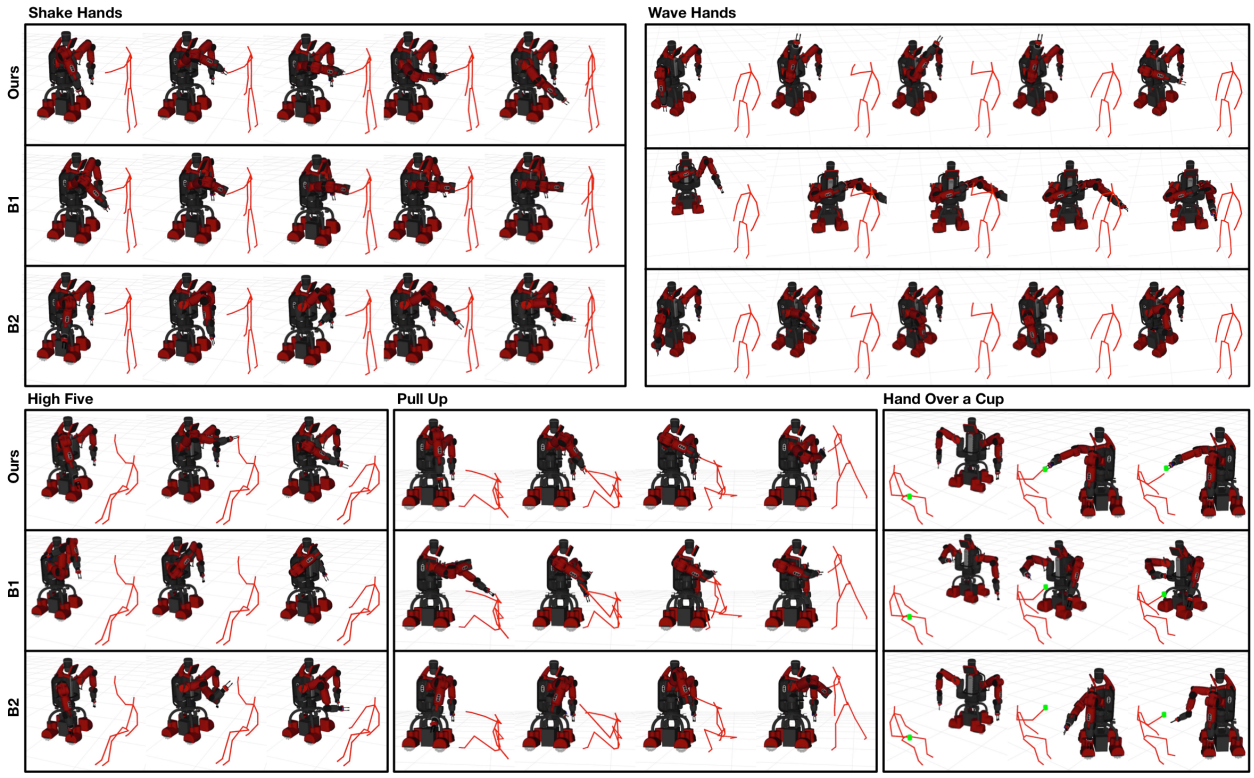


Figure 5.7: Qualitative results of our Baxter simulation.

Table 5.2: Mean joint angle difference (in radius degree) between the simulated Baxter and the ground truth skeletons.

Method	Shake Hands	High Five	Pull Up	Wave Hands	Hand Over
B1	0.939	0.832	0.748	0.866	0.867
B2	0.970	0.892	0.939	0.930	0.948
Ours	**0.779**	**0.739**	**0.678**	**0.551**	**0.727**

human skeletons (i.e., agent 2) captured from Kinect as reported in Table 5.2, which is one of the two common metrics of motion similarity [ARA17] (the other one, i.e., comparing the end-effector positions, is not suitable in our case since humans and robots have different arm lengths). Although the robot has a different structure than humans', ours can still generate arm gestures that are significantly closer to the GT skeletons than the ones by baselines are.

Table 5.3: Human subjects' ratings of Baxter simulation generated by the three methods based on the two criteria.

	Source	Shake Hands	High Five	Pull Up	Wave Hands	Hand Over
Q1	B1	3.22 ± 1.30	2.13 ± 1.09	2.75 ± 0.91	2.59 ± 1.20	2.19 ± 1.12
	B2	2.14 ± 0.56	3.07 ± 1.22	2.11 ± 0.94	2.47 ± 0.69	1.48 ± 0.52
	Ours	$\mathbf{4.45} \pm 0.61$	$\mathbf{4.79} \pm 0.41$	$\mathbf{4.53} \pm 0.61$	$\mathbf{4.82} \pm 0.52$	$\mathbf{4.63} \pm 0.53$
Q2	B1	2.89 ± 0.99	2.38 ± 0.96	2.75 ± 0.55	2.00 ± 1.17	2.45 ± 0.71
	B2	2.14 ± 0.83	2.93 ± 0.80	2.32 ± 1.00	1.60 ± 0.69	1.82 ± 0.63
	Ours	$\mathbf{4.20} \pm 0.75$	$\mathbf{4.17} \pm 0.62$	$\mathbf{4.25} \pm 0.79$	$\mathbf{4.65} \pm 0.72$	$\mathbf{3.97} \pm 0.61$

5.8.2 Experiment 2: Human Evaluation

To evaluate the quality of our human-robot interactions, we showed the simulation videos of three methods to 12 human subjects (UCLA students) who did not know that videos were from different methods. Subjects first watched two RGB videos of human interactions per category. Then for each testing instance, we randomly selected one method's simulation to a subject. The subjects only watched the assigned videos once and rated them based on two criteria: i) whether the purpose of the interaction is achieved (Q1), and ii) whether the robot's behavior looks natural (Q2). The ratings range from 1 (total failure/awkward) to 5 (successful/human-like).

The mean ratings and the standard deviations are summarized in Table 5.3. Our approach outperforms the baselines for both criteria and has smaller standard deviations, which manifests its advantages on accurately achieving critical latent goals (e.g., holding hands) while keeping human-like motion. The rigid representation and failing to learn explicit hand relations affect B1's ability to adapt the robot to various scenarios. It also appears that only using a simple IK (B2) is probably insufficient: its optimization is only based on the current target position, which often generate a very long path and may lead to an awkward gesture. This makes the future target positions hard to reach as the target (e.g., a human hand) is constantly moving.

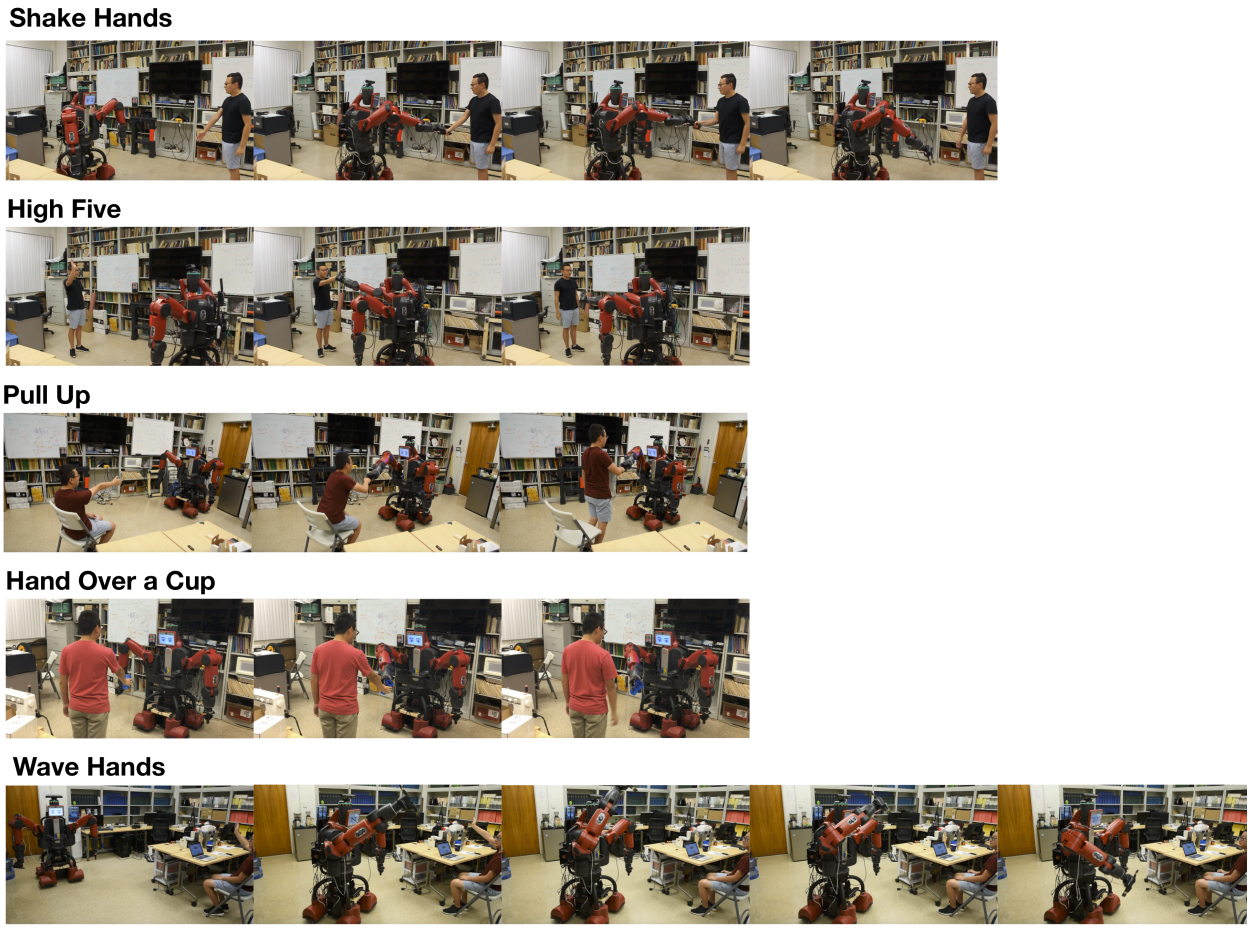


Figure 5.8: Qualitative results of the real Baxter test.

5.8.3 Experiment 3: Real Baxter Test

We test our approach on a Baxter research robot with a mobility base (Figure 5.8). A Kinect sensor is mounted on the top of the Baxter's head to detect and track human skeletons. To compensate the noise from Kinect, we further take advantage of the pressure sensors on the ReFlex TakkTile Hand (our Baxter's right hand) to detect holding relations between the agents' hands. Although the arm movement is notably slower than the simulation due to the mechanical limits, the interactions are generally successful and reasonably natural.

Since we only need joints on the upper body, the estimation of which is relatively reliable, the noisy Kinect skeletons usually do not greatly affect the control. In practice, temporal smoothing of the skeleton sequences is also helpful.

5.9 Conclusion

We propose a general framework of learning social affordance grammar as a ST-AOG from human interaction videos and transferring such knowledge to human-robot interactions in unseen scenarios by a real-time motion inference based on the learned grammar. The experimental results demonstrate the effectiveness of our approach and its advantages over baselines. In the future, it is possible to integrate a language model into the system to achieve verbal communications between robots and humans. In addition, human intention inference can also be added to the system.

CHAPTER 6

Perception of Human Interaction in Decontextualized Animations

6.1 Introduction

People are adept at perceiving goal-directed action and inferring social interaction from movements of simple objects. In their pioneering work, [HS44] presented video clips showing three simple geometrical shapes moving around, and asked human observers to describe what they saw. Almost all observers described the object movements in an anthropomorphic way, reporting a reliable impression of animacy and meaningful social interactions among the geometric shapes displayed in the decontextualized animation. Their results were replicated in other studies using similar videos for both human adults [OY85, RBL85] and preschoolers as young as five years old [SMB96].

To study what visual information drives the perception of interaction, [BMK92] generated new Heider-Simmel animations with either the structural aspect or dynamic aspect disrupted. They found that the motion patterns mainly determined the anthropomorphic description of videos. Later studies [DL94, ST00, TF00, TF06, GNS09, GMS10] used more controlled stimuli and systematically examined what factors can impact the perception of goal-directed actions in a decontextualized animation. These findings provided converging evidence that the perception of human-like interactions relies on some critical low-level motion cues, such as speed and motion direction. However, it remains unclear how the human visual system combines motion cues from different objects to infer interpersonal interactivity in the absence of any context cues.

To address this fundamental question, [BST09] developed a Bayesian model to reason about the intentions of an agent when moving in maze-like environments of the sort used by [HS44]. Other

studies [BGT08, UBM10, BST11, Bak12, SK12] developed similar models that could be generalized to situations with multiple agents and different contexts. These modeling studies illustrate the potential fruitfulness of using a Bayesian approach as a principled framework for modeling human interaction shown in decontextualized animations. However, these models have been limited to experimenter-defined movements, and by computational constraints imposed by the modelers for particular application domains.

In daily life, humans rarely observe Heider-Simmel-type animations. Although examining inferences about human interactions in videos of daily-life activities would be ecologically natural, challenges arise. Human interactions are usually accompanied by rich context information, such as language, body gestures, moving trajectories of multiple agents, and backgrounds in the environment. Hence, the complexity of information may make it difficult to pin down what critical characteristics in the input determine human judgments.

To address this problem, we used aerial video and employed advanced computer vision algorithms to generate experimental stimuli that were rigorously controlled but rooted in real-life situations. As an example, imagine that you are watching a surveillance video recorded by a drone from a bird's eye view, as shown in Fig. 6.1. In such aerial videos, changes in human body postures can barely be seen, and the primary visual cues are the noisy movement trajectories of each person in the scene. This situation is analogous to the experimental stimuli used in Heider and Simmel animations, but the trajectories of each entity are directly based on real-life human movements. Another advantage of using aerial videos is that they provide a window to examine whether a model trained with real-life motions can generalize its learned knowledge to interpret decontextualized movements of geometric shapes, without prior exposures. Such generalizability emulates humans' irresistible and automatic impressions when viewing the Heider-Simmel animations for the first time. If the generalization is successful, the cues used by the model in learning can shed light on the mechanisms underlying the human ability to recover the causal and social structure of the world from the visual inputs.

In the present study, we aimed to use real-life aerial videos to generate Heider-Simmel-type decontextualized animations and to assess how human judgments of interactivity emerge over time. We employed decontextualized animations generated from the aerial videos to measure how well

humans make online judgments about interpersonal interactions, and to gauge what visual cues determine the dynamic changes in human judgments. To account for human performance, we developed a hierarchical model with hidden layers. The model aimed to learn the representations of critical movement patterns that signal potential interactivity between agents. Furthermore, we assessed whether the learning component in the model can be generalized to the original animations used by [HS44].

Figure 6.1: Stimulus illustration. (Left) An example frame of an aerial video recorded by a drone. Two people were being tracked (framed by red and green boxes). (Right) A sample frame of an experimental trial. The two people being tracked in the aerial video are presented as two dots, one in red and one in green, against a black background. A video demonstration can be viewed at https://tshu.io/HeiderSimmel/CogSci17.

6.2 Computational Model

We designed a hierarchical model with three layers. As shown in Fig. 6.2, the first layer (the *X* layer) estimates spatiotemporal motion patterns within a short period of time. The second layer (the *S* layer) captures the involvement of various motion fields at different stages of interactivity over a long period by temporally decomposing interactivity into multiple latent sub-interactions. The last layer (the *Y* layer) indicates the presence or absence of interactivity between two agents.

The inputs to the model are motion trajectories of two agents, denoted as $\Gamma_a = \{\mathbf{x}_a^t\}_{t=0,\cdots,T}$, $a = 1, 2$. The position of agent a ($a = 1, 2$) at time t is $\mathbf{x}_a^t = (x, y)$. The total length of the trajec-

tory is T. Using the input of motion trajectories, we can readily compute the velocity sequence of agent a ($a = 1, 2$), i.e., $V_a = \{\mathbf{v}_a^t\}_{t=1,\cdots,T}$, where $\mathbf{v}_a^t = \mathbf{x}_a^t - \mathbf{x}_a^{t-1}$.

To capture the interactivity between two agents based on the observed trajectories of movements, the model builds on two basic components. (1) Interactivity between two agents can be represented by a sequence of latent motion fields, each capturing the relative motion between the two agents who perform meaningful social interactions. (2) Latent motion fields can vary over time, capturing the behavioral change of the agents over a long period of time. The details for quantifying the two key components are presented in the next two subsections.

6.2.1 Conditional Interactive Fields

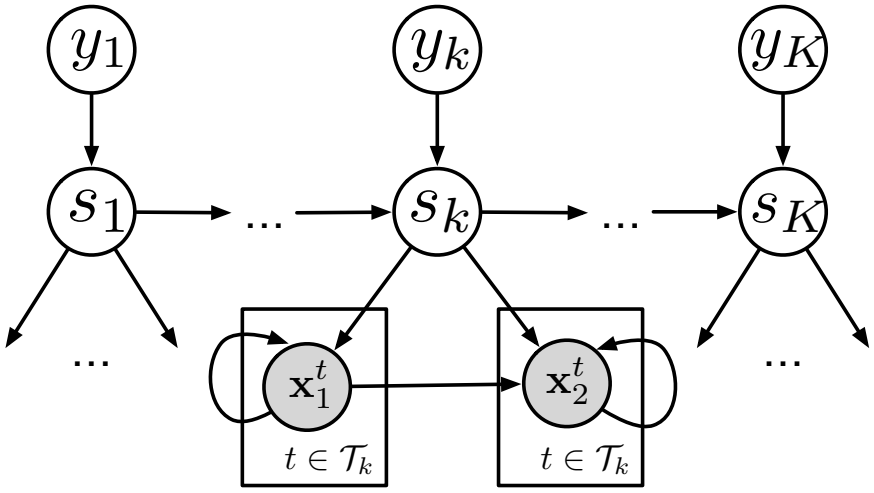


Figure 6.2: Illustration of the hierarchical generative model. The solid nodes are observations of motion trajectories of two agents, and the remaining nodes are latent variables constituting the symbolic representation of an interaction, i.e., the original trajectories are coded as a sequence of sub-interactions S and interaction labels Y.

As illustrated in Fig. 6.3, we use conditional interactive fields (CIFs) to represent how an agent moves with respect to a reference agent. This is analogous to the force fields in Physics, where the objects interact with each other through invisible fields (e.g., gravity). To derive the CIFs, we randomly select an agent to be the reference agent, and then model the partner agent's movement by estimating a vector field of the relative motion conditioned on a specific distribution of the

reference agent's motion.

To ensure that the fields are orientation invariant, we perform a coordinate transformation, as Fig. 6.3 illustrates. At each time point t, the transformed position of the reference agent is always located at $(0, 0)$, and its transformed velocity direction is always pointed to the norm of the upward vertical direction. Consequently, the position and velocity of the second agent after the transformation, i.e., $\tilde{\Gamma} = \{\tilde{\mathbf{x}}^t\}_{t=0,\cdots,T}$ and $\tilde{V} = \{\tilde{\mathbf{v}}^t\}_{t=1,\cdots,T}$, can be used to model the relative motion.

A sub-interaction s corresponds to interactivity in a relatively short time sharing consistent motion patterns, e.g., approaching, walking together, standing together. The model can infer its CIF using a potential function $U(\tilde{\mathbf{x}}^t, \tilde{\mathbf{v}}^t, \mathbf{v}_1^t)$, where the first two variables $(\tilde{\mathbf{x}}^t, \tilde{\mathbf{v}}^t)$ are used to model the relative motion as defined in the last paragraph and $\mathbf{v}_1^t$ is the reference agent's motion. The potential function is defined to yield the lowest potential value if the motion pattern fits the characteristics of s the best. In this way, the model considers the agents more likely to be interactive if the agents are moving in a specific way that can minimize the potential energy w.r.t. certain potential fields.

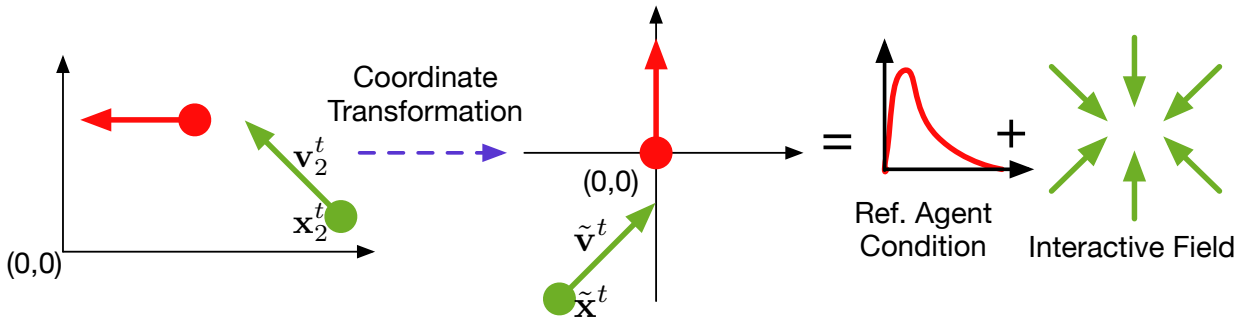


Figure 6.3: Illustration of a conditional interactive field (CIF): after a coordinate transformation w.r.t. the reference agent, we model the expected relative motion pattern $\tilde{\mathbf{x}}^t$ and $\tilde{\mathbf{v}}^t$ conditioned on the reference agent's motion.

6.2.2 Temporal Parsing by Latent Sub-Interactions

We assume that a long interactive sequence can be decomposed into several distinct sub-interactions each with a different CIF. For example, when observing that two people walk towards each other,

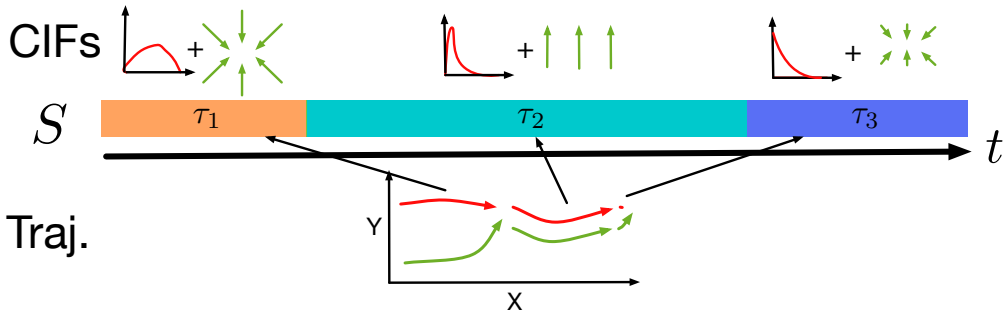


Figure 6.4: Temporal parsing by S (middle). The top demonstrates the change of CIFs in sub-interactions as the interaction proceeds. The bottom indicates the change of interactive behaviors in terms of motion trajectories. The colored bars in the middle depict the types of the sub-interactions.

shake hands and walk together, this long sequence can be segmented into three distinct sub-interactions. We represent meaningful interactivity as a sequence of latent sub-interactions $S = \{s_k\}_{k=1,\dots,K}$, where a latent sub-interaction determines the category of the CIF involved in a time interval $\mathcal{T}_k = \{t : t_k^1 \leq t \leq t_k^2\}$, such that $s^t = s_k, \forall t \in \mathcal{T}_k$. s_k is the sub-interaction label in the k-th interval representing the consistent interactivity of two agents in the relatively short interval. Fig. 6.4 illustrates the temporal parsing.

In each interval k, we define an interaction label $y_k \in \{0, 1\}$ to indicate the absence or presence of interactivity between the two agents. The interaction labels also constitute a sequence $Y = \{y^t\}_{t=1,\cdots,T}$. We have $y^t = y_k, \forall t \in \mathcal{T}_k$, where y_k denotes the interaction label in an interval $\mathcal{T}_k$.

6.3 Model Formulation

Given the input of motion trajectories Γ as defined in the above section, the model infers the posterior distribution of the latent variables S and Y using a Bayesian framework,

$$p(S, Y|\Gamma) \propto \underbrace{P(\Gamma \mid S, Y)}_{\text{likelihood}} \cdot \underbrace{P(S \mid Y)}_{\text{sub int. prior}} \cdot \underbrace{P(Y)}_{\text{int. prior}}. \tag{6.1}$$

The likelihood assesses how well the motion fields represented as a set of sub-interactions CIFs can account for relative motion observed in the video input, the spatial density of the relative

position, and the observed motion of the reference agent:

$$p(\Gamma \mid S, Y) = \prod_{k=1}^{K} \prod_{t \in \mathcal{T}_k} p(\tilde{\mathbf{v}}^t, \tilde{\mathbf{x}}^t, \mathbf{v}_1^t \mid s^t = s_k, y^t = y_k), \tag{6.2}$$

where the individual likelihood terms are defined by potential functions:

$$\log p(\tilde{\mathbf{v}}^t, \tilde{\mathbf{x}}^t, \mathbf{v}_1^t \mid s^t = s_k, y^t = y_k) \propto -U(\tilde{\mathbf{x}}^t, \tilde{\mathbf{v}}^t, \mathbf{v}_1^t | s_k, y_k). \tag{6.3}$$

Here, we assume that the potential function depends on the latent variables s_k and y_k to account for the variability in the motion patterns of different sub-interactions and to differentiate interactive motion from non-interactive motion. Eq. (6.3) also ensures that the expected interactive motion trajectories will move in the direction that minimizes the potential energy. We define the potential function in Eq. (6.3) as

$$U(\tilde{\mathbf{v}}^t, \tilde{\mathbf{x}}^t, \mathbf{v}_1^t \mid s^t = s_k, y^t = y_k) = \mathbf{w}_{s_k,y_k}^\top \phi(\tilde{\mathbf{x}}^t, \tilde{\mathbf{v}}^t, \mathbf{v}_1^t) + \beta_{s_k,y_k}, \tag{6.4}$$

where $\phi(\tilde{\mathbf{x}}^t, \tilde{\mathbf{v}}^t, \mathbf{v}_1^t) = [\tilde{\mathbf{x}}^{t\top}, \tilde{\mathbf{v}}^{t\top}, \mathbf{v}_1^{t\top}, \tilde{\mathbf{x}}^{t^\top} \tilde{\mathbf{v}}^t, ||\tilde{\mathbf{x}}^t||, ||\tilde{\mathbf{v}}^t||, ||\mathbf{v}_1^t||]^\top$ is the motion feature vector used to characterize the potential field, $\mathbf{w}_{s_k,y_k}$ and β_{s_k,y_k} are coefficients of the potential function learned for the specific latent variables s_k and y_k. There are certainly other ways to specify the potential function taking more motion patterns into account, such as acceleration, environment around the agents, and other possible factors of interest.

We model the prior term of sub-interactions $P(S|Y)$ using two independent components, i) the duration of each sub-interaction, and ii) the transition probability between two consecutive sub-interactions, as follows:

$$p(S \mid Y) = \prod_{k=1}^{K} \underbrace{p(|\mathcal{T}_k||s_k, y_k)}_{\text{duration}} \prod_{k=2}^{K} \underbrace{p(s_k|s_{k-1}, y_k)}_{\text{transition}}. \tag{6.5}$$

When $y_k = 1$, the two terms follow a log-normal distribution and a multinomial distribution, respectively; when $y_k = 0$, uniform distributions are used for the two terms instead.

Finally, we use a Bernoulli distribution to model the prior term of interactions $P(Y)$,

$$p(Y) = \prod_{k=1}^{K} \prod_{t \in \mathcal{T}_k} p(y^t = y_k) = \prod_{k=1}^{K} \prod_{t \in \mathcal{T}_k} \rho^{y^t} (1 - \rho)^{1 - y^t}. \tag{6.6}$$

6.4 Inference and Prediction

The model infers the current status of latent variables and produces an online prediction of future trajectories. Inference and prediction are performed for each time point from 1 to T sequentially (rather than offline prediction, which gives the labels after watching the entire video).

We denote trajectories from 0 to t as $\Gamma_{0:t}$, and the sub-interactions from 1 to $t-1$ as $S_{1:t-1}$. Without loss of generality, we assume there are K sub-interactions in $S_{1:t-1}$ with $\mathcal{T}_K$ being the last interval and $s^{t-1} = s_K$. We first infer s^t under the assumption of interaction (i.e., $y^t = 1$) by maximizing

$$p(s^t \mid \Gamma_{0:t}, S_{1:t-1}, y^t) \propto p(\tilde{\mathbf{v}}^t, \tilde{\mathbf{x}}^t, v_1^t \mid s^t, y^t) p(s^t \mid S_{1:t-1}, y^t), \tag{6.7}$$

where,

$$p(s^t \mid S_{1:t-1}, y^t) = \begin{cases} p(\tau \geq |\mathcal{T}_k| + 1 \mid s^t = s^{t-1}, y^t) & \text{if } s^t = s^{t-1} \\ p(\tau \geq 1 \mid s^t, y^t) p(s^t | s^{t-1}) & \text{otherwise} \end{cases}. \tag{6.8}$$

Then the posterior probability of $y^t = 1$ given $s^t \in \mathcal{S}$ is defined as

$$p(y^t \mid s^t, \Gamma_{0:t}, S_{1:t-1}) \propto p(s^t \mid \Gamma_{0:t}, S_{1:t-1}, y^t) p(y^t), \tag{6.9}$$

This computation makes it possible to perform the following inferences and online prediction: i) we maximize Eq. (6.7) to obtain the optimal s^t; ii) we use Eq. (6.9) to compute the posterior probability of two agents being interactive at t under the CIF of s^t as an approximation of the judgment of interaction/non-interaction provided by human observers; iii) the model can synthesize new trajectories using the following computation,

$$s^{t+1} \sim p(s^{t+1} \mid S_{1:t}, y^{t+1}), \tag{6.10}$$

$$\mathbf{x}_1^{t+1}, \mathbf{x}_2^{t+1} \sim p(\tilde{\mathbf{x}}^{t+1}, \tilde{\mathbf{v}}^{t+1}, v_1^{t+1} | s^{t+1}, y^{t+1}), \tag{6.11}$$

where $\tilde{\mathbf{v}}^{t+1}$, $\tilde{\mathbf{x}}^{t+1}$, and v_1^{t+1} are given by $\mathbf{x}_1^t$, $\mathbf{x}_1^{t+1}$, $\mathbf{x}_2^t$ and $\mathbf{x}_2^{t+1}$. By setting $y^{t+1} = 1$ or $y^{t+1} = 0$ in Eq. (6.10) and Eq. (6.11), we may synthesize interactive or non-interactive motion trajectories respectively.

6.5 Learning

To train the model, we used Gibbs sampling to find the S that maximizes the joint probability $P(Y, S, \Gamma)$. The implementation details are summarized below:

- Step 0: To initialize S, we first construct a feature vector for each time t (see the Appendix A). K-means clustering is then conducted to obtain the initial $\{s^t\}$, which also gives us the sub-interaction parsing S after merging the same consecutive s^t.
- Step 1: At each time point t of every training video, we update its sub-interaction label s^t by

$$s^t \sim p(\Gamma \mid S_{-t} \cup \{s^t\}, Y) p(S_{-t} \cup \{s^t\} \mid Y), \tag{6.12}$$

  where S_{-t} is the sub-interaction temporal parsing excluding time t, and $S_{-t} \cup \{s^t\}$ is a new sub-interaction sequence after adding the sub-interaction at t. Note that Y is always fixed in the procedure; thus we do not need $p(Y)$ term for sampling purpose.
- Step 2: If S does not change anymore, go to next step; otherwise, repeat step 1.
- Step 3: Since we do not include the non-interactive videos in the training set, we selected 22 videos in the first human experiment (a mixture of interactive and non-interactive videos) as a validation set to estimate coefficients of the potential functions under $y = 0$ by maximizing the correlation between the model prediction of Eq. (6.9) and the average human responses in the validation set. To simplify the search, we assume all potential functions under $y = 0$ share the same coefficients across all latent sub-interactions.

6.6 Model Simulation Results

We trained the model using two sets of training data, the UCLA aerial event dataset [SXR15] and the Heider-Simmel animation dataset.

6.6.1 Training with Aerial Videos

In the UCLA aerial event dataset collected by [SXR15], about 20 people performed some group activities in two scenes (a park or a parking lot), such as group touring, queuing in front of a vending machine or playing Frisbee. People's trajectories and their activities are manually annotated. The dataset is available at https://tshu.io/AerialVideo/AerialVideo.html.

One advantage of using aerial videos to generate decontextualized animations is that the technique provides sufficient training stimuli to enable the learning of representations of critical movement patterns that signal potential interactivity between agents. We selected training videos including interactivity from the database, so that the two agents always interact with each other in all training stimuli. Thus, for any training video, $y^t = 1, \forall t = 1, \cdots, T$. During the training phase, we excluded the examples used in human experiments. In total, there were 131 training instances.

In the implementation, we manually define the maximum number of sub-interaction categories to be 15 in our full model (i.e., $|\mathcal{S}| = 15$), which is over-complete for our training data according to learning (low frequency in the tail of Fig. 6.5). With simulated annealing [KGV83], Gibbs sampling converges within 20 sweeps (where a sweep is defined as all the latent sub-interaction labels being updated once). The frequencies of the top 15 CIFs are highly unbalanced. In fact, the top 10 CIFs account for 83.8% of the sub-interactions in the training data. The first row of Fig. 6.6 provides a visualization of the top 5 CIFs. Each of the top CIFs indicates some different behavioral patterns in the aerial videos. For example, the No.1 CIF signals the approaching behavior that one agent moves towards a reference agent. Interestingly, the converging point of the approaching is not at the center of the location of the reference agent. Instead, the agent heads towards the future location of the reference agent (above-the-center position in the flow figure), implying that the fundamental characteristic of human interactions is being predictive.

6.6.2 Training with Heider-Simmel Videos

The second dataset was created from the original Heider-Simmel animation (i.e., two triangles and one circle). We extracted the trajectories of the three shapes, and thus obtained 3 pairs of two-agent interactions. We truncated the movie into short clips (about 10 seconds) to generate a total of 27

videos. The same algorithm was used to train the model with 15 types of CIFs.

The most frequent five CIFs are visualized in the second row of Fig. 6.6. Clearly, the richer behavior in the Heider-Simmel animation yielded a variety of CIFs with distinct patterns compared to the CIFs learned from aerial videos. For example, the top CIF indicates that one agent moves around the reference agent, a common movement pattern observed in Heider-Simmel animations. The second CIF signals a "run away" movement to avoid the reference agent. The frequencies of CIFs are also more distributed in this dataset, as shown in Fig. 6.5.

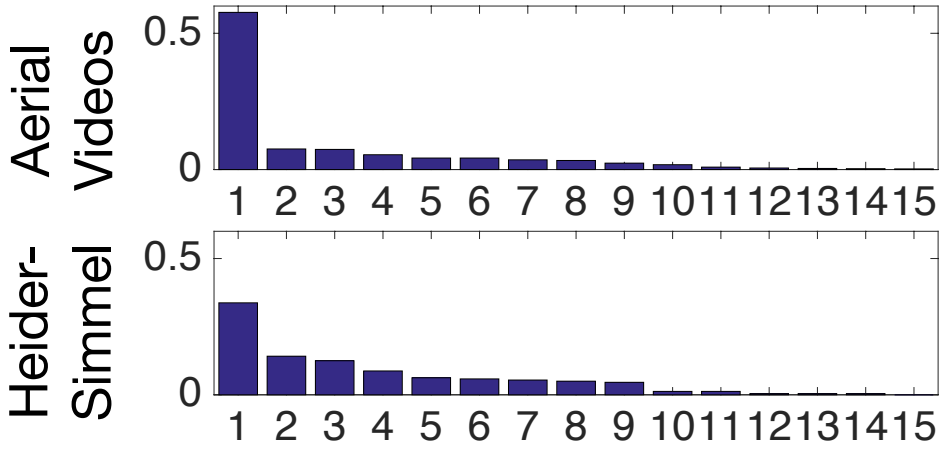


Figure 6.5: The frequencies of learned CIFs with the training data generated from aerial videos (top) and the Heider-Simmel movie (bottom). The numbers on the x axis indicate the IDs of CIFs, ranked according to the occurrence frequency in the training data.

6.6.3 Generalization: Training with Aerial Videos and Testing with Heider-Simmel Videos

We tested how well the model trained with the aerial videos ($|\mathcal{S}| = 15$) can be generalized to a different dataset, the Heider-Simmel animations. This generalization test aims to examine if the critical movement patterns learned from real-life situations can account for perceived interactiveness in laboratory stimuli. Fig. 6.7 shows the model simulation results for a few Heider-Simmel videos. We notice that the interactiveness ratings predicted by the model vary over time. Such variability is consistent with subjective impressions that the Heider-Simmel animations elicit different degrees of animacy and interactivity at different time points. In addition, most clips in Heider-Simmel animations are rated by the model as having a high probability of being interactive

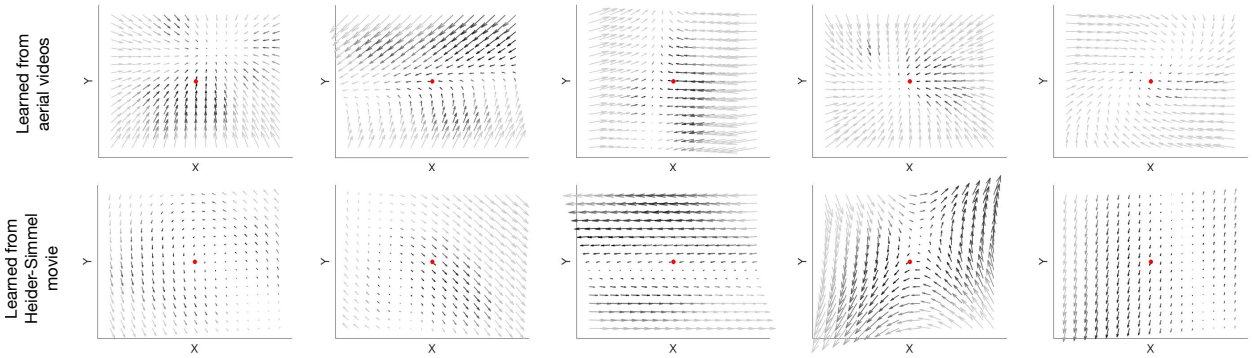


Figure 6.6: Interactive fields of the top five frequent CIFs learned from aerial videos (top) and Heider-Simmel movie (bottom) respectively. In each field, the reference agent (red dot) is at the center of a field i.e., (0,0), moving towards north; the arrows represent the mean relative motion at different locations and the intensities of the arrows indicate the relative spatial density which increases from light to dark. We observed a few critical CIFs that signal common interactions from the two simulation results. For instance, in aerial videos, we observed i) approaching, e.g., CIF 1, and ii) walking in parallel, or following, e.g., the lower part of CIF 2. The Heider-Simmel animation revealed additional patterns such as i) orbiting, e.g., CIF 1, and ii) leaving, e.g., CIF 4, iii) walking-by, e.g., CIF 5.

(i.e., mostly above 0.5), consistent with human observers' impression about the highly animate and interactive behaviors conveyed in the animations. Also, the model was able to give continuous online predictions based on the relative speeds and spatial locations of the two objects. For example, when the two objects approach each other or follow each other, the model yields higher interactive ratings.

The qualitative analysis of the model performance suggests that the model trained with aerial videos shows a certain degree of generalization to the Heider-Simmel animations. However, unsurprisingly, objects in aerial videos share different characteristics of motion patterns from the motions involved in Heider-Simmel animations (as illustrated in the training results of CIFs in Fig. 6.6). For example, orbiting behavior barely occurs in the aerial video dataset, and accordingly the model yields relatively low interactiveness predictions when observing such behavior, which is relatively common in the Heider-Simmel animations. In the next section, we will report human experiments that can quantitatively assess how well the model can account for human performance.

6.7 Experiment 1

6.7.1 Stimuli

24 interactive stimuli were generated from different pairs of human interactions in aerial videos. We selected two people interacting with each other in each aerial video. We then generated the decontextualized animations by depicting the two people as dots with different colors. The dots' coordinates were first extracted from the aerial videos by human annotators. Note that the two dots were first re-centered to localize the midpoint at the center of the screen in the first frame. The coordinates were temporally smoothed by averaging across the adjacent 5 frames.

24 non-interactive stimuli were generated by interchanging motion trajectories of two people selected from two irrelevant interactive videos (e.g., the motion of one dot in video 1 recombined with the motion of a dot in video 2). The starting distances between two dots in non-interactive stimuli were kept the same as in the corresponding interactive stimuli.

The duration of stimuli varied from 239 frames to 500 frames (mean frame = 404), corresponding to 15.9 to 33.3 seconds, with a recording refresh rate of 15 frames per second. The diameters of dots were 1° of visual angle. One dot was displayed in red (1.8 cd/m^2) and the other in green (30 cd/m^2) on a black background (0 cd/m^2). Among the 48 pairs of stimuli, four pairs of actions (two interactive and two non-interactive) were used as practice.

6.7.2 Participants

33 participants (mean age = 20.4; 18 female) were enrolled from the subject pool at the Department of Psychology, University of California, Los Angeles (UCLA). They were compensated with course credit. All participants had normal or corrected-to-normal vision.

6.7.3 Procedures

Participants were seated 35 cm in front of a screen, which had a resolution of 1024×768 and a 60 Hz refresh rate. First, participants were given a cover story: "Imagine that you are working for a

company to infer whether two people carry out a social interaction based on their body locations measured by GPS signals. Based on the GPS signal, we generated two dots to indicate the location of the two people being tracked." The task was to determine when the two dots were interacting with each other and when they were not. Participants were asked to make continuous responses across the entire duration of the stimuli. They were to press and hold the left-arrow or right-arrow button for interactive or non-interactive moments, respectively, and to press and hold the down-arrow button if they were unsure. If no button was pressed for more than one second, participants received a 500 Hz beep as a warning.

Participants were presented with four trials of practice at the beginning of the session to familiarize them with the task. Next, 44 trials of test stimuli were presented. The order of trials was randomized for each participant. No feedback was presented on any of the trials. The experiment lasted for about 30 minutes in total.

6.7.4 Results

Interactive, unsure and non-interactive responses were coded as 1, 0.5, and 0, respectively. Frames with no responses were removed from the comparison. Human responses are shown in Fig. 6.8. A paired-sample t-test revealed that the average ratings of non-interactive actions (M = 0.34, SD = 0.13) were significantly lower than interactive actions (M = 0.75, SD = 0.13), $t(32) = 13.29$, $p < 0.001$. This finding indicates that human observers are able to discriminate interactivity based on decontextualized animations generated from the real-life aerial videos.

To compare the model predictions with human continuous judgments, we computed the average human ratings, and ran the model to simulate online predictions of sub-interaction and interaction labels on the testing videos (excluding the ones in the validation set). Specifically, we used Eq. (6.9) to compute the probability of two agents being interactive with each other at any time point t. The model simulation used the hyper-parameters $\rho = 10^{-11}$ and $\sigma_0 = 1.26$.

Table 6.1 summarizes the Pearson correlation coefficient r and root-mean-square error (RMSE) between the model predictions and the human ratings using aerial videos as training data. We compared our hierarchical model with two baseline models: i) Hidden Markov Model (HMM),

Table 6.1: The quantitative results of all methods in Experiment 1 using aerial videos as training data.

Method	HMM	One-Interaction	Hierarchical Model		
			$\|\mathcal{S}\| = 5$	$\|\mathcal{S}\| = 10$	$\|\mathcal{S}\| = 15$
r	0.739	0.855	0.882	0.911	0.921
RMSE	0.277	0.165	0.158	0.139	0.134

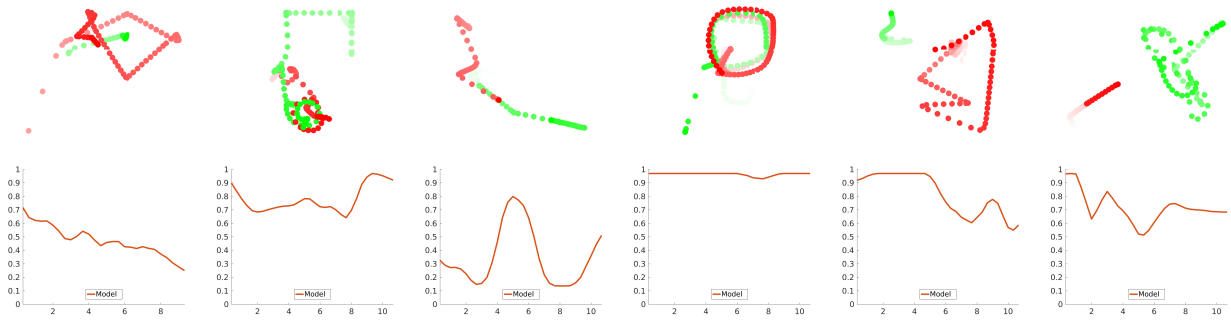


Figure 6.7: (Top) Examples of moving trajectories of selected objects in the Heider-Simmel animation dataset. One object is plotted in red and the other one is plotted in green. The intensity of colors increases with time lapse, with darker color representing more recent coordinates. (Bottom) Corresponding online predictions on the example Heider-Simmel videos by our full model ($|\mathcal{S}| = 15$) trained on aerial videos over time (in seconds).

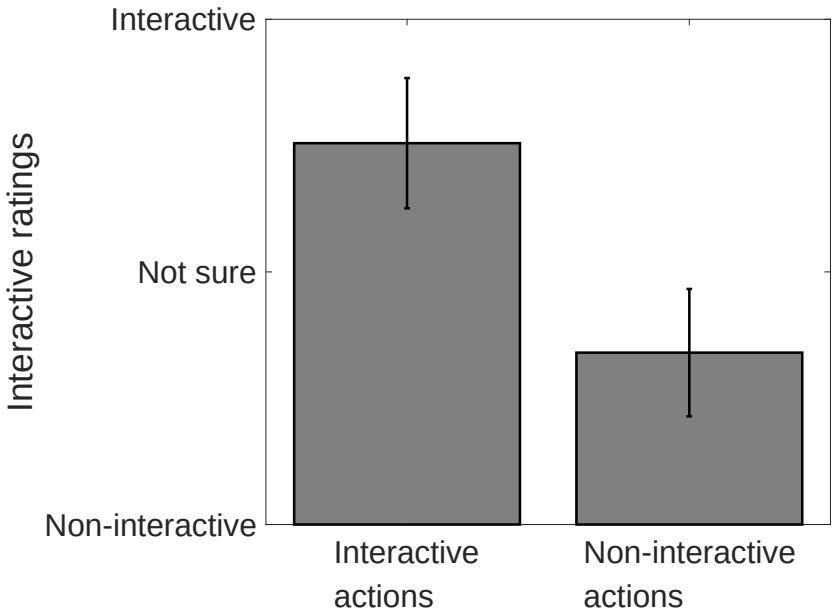


Figure 6.8: Mean ratings of the interactive versus non-interactive actions in the experiment 1. Error bars indicate +/- 1 SEM.

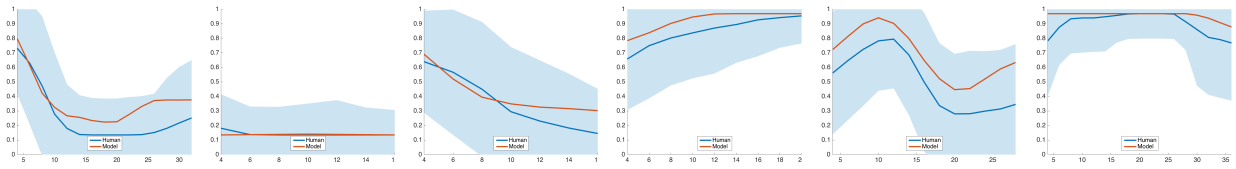


Figure 6.9: Comparison of online predictions by our full model trained on aerial videos ($|\mathcal{S}| = 15$) (orange) and humans (blue) over time (in seconds) on testing aerial videos. The shaded areas show the standard deviations of human responses at each moment.

where the latent variables s^t and y^t only depend on their preceding variables s^{t-1} and y^{t-1}; ii) a model with only one type of sub-interaction. Both models yielded poorer fits to human judgments (i.e., lower correlation and higher RMSE) than the hierarchical model. In addition, we changed the number of sub-interaction categories to examine how sensitive our model is to this parameter. The results clearly show that i) only using one type of sub-interaction provides reasonably good results, $r = .855$, and ii) by increasing the number of sub-interactions $|\mathcal{S}|$, the fits to human ratings were further improved until reaching a plateau with a sufficiently large number of sub-interactions.

Fig. 6.9 shows results for a few videos, with both model predictions and human ratings. The model predictions accounted for human ratings quite well in most cases. However, the model predictions were slightly higher than the average human ratings, which may be due to the lack of negative examples in the training phase. We also observed high standard deviations in human responses, indicating large variability of the online prediction task for every single frame in a dynamic animation. In general, the difference between our model's predictions and human responses are seldom larger than one standard deviation relative to human responses.

We also used the model trained from the Heider-Simmel animation and tested it on the stimuli generated from the aerial videos. This procedure yielded a correlation of 0.640 and RMSE of 0.227. The reduced fit for this simulation indicates the discrepancy between moving patterns of the two types of training datasets. The CIFs learned from one dataset may be limited in generalization to the other dataset.

6.8 Experiment 2

One advantage of developing a generative model is that it enables the synthesis of new videos by Eq. (6.10) and Eq. (6.11), based on randomly sampled initial positions of the two agents ($\mathbf{x}_1^0$, $\mathbf{x}_2^0$) and the first sub-interaction s^1. By setting the interaction labels to be 1 or 0, the synthesized stimuli can be controlled to vary the degree of interactiveness. In Experiment 2, we aimed to use the model to synthesize new animations and see if interactiveness can be accurately perceived by human observers.

We used the model trained on aerial videos to synthesize 10 interactive and 10 non-interactive

animation clips. 17 participants were enrolled from the subject pool at UCLA. The procedure of Experiment 2 was similar to that of Experiment 1. The 20 synthesized videos were presented to human observers in random orders. The task was to press one of the two buttons at the end of the action to judge if the two dots were interacting or not.

The interactiveness between the two agents in the synthesized videos was judged accurately by human observers, with the average ratings of the synthesized non-interactive actions (M = 0.15, SD = 0.15) significantly lower than the synthesized interactive actions (M = 0.85, SD = 0.20), $t(16) = 14.00$, $p < 0.001$. The model prediction of a whole video is set to be the average predictions of Eq. (6.9). The correlation between model predictions and average human responses was high, 0.94. The results suggested that humans reliably perceived interactiveness from the synthesized stimuli, and were sensitive to model-controlled degree of interactivity.

6.9 Discussion

In this paper, we examined human perception of social interactions using decontextualized animations based on movement trajectories recorded in aerial videos of a real-life environment, as well as Heider-Simmel-type animations. The proposed hierarchical model built on two key components: conditional interactive fields of sub-interactions, and temporal parsing of interactivity. The model fits human judgments of interactiveness well, and suggests potential mechanisms underlying our understanding of meaningful human interactions. Human interactions can be decomposed into sub-interactions such as approaching, walking in parallel, or standing still in close proximity. Based on the transition probabilities and the duration of sub-components, humans are able to make inferences about how likely the two people are interacting.

Our study indicates that rapid judgments on human interactivity can be elicited by the detection of critical visual features such as CIFs, without the involvement of a high-level reasoning system. The fairly fast, automatic, irresistible and highly stimulus-driven impressions about animacy and interactivity are largely perceptual in nature. This result is consistent with the literature on causal perception [ST00, PTL17, Joh73, BL11, BL12, TL13, TL14, SBL16]. Hence the detection of interactivity between agents is likely to be processed as in the proposed model without the explicit

modeling of intention and goals. This process is efficient, but not sufficient to address questions such as why and how the interactions are carried out between the agents. When these questions are important for a particular task in the social context, the reasoning system and the theory-of-mind system will be called upon after the perception of interactivity has been signaled. Future work should focus on the interplay between the two general systems involved in perception and in inference of human interactions.

The model provides a general framework and can be extended to include hidden intentions and goals. By modifying the potential function in the model, the computational framework can be applied to more sophisticated recognition and understanding of social behavioral patterns. While previous work has focused on actions of individuals based on detecting local spatial-temporal features embedded in videos [DRC05], the current work can deal with multi-agent interactions. Understanding the relation between agents could facilitate the recognition of individual behaviors by putting single actions into meaningful social contexts. The present model could be further improved to enhance its flexibility and broaden its applications. The parametric linear design of CIFs provides computational efficiency, and temporally parsing an interaction into multiple sub-interactions enhances the linearity in each sub-interaction. However, this design may not be as flexible as non-parametric or non-linear models, such as a Gaussian process. In addition, the current model is only based on visual motion cues. The model could be enhanced by incorporating a cognitive mechanism (e.g., a theory-of-mind framework) to enable explicit inference of intentions.

CHAPTER 7

A Unified Computational Framework for Modeling Physical and Social Events

7.1 Introduction

Imagine you are playing a multi-player video game with open or free-roaming worlds. You will encounter many physical events, such as blocks collapsing onto the ground, as well as social events, such as avatars constructing buildings or fighting each other. All these physical and social events are depicted by movements of simple geometric shapes, which suffice to generate a vivid perception of rich behavioral, including interactions between physical entities, interpersonal activities between avatars engaged in social interactions, or actions involving both humans and objects.

This type of rich perception elicited by movements within simple visual displays has been extensively studied in psychology. Prior work (e.g., [HS44, Mic63, Kas81, PG89, ST00, GNS09, GMS10]) have provided convincing evidence that humans possess a remarkable ability to perceive and reconstruct both physical events and social events from simple very limited visual inputs in an efficient and robust way.

Although many studies of both intuitive physics and social perception examined dynamic stimuli consisting of moving shapes, these research areas have largely been isolated from one another, with different theoretical approaches and experimental paradigms. In the case of physical events, research has been focused on the perception and interpretation of physical objects and their dynamics, aiming to determine whether humans use heuristics or mental simulation to reason about intuitive physics (see a recent review by [KHL17]). For social perception, some research has aimed to identify critical cues based on motion trajectories that determine the perception of animacy and

social interactions [DL94, ST00, GNS09, SPF17, SPF18]. Chapter 6 falls into this category. There has also been work focusing on inferences about agents' intentions [BST09, UBM10, PBC14]. In contrast to the clear separation between the two research topics, human perception integrates the perception of physical and social events. Hence, it is important to develop a common computational framework applicable to both intuitive physics and social perception to advance our understandings on how humans perceive and reason about physical and social events.

In this chapter, we propose a unified computational framework for modeling both physical events and social events based on movements of simple shapes. In particular, we unify the physical and social modeling in three ways.

First, we design a unified physical and social simulation for generating Heider-Simmel animations in which simple moving shapes vary in degrees of physical violation and the involvement of intention. Prior work usually created Heider-Simmel-type stimuli using manually designed interactions [GNS09, GMS10, IKB17], rule-based behavior simulation [KC10, PBC14], and trajectories extracted from human activities in aerial videos [SPF18]. It is challenging to manually create many motion trajectories, and to generate situations that violate physical constraints. Accordingly, we develop a joint physical-social simulation-based approach built upon a 2D physics engine (Figure 7.1). A similar idea has been previously instantiated in a 1D environment, Lineland [Ull15]. By generating Heider-Simmel-type animations in a 2D environment with the help of deep reinforcement learning, our simulation approach is able to depict a richer set of motion patterns in animations. This advanced simulation provides well-controlled Heider-Simmel stimuli enabling the measurement of human perception of physical and social events for hundreds of different motion patterns.

Second, we propose a unified physical and social concept learning paradigm by formulating the concept learning process as the pursuit of generalized coordinates and the corresponding parsimonious potential energy functions. This is inspired by Lagrangian mechanics, where the dynamics of a complex system can be fully captured by a few simple scalar functions, i.e., potential energy functions based on generalized coordinates that are intuitive to humans. We show that from a handful of examples of simple shapes' movements generated by our joint physical-social simulation engine, this learning paradigm can discover interpretable physical and social concepts (as

generalized coordinates) and model physical laws as well as social behaviors (by potential energy functions). Based on the learned physical laws and social behaviors, we also develop general metrics of the physical violation and the likelihood of pursuing certain goals for entities in the generated animations given their motion trajectories.

Third, we aim to construct a unified psychological space that may reveal the partition between the perception of physical events involving inanimate objects and the perception of social events involving human interactions with other agents. Specifically, we hypothesize that this unified space includes two prominent dimensions: an intuitive sense regarding whether physical laws are obeyed or violated; and an impression of whether an agent possesses intentions in the display. In this work, these two indices are computed based on the metrics proposed in our computational model to measure how well the motion patterns in an animation satisfy physics, and the likelihood that entities are human agents showing intentions. Note that the intuitive sense of physical violation may result from observable physical forces that can not be explained by perceived entity properties (such as motion, size, etc.) in a scene. The development of this unified space may shed light on many fundamental problems in both intuitive physics and social perception.

To construct such space, we project a video rendered by our simulation engine as a whole onto the space. Hence, a large range of videos can provide a distribution of observed events. We can also project individual entities in one physical or social event onto the same space, and then examine pairwise relations between the projected locations of entities in the space, which could serve as an informative cue for judging social/physical roles of entities (e.g, as an human agent or an inanimate object). In two experiments, we combined model simulations with human responses to validate the proposed psychological space.

7.2 Stimulus Synthesis in Flatland

7.2.1 Overview

Can we have a unified view of physical events with inanimate objects and social events with human agents? Can we create a continuous transition from objects to agents, and from agents back to

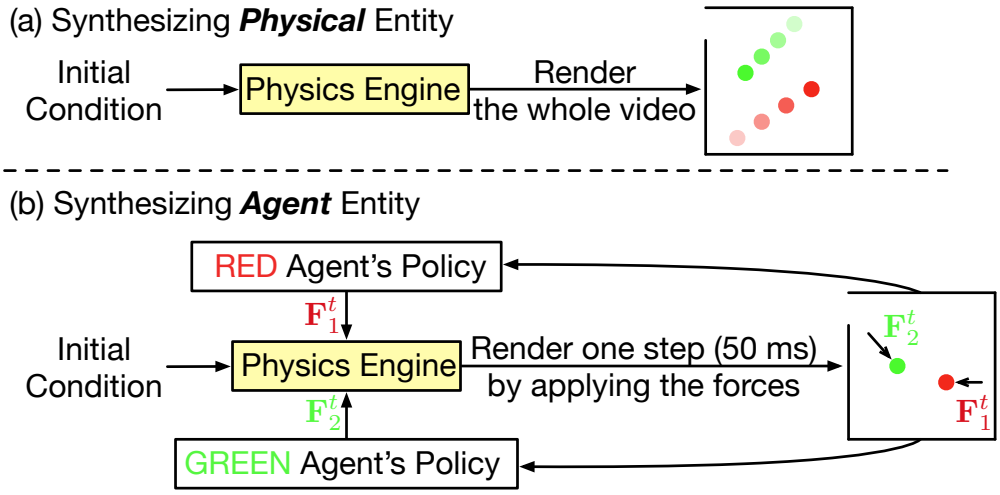


Figure 7.1: Overview of our joint physical-social simulation engine. For a dot instantiating a physical object, we randomly assign its initial position and velocity and then use physics engine to simulate its movements. For a dot instantiating a human agent, we use policies learned by deep reinforcement learning to guide the forces provided to the physics engine.

objects? In other words, can we bridge physics and social behaviors? We believe that the first step towards addressing these questions should be building a simulation engine that can generate both physical interactions and social interactions in a principled manner, so that the two types of interactions can emerge in the same world. Therefore, we propose a joint physical and social simulation engine, Flatland[1], where entities are moving in a 2D environment w.r.t. the motion rendered by a physics engine.

Figure 7.1 gives an overview of our joint physical-social simulation engine. Each video included two dots (red and green) and a box with a small gap indicating a room with a door. The movements of the two dots were rendered by a 2D physics engine (pybox2d[2]). If a dot represents an object, we randomly assigned the initial position and velocity, and then used the physics engine to synthesize its motion. Note that our simulation incorporated the environmental constraints (e.g., a dot can bounce off the wall, the edge of the box), but did not include friction. If a dot represents an agent, it was assigned with a clearly-defined goal (e.g., leaving room) and pursued its goal

[1]Inspired by the classic mathematical fiction, *Flatland: A Romance of Many Dimensions* by Edwin A. Abbott.

[2]https://github.com/pybox2d/pybox2d

by exerting self-propelled forces (e.g., pushing itself towards the door). The self-propelled forces were sampled from agent policy learned by deep reinforcement learning (see more details in a later subsection). Specifically, at each step (every 50 ms), the agent observed the current state rendered by the physics engine, and its policy determined the best force to advance the agent's pursuit of its goal. We then programmed the physics engine to apply this force to the dot, and rendered its motion for another step. This process was repeated until the entire video was generated.

7.2.2 Interaction Types

As summarized in Figure 7.2, we consider three types of interactions, including human-human (HH), human-object (HO) and object-object (OO) interactions, all of which are generated by the approach depicted in Figure 7.1. Note that in this work we treat the terms "human" and "agent" interchangeably. When synthesizing the agents' motion, we set two types of goals for the agents, i.e., "leave the room" (g_1) and "block the other entity" (g_2). Specially, in HH stimuli, one agent has a goal of leaving the room (g_1), and the other agent aims to block it (g_2); in HO stimuli, an agent always attempts to keep a moving object within the room (g_2) and the object has an initial velocity towards the door. By randomly assigning initial position and velocity to an agent, we can simulate rich behaviors that can give the impression such as blocking, chasing, attacking, pushing, etc.

In addition to the three general types of interactions, we have also created sub-categories of interactions to capture a variety of physical and social events. For OO animations, we included four events, as collision, connections with rod, spring and soft rope. Since these connections were invisible in the displays, the hidden physical relations may result in a subjective impression of animacy or social interactions between the entities. In addition, the invisible connections between objects (rod, spring, and soft rope) introduce different degrees of violation of physics in the motion of the corresponding entities if assuming the two entities are independent. For HH animations, we varied the "animacy degree" (AD) of the agents by controlling how often they exerted self-propelled forces in the animation. In general, a higher degree of animacy associates with more frequent observations about violation of physics, thus revealing self-controlled behaviors guided by the intention of an agent. The animacy manipulation introduced five sub-categories of HH

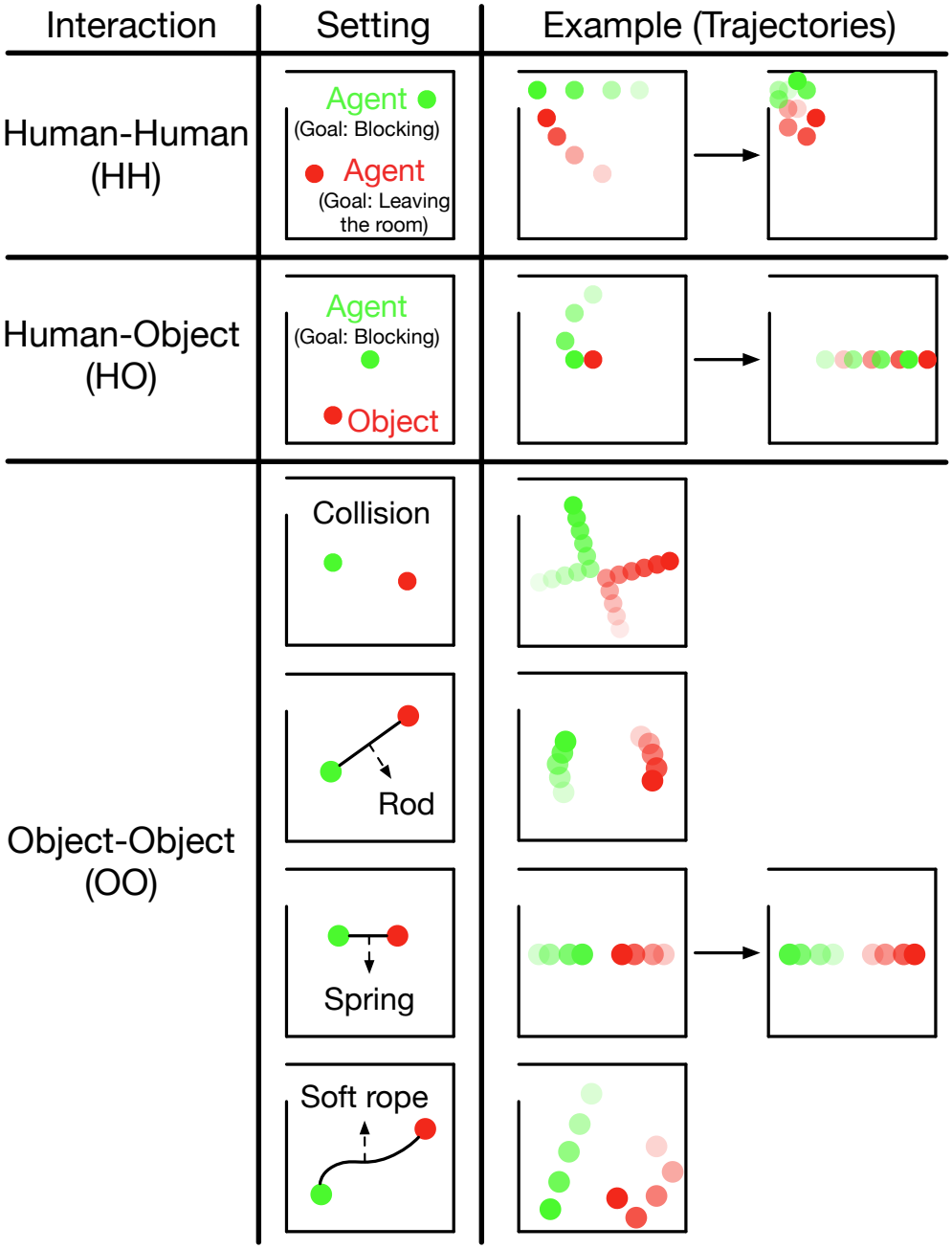


Figure 7.2: An illustration of three types of synthesized interactions for physical and social events. A few examples are included by showing trajectories of the two entities. The dot intensities change from low to high to denote elapsed time. Note that the connections in OO stimuli (i.e., rod, spring, and soft rope) are drawn only for illustration purpose. Such connections were invisible in the stimuli. Examples of stimuli are available at: https://tshu.io/HeiderSimmel/CogSci19.

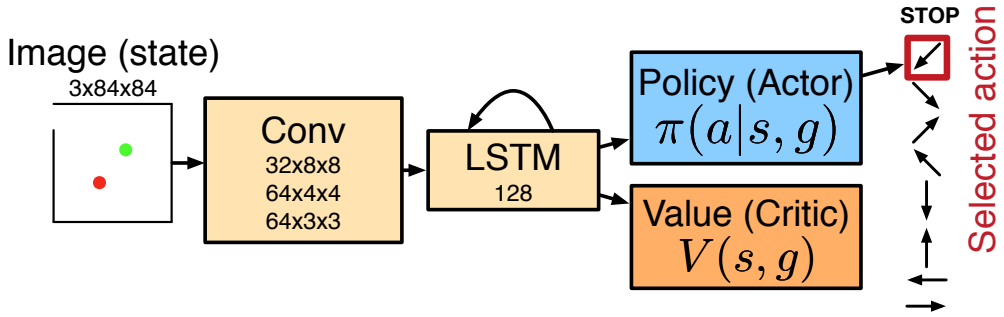


Figure 7.3: The deep RL network architecture for learning policy for goal-directed movements of an agent. For each goal, we train a separate network with the same architecture.

stimuli with five degrees of animacy – 7%, 10%, 20%, 50%, and 100%, respectively corresponding to applying force once for every 750, 500, 250, 100, and 50 ms. In an HH animation, we assigned the same level of animacy degree to both dots.

7.2.3 Training Policies

As shown in Figure 7.1, in order to generate social events, we need sensible policies to infer the self-propelled forces for pursuing goals. However, searching for such policies in a physics engine is extremely difficult. In this study, we use deep reinforcement learning (RL) to acquire such policies, which has been shown to be a powerful tool for learning complex policies in recent studies [SSS17]. Formally, an agent's behavior is defined by an Markov decision process (MDP), $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, R, \mathcal{G}, \gamma \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state space (raw pixels as in Figure 7.3) and action space, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{S}$ are the transition probabilities of the environment (in our case, deterministic transitions defined by physics), R is the reward function associated with the intended goals $g \in \mathcal{G}$, and $0 < \gamma \leq 1$ is a discount factor. To match to the experimental setup, we define two reward functions for the two goals: i) for "leaving the room", the agent receives a reward, $r^t = R(s^t, g_1) = \mathbb{1}(\text{out of the room})$, at step t; ii) for "blocking", the reward at step t is $r^t = R(s^t, g_2) = -\mathbb{1}(\text{opponent is out of the room})$. To simplify the policy learning, we define a discrete action space, which corresponds to applying forces with the same magnitude in one of the eight directions and "stop" (the agent's speed decreases to zero after applying necessary force).

The objective of the deep RL model is to train the policy network shown in Figure 7.3 to

maximize the expected return $E[\sum_{t=0}^{\infty} \gamma^t r^t]$ for each agent. The optimization was implemented using advantage actor critic (A2C) [MBM16] to jointly learn a policy (actor) $\pi : \mathcal{S} \times \mathcal{G} \mapsto \mathcal{A}$ which maps an agent's state and goal to its action, and a value function (critic) $V : \mathcal{S} \mapsto \mathbb{R}$. The two functions were trained as follows (assuming that entity i is an agent):

$$\nabla_{\theta_\pi} J(\theta_\pi) = \nabla_{\theta_\pi} \log \pi(a_i^t | s_i^t, g_i; \theta_\pi) A(s_i^t, g_i), \tag{7.1}$$

$$\nabla_{\theta_V} J(\theta_V) = \nabla_{\theta_V} \frac{1}{2} \left(\sum_{\tau=0}^{\infty} \gamma^\tau r_i^{t+\tau} - V(s_i^t, g_i; \theta_V) \right)^2, \tag{7.2}$$

where $A(s_i^t, g_i) = \sum_{\tau=0}^{\infty} \gamma^\tau r_i^{t+\tau} - V(s_i^t, g_i)$ is an estimate of the advantage of current policy over the baseline $V(s_i^t, g_i)$. We set $\gamma = 0.95$ and limit the maximum number of steps in an episode to be 30 (i.e., 1.5 s). Note that we train a network for each goal with the same architecture. In HH animations, an agent's policy depends on its opponent's policy. To achieve a joint policy optimization for both agents, we adopt an alternating training procedure: at each iteration, we train the policy of one of the agents by fixing its opponent's policy. In practice, we trained the polices by 3 iterations.

7.3 Unified Physical and Social Concept Learning

7.3.1 Inspiration from Lagrangian Mechanics

Why do we want to construct potential energy functions to model physical and social systems and learn the underlying physical and social concepts? To answer this, let us first look at the comparison between Lagrangian mechanics (based on potential energy) and Newtonian mechanics (direct force analysis).

Consider a system of N particles with the same mass (i.e., $m_i = m$, $\forall i = 1, \cdots, N$) where their positions are $(\mathbf{x}_1(t), \mathbf{x}_2(t), \cdots, \mathbf{x}_N(t))$ in Cartesian coordinates at time t. The surrounding environment (context) is denoted as c. The Lagrangian of this system is defined as

$$L = L(\mathbf{x}_1, \cdots, \mathbf{x}_N, \dot{\mathbf{x}}_1, \cdots, \dot{\mathbf{x}}_N, t) = T - U, \tag{7.3}$$

where $T = T(\dot{\mathbf{x}}_1, \cdots, \dot{\mathbf{x}}_N, t) = \sum_{i=1}^{N} \frac{1}{2} m \dot{\mathbf{x}}_i(t)^2$ is the kinetic energy of all entities and U is the potential energy. When there are only conservative forces in the system, the potential energy solely

depends on the coordinates of the entities, i.e., $U = U(\mathbf{x}_1, \cdots, \mathbf{x}_N, t)$. For convenience, we may drop the notation t sometimes.

From the Euler-Lagrange equation, we may derive the motion of equations for each entity:

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{\mathbf{x}}_i} - \frac{\partial L}{\partial \dot{\mathbf{x}}_i} = 0, \quad \forall i = 1, \cdots, N. \tag{7.4}$$

By plugging in T and U, this in fact gives us Newton's second law:

$$m\ddot{\mathbf{x}}_i = \mathbf{F}_i = -\frac{\partial U(\mathbf{x}_1, \cdots, \mathbf{x}_N)}{\partial \mathbf{x}_i}. \tag{7.5}$$

This implies that as an alternative approach to conducting explicit force analysis which is often extremely difficult in complex systems, we can instead derive forces from a few scalar functions, i.e., potential energy functions. This advantage becomes more significant when we adopt suitable generalized coordinates which constitutes potential energy functions in simple forms.

7.3.2 Parsimonious Models from Generalized Coordinates

Formally, we may convert the Cartesian coordinates of the N entities into a generalized coordinate system $\mathbf{q} = (q_j)_{j=1}^D$, where D is usually the number of degrees of freedom in the system. Each dimension is derived from a transformation function $q_j = \phi_j(\mathbf{x}_1, \cdots, \mathbf{x}_N, c)$, where c is the context (e.g., surrounding environment) of the current system. These coordinates' first-order derivatives $\dot{\mathbf{q}} = (\dot{q}_j)_{j=1}^D$ become generalized velocities accordingly. Here, ϕ_j could be understood as a type of state representation extracted from the raw observations. Based on the generalized coordinates, we can redefine the Lagrangian:

$$L = L(\mathbf{q}, \dot{\mathbf{q}}) = T - U, \tag{7.6}$$

where $T = T(\mathbf{q}, \dot{\mathbf{q}}) = T(\dot{\mathbf{x}}_1, \cdots, \dot{\mathbf{x}}_N)$ is the kinetic energy, and V is the potential energy. Again, if we only consider conservative forces, we will have $U = U(\mathbf{q})$. The Euler-Lagrange equation still holds for the generalized coordinates:

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_j} - \frac{\partial L}{\partial q_j} = 0, \quad \forall j = 1, \cdots, D. \tag{7.7}$$

The resulting equations of motion describe the dynamics of the system as a whole in terms of how generalized coordinates (i.e., the physical quantities of interest) change over time. We can map the

(a) (b)

Figure 7.4: Systems with circles and springs. (a) Two entities (circles) connected by a massless spring. The Cartesian coordinates of the two entities are $\mathbf{x}_1$ and $\mathbf{x}_2$. The potential energy of this system can be defined by using just one variable, i.e., the distance between the two entities. (b) Three entities connected by two massless springs.

motion back to individual entity's Cartesian coordinates based on the transformation functions ϕ_j:

$$m\ddot{\mathbf{x}}_i = \mathbf{F}_i = -\sum_{j=1}^{D} \frac{\partial U(\mathbf{q})}{\partial q_j} \frac{\partial \phi_j}{\partial \mathbf{x}_i} \quad \forall i = 1, \cdots, N. \tag{7.8}$$

The use of generalized coordinates allows us to greatly simplify the derivation of forces (or dynamics) for entities in a system, which ultimately results in a parsimonious model to describe the dynamics of a system. Therefore, by constructing the most suitable generalized coordinates, the key characteristics of a system may naturally emerge from raw observations. Consider the spring system shown in Figure 7.4a as an example. Assume the equilibrium length of the spring is l and its constant is k, then potential energy of this system can be conveniently defined by only one variable – the distance between the two entities (or equivalently the length of the spring). Let $q = \phi(\mathbf{x}_1, \mathbf{x}_2) = ||\mathbf{x}_1 - \mathbf{x}_2||$, the potential is $U(q) = \frac{1}{2}k(q-l)^2$, which is a simple quadratic function of q. Based on Eqn. 7.8, we can derive the forces applied to the two entities accordingly: $\mathbf{F}_1 = -k(q-l)(\mathbf{x}_1 - \mathbf{x}_2)/q$, $\mathbf{F}_2 = -k(q-l)(\mathbf{x}_2 - \mathbf{x}_1)/q$.

7.3.3 Modular Models and Triggering Conditions

Multiple independent potential energy functions may coexist in a complex system, and the overall potential energy is simply the sum of all individual potential energy functions. This naturally leads to a modular design, where the potential energy of any system is a combination of atomic potential

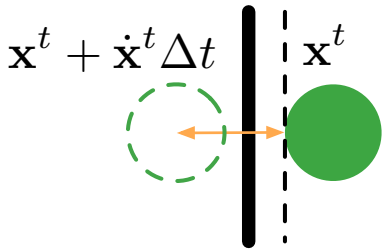
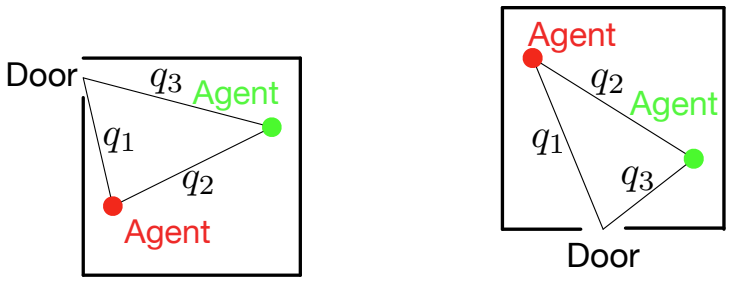


Figure 7.5: A circle bouncing off a wall. The generalized coordinate in this case can be derived as the expected violation after a short period of time Δt based on the entity's current position $\mathbf{x}^t$ and velocity $\dot{\mathbf{x}}^t$.

energy functions as bases. For instance, in Figure 7.4b, by defining generalized coordinates $q_1 = ||\mathbf{x}_1 - \mathbf{x}_2||$ and $q_2 = ||\mathbf{x}_1 - \mathbf{x}_3||$, the overall potential energy can be decomposed into two functions associated with the two springs: $U(\mathbf{q}) = U_1(q_1) + U_2(q_2)$. If the two springs have the same property, then the potential energy can be further simplified by reusing the same atomic function: $U(\mathbf{q}) = U(q_1) + U(q_2)$.

To enforce sparsity, we assume a polynomial form for each potential function. Specifically, we consider a potential function such as $U_j(q_j) = \mathbf{w}_j^\top [q_j^{-1}, 1, q_j, q_j^2]$, where $\mathbf{w}_j$ are parameters of the polynomial function.

When we have multiple atomic potential energy functions in a system, it is often important to identify when each function will be present or effective in terms of yielding forces to the entities. Some potential energy functions like the ones in Figure 7.4 are always effective. But there are also functions with limited effective spatial ranges. For instance, to approximate the force an entity receives when bouncing off a wall (here we assume perfectly elastic collision) as shown in Figure 7.5, we can imagine that when the entity is expected to violate the non-overlapping constraint (the distance between the entity and the wall can not be smaller than a threshold) in a very short period of time (Δt) based on its current position and velocity, there will be an effective potential energy function applied to the entity. In fact, this potential can be approximated by a spring (with a very large constant $k \gg 1$ and a equilibrium length of distance threshold) connecting the contact point and the entity. This type of approximation has been previously introduced in robotics literature as well [EB15].



(a) (b)

Figure 7.6: Illustration of social concepts as generalized coordinates. (a) An example of generalized coordinates in social systems. The (q_1, q_2, q_3) here are potentially the most critical variables in describing this social system. q_1 and q_3 here reveal the potential goal (i.e., the door) for both agents, so an attraction potential term could explain the behavior of "leaving the room". q_2 can represent the relation between the agents. E.g., the "chasing" behavior could be modeled by a potential term that only depends on q_2. (b) The generalization of (a) where the generalized coordinates and the potential energy function can be preserved; we only need to modify the transformation from raw observations to the generalized coordinates.

If we denote $\delta_j(q_j)$ to be the triggering condition function, then we may define the complete potential energy as

$$U(\mathbf{q}) = \sum_{j=1}^{D} \delta_j(q_j) U_j(q_j). \tag{7.9}$$

7.3.4 Goal-oriented Potentials for Social Behaviors

In our joint simulation engine, everything is generated in a physics engine. It is natural to derive the generalized coordinates and the corresponding potentials regardless of whether an entity is an object or an agent. Consequently, similar ideas discussed for modeling physical systems may also be applied to modeling the goals and relations in social behaviors as illustrated in Figure 7.6a. Suppose an agent with free will can exert self-propelled forces to purse its goal. Then its plan or policy w.r.t. a certain goal can be represented as the force exerted by itself given its current state and the context. By assuming rationality of the agent's plan, the force should be explained by certain potential function associated with its goal and its relations with the environment and other

agents, which can be seen as a form of social potential energy defined on semantically meaningful measurements (i.e., generalized coordinates) such as the distance between its current position and its goal position, or the relative spatial displacement between itself and other agents. By seeking the simplest generalized coordinates and the corresponding sparse functions of potential energy, important concepts in social behaviors, such as goals and relations could naturally emerge as well.

With this analogy, the Cartesian coordinates $(\mathbf{x}_i)_{i=1}^N$ coupled with the context c are the states of the agents, and the generalized coordinates q_j are equivalent to the sufficient statistics in describing the observed social scenario. Let the agents' goals be $g_i \in \mathcal{G}$, where $\mathcal{G}$ is a set of all possible goals, then an agent's behavior is guided by a potential energy function defined in Cartesian coordinates, i.e., $U_i(\mathbf{x}_1, \cdots, \mathbf{x}_N, G, c)$. We then use a potential energy function defined in generalized coordinates to equivalently represent the goal directed potential for agent i as follows

$$U_i(\mathbf{q}, g_i) = U_i(\mathbf{x}_1, \cdots, \mathbf{x}_N, G, c). \tag{7.10}$$

Let $X = \{\mathbf{x}_i\}_{i=1}^N$. The plan of agent i can be derived in a step-by-step manner by Eq. 7.8, i.e.,

$$\mathbf{F}_i(\mathbf{x}_i | X_{-i}, g_i) = -\sum_{j=1}^{D} \frac{\partial U(\mathbf{q}, g_i)}{\partial q_j} \frac{\partial \phi_j}{\partial \mathbf{x}_i}, \quad \forall i = 1, \cdots, N. \tag{7.11}$$

For instance, in Figure 7.6a, if the red agent tries to leave the room, then its motion will be driven by potential $U(q_1)$. Similarly, if the green agent aims to catch the red agent, then it is driven by a potential $U(q_2)$.

Thus, learning sparse potential energy functions through generalized coordinates takes a straightforward approach in explaining the rational behaviors demonstrated by the agents since it allows us to derive the optimal policy directly from the inferred potential energy in addition to discovering the goals. This method may also help us discover sub-goals (i.e., different potential energy terms) in the optimal plans. Finally, the explicit modeling of generalized coordinates can potentially improve the generalization of the learned optimal plans as well since we can simply remap any new environment to the same coordinate system by only changing $\phi_j(\cdot)$; the previously learned potential energy functions and the corresponding optimal plans can be preserved. For instance, the generalized coordinates and potential energy functions constructed based on the environment in

Figure 7.6a can be transferred to the new scenario in Figure 7.6b where the new position of the door will only affect the coordinate transformation for q_1 and q_3.

In summary, we aim to learn the following concepts by constructing generalized coordinates and the corresponding potential energy terms:

- Discovery of meaningful **goals** and **relations** through the generalized coordinates;
- Deriving optimal plans directly through the learned potential energy (with recursive reasoning when involving the anticipation of other agents' moves).

7.3.5 Summary of Main Advantages

We summarize the main advantages of constructing generalized coordinates and the corresponding potential energy functions as follows:

- **Generalized coordinates as effective representations of a system**. The change in $\mathbf{q}$ are the effective change of a system, i.e., $\partial U(\mathbf{q})/\partial \mathbf{q}$. By pursing the coordinates that results in the simplest $U(\mathbf{q})$, we are essentially pursuing a sparse model for the system. For physical systems, such representations will reveal physical concepts, whereas in social systems, they may denote important concepts of goals and social relations.
- **"Compression" of optimal planning**. Optimal planning is complex and time consuming. However, given demonstrations (observed trajectories of agents), we may compress these optimal plans into a few potential energy functions. Consequently, instead of searching for an optimal plan from scratch every time, we may derive forces from the potential energy functions and roll out the whole plan step-by-step starting from the initial state. We may deploy this plan directly, or use it as a starting point and further refine it to compensate the errors in the learned potential energy functions. Similarly, we can also take advantage of the derived forces to conduct inverse planning for Bayesian goal inference.
- **Knowledge transfer**. When the surrounding environment changes, the potential energy defined on generalized coordinates, $U(\mathbf{q})$, may be preserved. In order to derive forces for

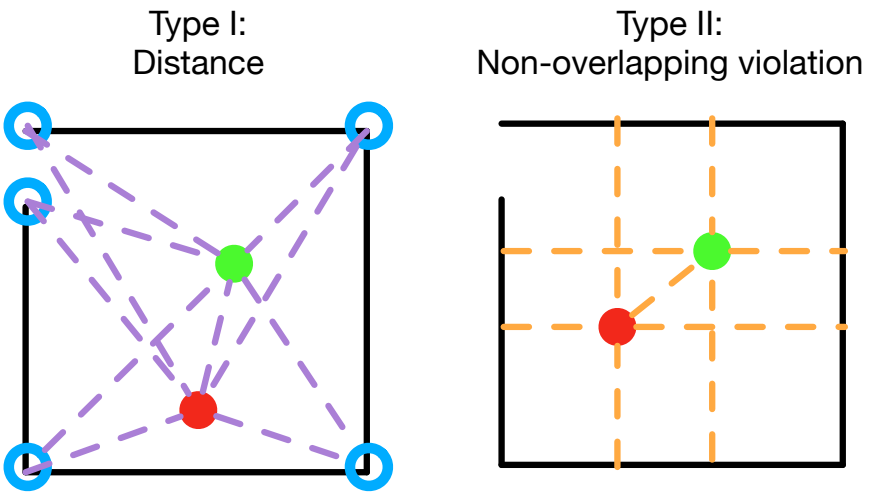


Figure 7.7: Two types of candidates of generalized coordinates shown as the purple and orange dashed lines respectively. The blue circles highlight the reference points used for extracting the first type of candidate coordinates.

the entities in the new environment, we only need to change the coordinate transformations, i.e., $q_j = \phi_j(\mathbf{x}_1, \cdots, \mathbf{x}_N, c)$.

7.3.6 A Sketch of the Learning Algorithm

Problem setup. In an N-entity system, we may observe the context (environment) c, and the trajectories of all entities $\Gamma_i = \{(\mathbf{x}_i^t, \dot{\mathbf{x}}_i^t)\}_{t=1}^T$, where the length of each step is Δt, and the total length is $T\Delta t$. We assume that all entities have the same mass m and there are only conservative forces in the system. From the trajectories, we may also compute the ground-truth force each agent i receives at time step t, i.e., $\mathbf{F}_i^t$. The goal is to learn a model (generalized coordinates and potential energy functions) which can predict the forces given the observations.

Proposals of generalized coordinates. From bottom-up proposals, we obtain a pool of candidates for generalized coordinates, $\mathbb{Q} = \{q_j\}_{j=1}^D$. Note that many of them may be redundant and will not be selected by the final model. In particular, these candidates can arise from two types of proposals:

i) Distance between two geometric shapes. As shown in Figure 7.7, this can be the distance

between two entities (e.g., the one in Figure 7.4) or the distance between an entity and a part of the context (e.g., the one in Figure 7.5). The corresponding potential energy functions are always triggered, i.e., $\delta_j(q_j) = 1$.

ii) Expected constraint violation as illustrated in Figure 7.5. When there is violation, q_j represents the expected overlapped length; otherwise $q_j = 0$. The triggering condition is consequently defined as $\delta_j(q_j) = \mathbb{1}(q_j > 0)$.

Note that for social behaviors, we do not consider the second type of the generalized coordinates.

Pursuing a set of atomic potential energy functions. The final potential energy function consists of a set of atomic potential energy functions, each of which is defined as $U_k(q_k)$, $k \in \mathbb{S} \subset \mathbb{Q}$, where $\mathbb{S}$ is a set of generalized coordinates selected from the candidate pool $\mathbb{Q}$. The final potential energy will be used for predicting the forces for each entity:

$$\hat{\mathbf{F}}_i^t = -\sum_{k \in \mathbb{S}} \delta_k(q_k^t) \frac{\partial U_k(q_k^t)}{\partial q_k^t} \frac{\partial \phi_k^t}{\partial \mathbf{x}_i^t}. \tag{7.12}$$

Finally, we define an MSE loss for the force prediction as the learning objective function:

$$L(\mathbb{S}, \Omega = (w_k)_{k=1}^K) = \mathbb{E}\left[\frac{1}{2}||\mathbf{F}_i^t - \hat{\mathbf{F}}_i^t||_2^2\right]. \tag{7.13}$$

The pursuit of the final model is essentially the search of the optimal generalized coordinates $\mathcal{S}$ and the parameters Ω of the corresponding potential energy functions that minimize the above loss (along with some regularization for sparsity). For computational efficiency, we adopt a greedy pursuit, where we start from an empty set of generalized coordinates, then at each iteration, we augment the final model with the candidate generalized coordinate that has not yet been selected in previous iterations and yields a fitted potential energy function with the largest loss reduction. The iterative pursuit is repeated until there is no significant loss reduction anymore.

7.3.7 Learning Results

We generated collision and spring (with several different spring lengths) physical systems shown in Figure 7.2, each had 50 videos as training examples. Figure 7.8 shows the learning process of

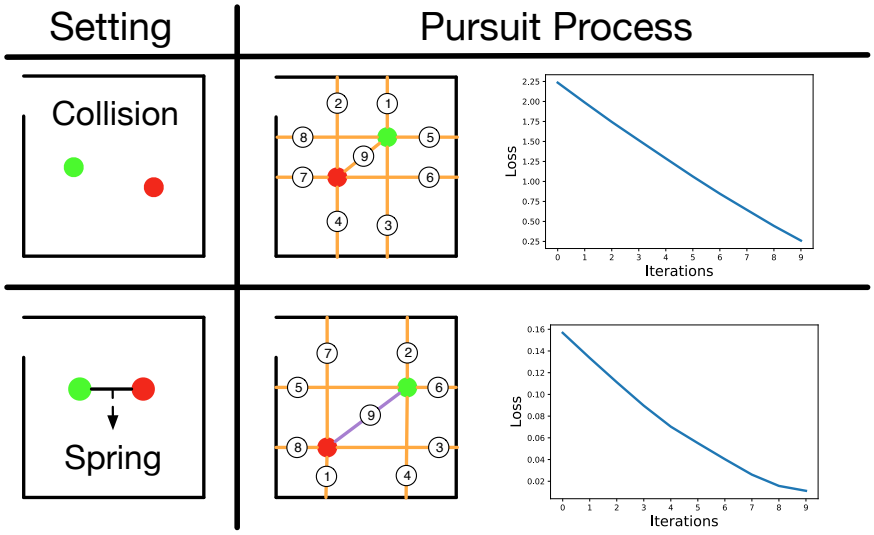


Figure 7.8: Learning process of two physical systems. The purple and orange lines are the selected generalized coordinates from the first and the second type of candidates respectively; each number indicates the iteration when the corresponding generalized coordinate was selected.

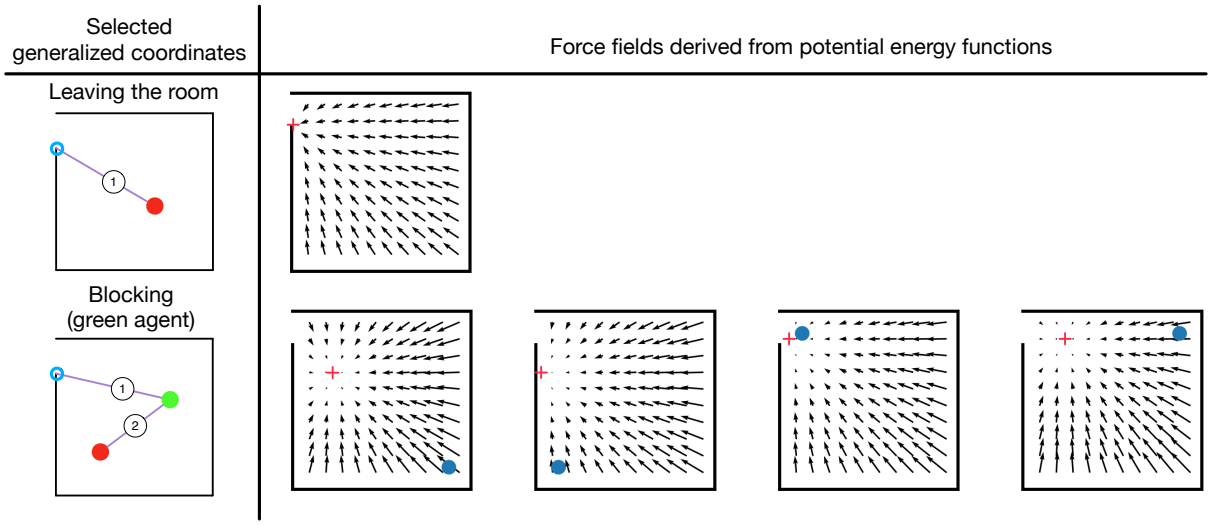


Figure 7.9: Learning results of two goals. Left: selected generalized coordinates; right: force fields derived from the learned potential energy functions, where the blue circle represents the position of the other agent, and the red cross shows the location with the lowest potential energy in the current field.

two systems.

We also used the same approach to pursue potential energy functions for two goals depicted in

Figure 7.2 for HH videos. In practice, we used 45 videos of an agent fleeing the room successfully to learn the potential energy functions for the goal of "leaving the room", and used another 45 videos of an agent successfully blocking another agent or attempting to block it without success for the goal of "blocking". Figure 7.9 shows generalized coordinates and the derived forces fields based on the learned model for both goals. We find that using Lasso can help discover more meaningful goal-directed potentials for social behaviors by enforcing sparsity for the potential energy function of each generalized coordinate.

7.3.8 Physics Inference

By giving the positions and velocities of the two entities at time t, i.e., $\mathbf{x}_i^t$, $\dot{\mathbf{x}}_i^t$, $i = 1, 2$, we can predict the physical forces each entity receives at t and consequently their future velocities at $t+1$, $\hat{\dot{\mathbf{x}}}_i^{t+1}$, $i = 1, 2$. By comparing with the ground truth $\dot{\mathbf{x}}_i^{t+1}$, we can evaluate to what degree an entity's motion is inconsistent with physics predictions:

$$\mathcal{D}_i = \frac{1}{T} \sum_{t=1}^{T} ||\dot{\mathbf{x}}_i^t - \hat{\dot{\mathbf{x}}}_i^t||_2^2, \quad \forall i = 1, 2. \tag{7.14}$$

In practice, there are multiple physical systems, each of which will give different predictions. Since we do not know which system an observation belongs to, we can enumerate all learned physical systems and select the one that yields the lowest prediction error, which we may use as the physical violation measurement.

7.3.9 Intention Inference

The force fields illustrated in Figure 7.9 give us the expected moving direction at each location given the goal of the agent and the position of the other agent. Inspired by the classic FRAME model [ZWM98, XHZ15] which was originally used for modeling texture and natural images, we may treat a field derived from our learned model as filters of motion for a given goal at different locations. The basic idea is illustrated in Figure 7.10. Specifically, the filter response at location $\mathbf{x}_i$

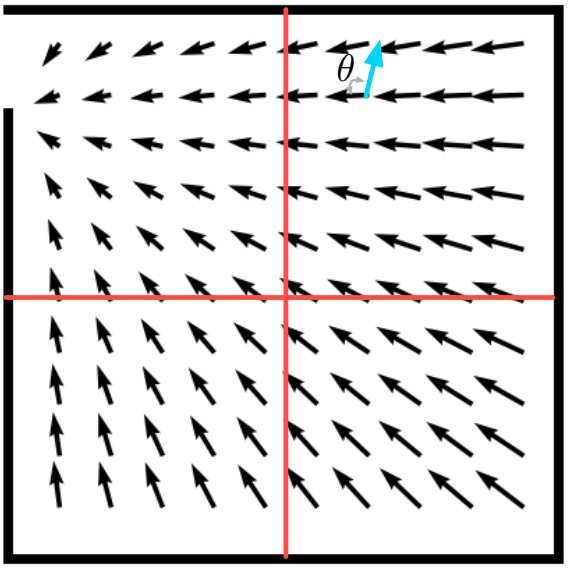


Figure 7.10: Illustration of the idea of motion filters. Suppose the blue arrow is the observed velocity of an agent at a given moment, then we may use the angle θ between to measure the fitness of the observed motion and the expected goal-directed motion (i.e., using $cos(\theta)$ as the filter response). We divide the space into four regions to compute the likelihood of an agent is pursuing a specific goal.

for agent i with goal g_i and the other agent being at $\mathbf{x}_j$ can be defined as

$$h(\dot{\mathbf{x}}_i|\mathbf{x}_i, \mathbf{x}_j, g_i) = cos(\theta) = \frac{\hat{\mathbf{F}}_i(\mathbf{x}_i|\mathbf{x}_j, g_i)^\top \dot{\mathbf{x}}_i}{||\hat{\mathbf{F}}_i(\mathbf{x}_i|\mathbf{x}_j, g_i)|| \cdot ||\dot{\mathbf{x}}_i||}, \tag{7.15}$$

where θ is the angle between the observed moving direction $\dot{\mathbf{x}}_i$ and the expected moving direction from the predicted force $\hat{\mathbf{F}}_i$ in Eq. 7.11. By dividing the whole space into R discrete regions ($R = 4$ in this work), where each region has a location set $\mathbb{X}_r$, we can define the likelihood of observing an agent with a goal having a certain trajectory Γ_i as

$$p(\Gamma_i|g_i, \Gamma_j) = \frac{1}{Z(\Lambda)} \exp\left\{\frac{1}{T}\sum_{t=1}^{T}\sum_{r=1}^{R} \mathbb{1}(\mathbf{x}_i^t \in \mathbb{X}_r)\lambda_r h(\dot{\mathbf{x}}_i^t|\mathbf{x}_i^t, \mathbf{x}_j^t, g_i)\right\} q(\Gamma), \tag{7.16}$$

where $q(\Gamma_i) = \prod_{t=1}^{T} q(\dot{\mathbf{x}}_i^t)$ is a background model for all moving directions without pursuing a specific goal (we assume a uniform distribution for $q(\dot{\mathbf{x}}_i^t)$), $\Lambda = (\lambda_1, \cdots, \lambda_R)$ is the parameter for the likelihood corresponding to the R regions, and $Z(\Lambda)$ is the normalization term. We may write

$Z(\Lambda)$ as

$$Z(\Lambda) = E_{q(\Gamma)}\left[\exp\left\{\frac{1}{T}\sum_{t=1}^{T}\sum_{r=1}^{R}\mathbb{1}(\mathbf{x}_i^t \in \mathbb{X}_r)\lambda_r h(\dot{\mathbf{x}}_i^t|\mathbf{x}_i^t, \mathbf{x}_j^t, g_i)\right\}\right]. \tag{7.17}$$

Since we assume a uniform distribution for the background velocity, it is easy to show that $Z(\Lambda) = 1$. Then parameter λ_r in the likelihood can be estimated as the every filter responses of trajectories in training examples in region r. Finally, we define the intention measurement as the log-likelihood ratio of a trajectory following the optimal plan for pursuing *any* goal over the background trajectory model:

$$\mathcal{L}_i = \max_{g\in\mathcal{G}} \log p(\Gamma_i|g, \Gamma_j) - \log q(\Gamma_i), \quad \forall i = 1, 2. \tag{7.18}$$

7.4 Experiment 1

7.4.1 Participants

30 participants (mean age = 20.9; 19 female) were recruited from UCLA Psychology Department Subject Pool. All participants had normal or corrected-to-normal vision. Participants provided written consent via a preliminary online survey in accordance with the UCLA Institutional Review Board and were compensated with course credit.

7.4.2 Stimuli and Procedure

850 videos of Heider-Simmel animations were generated from our synthesis algorithm described above, with 500 HH videos (100 videos for each AD level), 150 HO videos, and 200 OO videos (50 videos for each sub-category). Videos lasted from 1 s to 1.5 s with a frame rate of 20 fps. By setting appropriate initial velocities, the average speeds of dots in OO videos were controlled to be the same as the average speeds of dots in HH with 100% ADs (44 pixel/s). The dataset was split into two equal sets; each contained 250 HH, 75 HO, and 100 OO videos. 15 participants were presented with set 1 and the other 15 participants were presented with set 2.

Stimuli were presented on a 1024×768 monitor with a 60 Hz refresh rate. Participants were given the following instructions: "In the current experiment, imagine that you are working for a

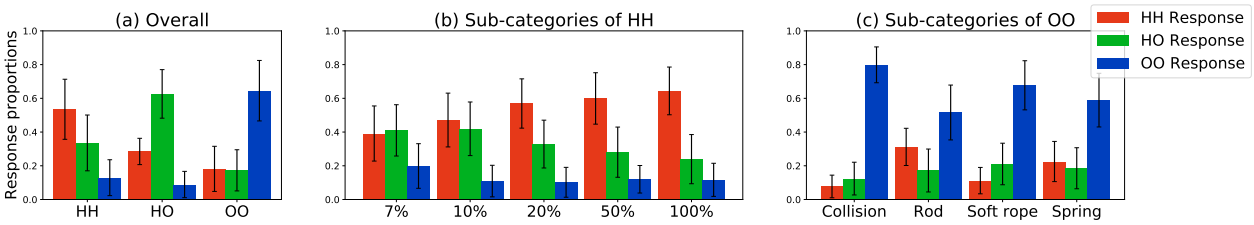


Figure 7.11: Human response proportions of interaction categories (a) and of the sub-categories (b,c) in Experiment 1. Error bars indicate the standard deviations across stimuli.

security company. Videos were recorded by bird's-eye view surveillance cameras. In each video, you will see two dots moving around, one in red and one in green. Your task is to 'identify' these two dots based on their movement. There are three possible scenarios: human-human, human-object, or object-object." Videos were presented in random orders. After the display of each video, participants were asked to classify the video into one of the three categories.

7.4.3 Results

Human response proportions are summarized in Figure 7.11. Response proportion of human-human interaction swas ignificantly greater than the chance level 0.33 ($t(499) = 25.713, p < .001$). For HO animations, response proportion of human-object interaction was significantly greater than the other two responses ($p < .001$). Similarly, response proportion of object-object was greater than the other two responses ($p < .001$) for OO animations. These results reveal that human participants identified the main characteristics of different interaction types based on dot movements.

Next, we examined human responses to the sub-categories within the HH and OO animations. We first used the animacy degree as a continuous variable and tested its effect on human responses in the HH animations. With increases in degree of animacy in HH, the response proportion of human-human interaction increased significantly as revealed by a positive correlation ($r = .42$, $p < .001$). This finding suggests that humans are sensitive to the animacy manipulation in terms of the frequency with which self-propelled forces occurred in the stimuli. For the OO animations, the response proportion for object-object interaction among the four sub-categories yielded significant differences ($F(3, 196) = 34.42$, $p < .001$ by an ANOVA), with the most object-object responses

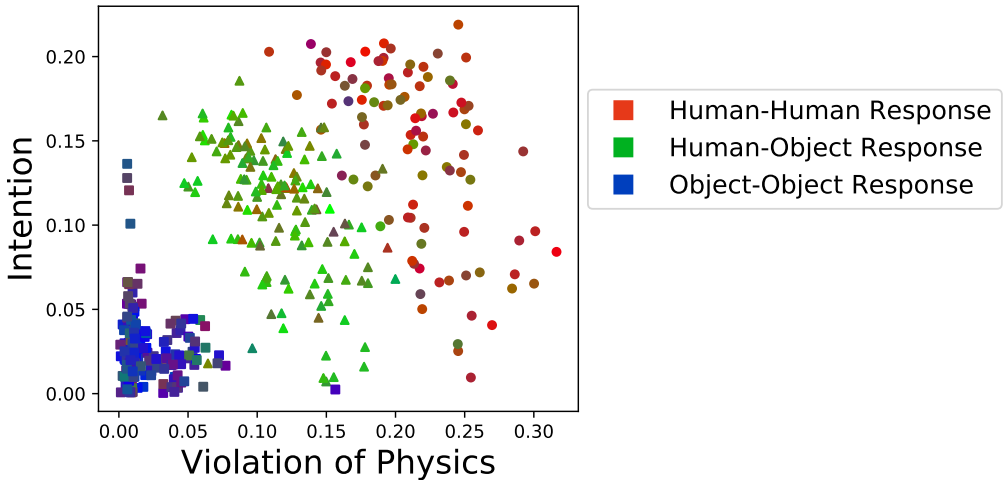


Figure 7.12: Constructed psychological space including HH animations with 100% animacy degree, HO animations, and OO animations. In this figure, a stimulus is depicted by a data point with coordinates derived by the model, and the colors of data points indicate the average human responses of this stimulus. The two coordinates of the space are the averaged measures between the two entities, as the measure of the degree of violation of physical laws (horizontal) and the measure of maximum log-likelihood ratio of goal-directed trajectory over the background model indicating the presence of intention. The mark shapes of data points correspond to the interaction types used in the simulation for generating the corresponding stimuli (circle: HH, triangle: HO, square: OO).

in the collision condition, and the least in the rod condition. Pairwise comparisons among the four-categories show significant difference between collision and everything else ($p < .001$), between soft rope and rope ($p < .001$), and also between soft rope and string ($p = .018$); there is a marginally significant difference between rod and string ($p = .079$).

We then combined human responses and the model-derived measures for each animation stimulus to depict the unified psychology space for the perception of physical and social events. Figure 7.12 presents the distributions of 100 HH videos with 100% animacy degree, 150 HO videos, and 200 OO videos, all in this unified space. In this figure, an animation video is indicated by a data point with coordinates derived by the model, and the colors of data points indicate the average human responses of this stimulus. Specifically, the values of its RGB channels are determined by the average human-human responses in red, human-object responses in green, and object-object

responses in blue. The mark shapes of data points correspond to the interaction type used in the simulation for generating the synthesized animations. The coordinates of each data point were calculated as the model-derived measures averaged across the two entities in an animation, i.e., Eq. (7.14) for physical violation and Eq. (7.18) for the log-likelihood ratio of the trajectory of an entity is driven by a goal. The resulting space showed clear separations between the animations that were judged as three different types of interactions. Animations with more human-human interaction responses (red marks) clustered at the top-right corner, corresponding to great values of intention and strong evidence signaling the violation of physics. Animations with high responses for object-object interactions (blue marks), located at the bottom left of the space, show low values of intention index and little evidence of violation of physics. Animations with high responses for human-object interactions (green marks) fell in the middle of the space.

To quantitatively evaluate how well the model-derived space accounts for human judgments, we trained a classifier using the coordinates derived in the space shown in Figure 7.12 as input features ($\mathcal{D}$ and $\mathcal{L}$ for the indices of physical violation and intention respectively). For each ground-truth type of interactions $y \in \{\text{HH}, \text{HO}, \text{OO}\}$, we fit a 2D Gaussian distribution $p_y(\mathcal{D}, \mathcal{L})$, using half of the stimuli as training data. Then for a given animation with the coordinates of $(\mathcal{D}, \mathcal{L})$, the classifier predicts $p(y|\mathcal{D}, \mathcal{L}) = \frac{p_y(\mathcal{D},\mathcal{L})}{\sum_y p_y(\mathcal{D},\mathcal{L})}$ for animations in the remaining half of the stimuli. The correlation between the model predictions and average human responses was 0.815 ($p < .001$) based on 2-fold cross-validation. Using a split-half reliability method, human participants showed an inter-subject correlation of 0.728 ($p < .001$). Hence, the response correlation between model and humans closely matched inter-subject correlations, suggesting a good fit of the unified space as a generic account of human perception of physical and social events based on movements of simple shapes.

We examined the impact of different degrees of animacy on the perception of social events, and how different subcategories of physical events affect human judgments on interaction types. The unified space provides a platform to compare these fine-grained judgments. Figure 7.13 shows the centers of the coordinates and the average responses for each of the sub-categories. We first found that, with a decreased degree of animacy, the intention index in HH animations was gradually reduced towards the level of HO animations. Meanwhile, human judgments of these stimuli

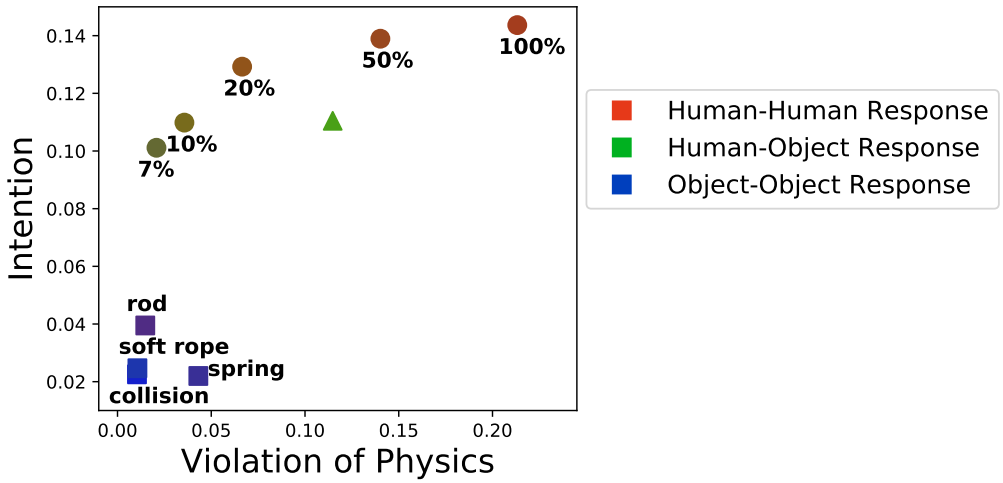


Figure 7.13: Centers of all types of stimuli.

varying from low to high degree of animacy transited gradually from human-object responses to human-human responses, consistent with the trend that the data points moved along the physics axis. Among all physical events, the rod and spring conditions showed the highest intention index and the strongest physical violation, respectively, resulting in a greater portion of human-human interaction responses than the other categories.

7.5 Experiment 2

In Experiment 1, human participants were asked to classify the three interaction types. But for human-object responses, the assignment of the roles to individual entities was not measured. In Experiment 2, we focused on stimuli that elicited the classification of human-object responses, and asked participants to report which dot was a human agent, and which dot was an inanimate object. Specifically, the role assignment in the human-object responses helps us identify some key characteristics in the psychological space that signal a human-object interaction.

7.5.1 Methods

25 participants (mean age = 20.2; 19 female) were recruited from the UCLA Psychology Department Subject Pool. The top 80 HH videos and the top 80 HO videos that got the highest response

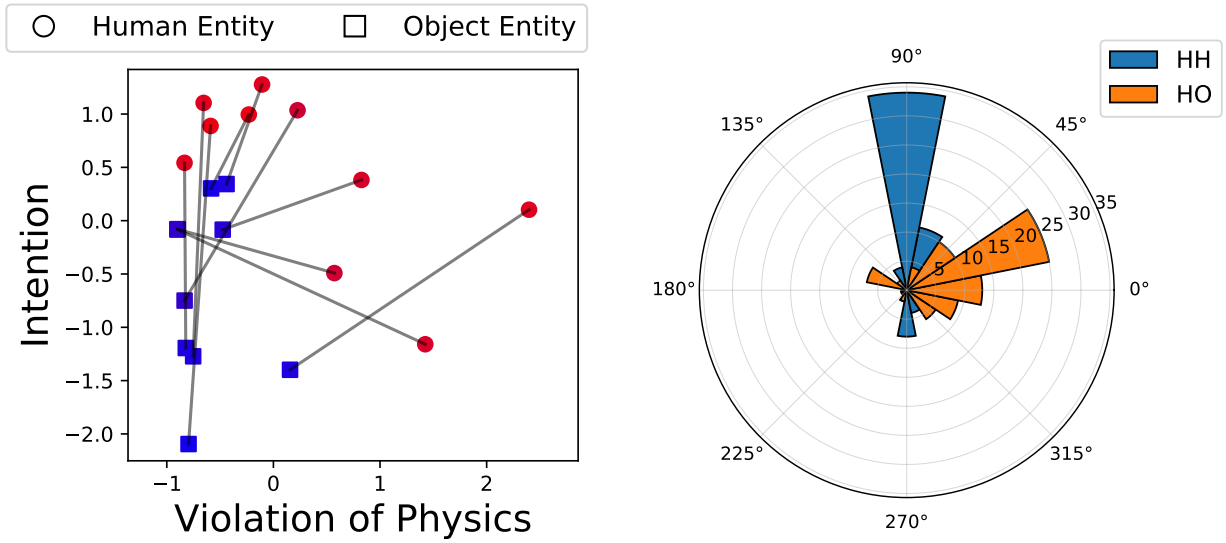
proportions of being judged as HO category in Experiment 1were selected for Experiment 2. The procedure of Experiment 2 was the same as Experiment 1 except that on each trial, subjects were asked to judge among two dots, which dot represented a human agent and which dot represented an object. One dot was red and the other was green and the colors were randomly assigned to the two dots in each trial.

7.5.2 Results

We projected all entities onto the psychological space based on the model-derived measures for each individual entity, and connected a pair of the two entities that appeared in the same video. To make the scale of the two indices directly comparable, each of them was standardized to have a mean of 0 and a standard deviation of 1. We visualized 5 HH animations and 5 HO animations that yielded high human-object response proportions and the most consistent role judgment among participants as shown in Figure 7.14a, where circles represent the dots that were frequently identified as humans, and squares represent the dots identified as objects. The resulting segments showed a common feature in that the human dot has higher degree of physical violation and/or a higher intention measure compared to the object dot. To further examine the orientations in the space for the human-object responses, we calculated the histogram of the orientations for animations judged as human-object interactions, which shows a high concentration around 90 degrees for HH videos and a high conentration between 0 and 45 degrees for HO videos (see Figure 7.14b). This finding suggests that both physical violation and intention contribute to the subjects' role judgment.

7.6 Conclusion

In this chapter, we propose a unified framework for modeling physical and social events from movements of simple shapes in Heider-Simmel animations. We first build a joint physical-social simulation engine and propose a new paradigm for unified physical and social concept learning. Based on the Heider-Simmel animations generated by the simulation engine as well as the metrics for measuring physical violation and impression of intention using the learned computational model, we then construct a unified psychological space to account for human perception of phys-



(a) (b)

Figure 7.14: Human and model-simulation results in Experiment 2. (a) Representative cases of animations that elicited the human-object responses, located in the space with model-derived coordinates. The colors reflects average human responses of assigning a dot to the human role (red) and to the object role (blue). (b) Orientation histogram of the segments connected by the concurrent pairs of entities in an animation.

ical and social events. The space consists of two primary dimensions: the intuitive sense of violation of physics, and the impression of intentions. We tested the space by measuring human responses when viewing a range of synthesized stimuli depicting human-human, human-object, and object-object interactions in the style of Heider-Simmel animations. We found that the constructed physics-intention space revealed clear separations between social and physical events as judged by humans. Furthermore, we trained a classification model based on the coordinates of each stimulus in this space. The resulting model was able to predict human classification responses at the same level as human inter-subject reliability.

The present study provides a proof of concept that the perception of physical events and social events can be integrated within a unified space. Such common representation enables the development of a comprehensive computational model of how humans perceive and reason about physical and social scenes. Perhaps the most surprising finding in our work is that the classification result

based on just the two measures reflecting the violation of physical laws and the estimate of intention can predict human judgment very well, reaching the same level as inter-subject correlation. The good fit to human responses across a range of Heider-Simmel stimuli demonstrates the great potential of using a unified space to study the transition from intuitive physics to social perception.

The main benefit of constructing this psychological space is to provide an intuitive assessment for general impressions of physical and social events. To build up such representation, humans or a computation model may use various cues to detect intentions and/or physical violations; such cue-based detection is usually subjected to personal preferences. Instead of discovering a list of cues for distinguishing between physical events and social events, the proposed space offers an abstract framework for gauging how humans' intuitive senses of physics and intentions interplay in their perception of physical and social events.

This work provides a first step toward developing a unified computational theory to connect human perception and reasoning for both physical and social environments. However, the model has limitations. For example, the simulations are limited by a small set of goals, and the model requires predefined goals and good knowledge about the constrained physical environment. Future work should aim to extend the analysis to a variety of goals in social events [TL14], to develop better goal inference, and to support causal perception in human actions [PTL17]. A more complete model would possess the ability to learn about physical environments based on partial knowledge, and to emulate a theory of mind in order to cope with hierarchical structures in the goal space. In addition, we have only examined human perception of physical and social events on short stimuli with only two entities. Generating longer stimuli with more entities and analyzing human perception on them will further help reveal the mechanisms underlying humans' physical and social perception.

CHAPTER 8

Conclusion

This dissertation aims at addressing three core problems in social scene understanding: group activity parsing, human-robot interactions, and perception of animacy. For each of them, we have proposed new formulations, new representations, and new algorithms for learning and inference. We summarize the key contributions and findings in each problem as follows.

Group activity parsing. We proposed a framework for a joint parsing of groups, events and human roles in group activities from noisy observations. We designed a stochastic grammar model, spatiotemporal AND-OR graph (ST-AOG), as a new hierarchical representation for the underlying structures of group activities in social scenes. For evaluation, an aerial event dataset with rich social behaviors was collected with extensive annotations. Inspired by the recent success of deep neural nets, we also extended the joint inference to structured neural nets to form a deep energy based model. In both cases, experimental results on our aerial video dataset or public group activity recognition benchmarks demonstrate that joint parsing and structured representations greatly can greatly improve model performance on reasoning social concepts from real world videos.

Human-robot interactions. In the first part of the dissertation, we have shown the power of grammar models in representing the knowledge learned from group activities. The nature of such grammar models is to inform a computational model what humans typically do in social interactions, namely social affordances. Hence, in the second part, we formulated social affordances based on this representation (i.e., ST-AOG for human-human interactions), which can be learned from a handful of human demonstrations. We then proposed a real-time motion inference to transfer the social affrodance knowledge from ST-AOG to enable human-robot social interactions. Experiments on simulation and a real Baxter robot show that symbolic plans derived from our ST-AOG are able to capture the essence of a human interaction without over-imitation and

consequently enhance the generalization of the motion transfer in unseen social scenarios.

Perception of Animacy. The final part of the dissertation outlined a unified framework of modeling both physical events and social events based on simple visual input (geometric shapes' motion trajectories), and showed how this framework can account for human perception of both physical events and social events. In particular, under this framework, we i) built a joint physical-social simulation engine, Flatland, to generate Heider-Simmel animations with rich and diverse physical interactions and social behaviors, ii) formulated physical and social concept learning as the pursuit of generalized coordinates and the corresponding parsimonious potential energy functions, iii) constructed a unified psychological space with a dimension of the degree of physical violation and a dimension of the impression of intention. Results from multiple human experiments suggest that this framework is able to shed lights on how humans' perception of animacy integrates intuitive physics and intuitive psychology and how a computational model can reverse engineer this integration.

We are still far from adequately solving these challenges. However, we hope that this dissertation can provide new insights into social scene understanding and inspire more research in this fascinating area. For the future work, we may further explore the following aspects of social scene understanding:

- Computational models for Theory of Mind. Inferring human mental states is a crucial part of social perception, and humans can infer others' mental states fairly efficiently. Hence, it is essential to build computational models to robustly and efficiently infer human mental states based on limited and noisy observations if we want machines to be able to live, work, and communicate with humans. There has been some prior work on this [BST09, UBM10, RPS18, WLS18], but they were designed to model human minds in very restricted environments. Furthermore, for multi-agent systems, it is also important for a machine agent to build mental models of other agents which may be machines too, since an accurate estimate of other agents' intents, beliefs, and plans can significantly improve multi-agent planning or reinforcement learning [QZ18, SKT18, SXW18, ST19].

- The emergence of social norms. How can we mathematically define social norms? Can we

build a computational framework to explain the emergence of various social norms that have been observed in human societies? These are key questions about understanding human social behaviors at a large scale, which remain unanswered. Besides theoretical values, they also have direct applications. For instance, once we have a better understanding of the emergence of social norms in a computational sense, perhaps we can build a reasonably accurate social simulation engine, just like how we can build realistic physics engines on top of computational models that simulate physical laws.

- Representation learning for complex social scenes. We may represent social scenes by a graph constructed by entities and their relations, by natural languages (e.g., telling a story based on Heider-Simmel animations), by potential energy functions, by stochastic grammars, by Bayesian networks, etc. But what are the best representations for describing complex social scenes? How can we learn such representations and use them as transferable and actionable knowledge for applications such as human-robot interactions, education, and social sciences where understanding social behaviors is critical? These are fundamental questions that we need to answer in order to build interpretable models for social scene understanding.

REFERENCES

[ALT14] Mohamed R. Amer, Peng Lei, and Sinisa Todorovic. "Hirf: Hierarchical random field for collective activity recognition in videos." In *European Conference on Computer Vision (ECCV)*, pp. 572–585, 2014.

[AO14] Borislav Antic and Björn Ommer. "Learning Latent Constituents for Recognition of Group Activities in Video." In *ECCV*, 2014.

[ARA17] Mina Alibeigi, Sadegh Rabiee, and Majid Nili Ahmadabadi. "Inverse Kinematics Based Human Mimicking System using Skeletal Tracking Technology." *JIRS*, **85**:27–45, 2017.

[ARS07] Saad Ali, Vladimir Reilly, and Mubarak Shah. "Motion and Appearance Contexts for Tracking and Re-Acquiring Targets in Aerial Videos." In *CVPR*, 2007.

[ASS12] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. "SLIC Superpixels Compared to State-of-the-art Superpixel Methods." *IEEE TPAMI*, **34**(11):2274–2282, 2012.

[AXZ12] Mohamed R. Amer, Dan Xie, Mingtian Zhao, Sinisa Todorovic, and Song-Chun Zhu. "Cost-Sensitive Top-down/Bottom-up Inference for Multiscale Activity Recognition." In *ECCV*, 2012.

[Bak12] Chris L. Baker. *Bayesian theory of mind: modeling human reasoning about beliefs, desires, goals, and social relations*. PhD thesis, Massachusetts Institute of Technology, 2012.

[BGT08] Chris L. Baker, N. D. Goodman, and J. B. Tenenbaum. "Theory-based social goal inference." In *Proceedings of The Thirtieth Annual Conference of the Cognitive Science Society*, pp. 1447–1452, 2008.

[BL11] Jeroen JA van Boxtel and Hongjing Lu. "Visual search by action category." *Journal of Vision*, **11**(7):19–19, 2011.

[BL12] Jeroen JA van Boxtel and Hongjing Lu. "Signature movements lead to efficient search for threatening actions." *PLoS One*, **7**(5):e37085, 2012.

[BM16] David Belanger and Andrew McCallum. "Structured prediction energy networks." In *International Conference on Machine Learning (ICML)*, 2016.

[BMK92] Diane S Berry, Stephen J Misovich, Kevin J Kean, and Reuben M Baron. "Effects of disruption of structure and motion on perceptions of social causality." *Personality and Social Psychology Bulletin*, **18**(2):237–244, 1992.

[BST09] Chris L. Baker, R. Saxe, and J. B. Tenenbaum. "Action understanding as inverse planning." *Cognition*, **113**(3):329–349, 2009.

[BST11] Chris L. Baker, Rebecca Saxe, and Joshua Tenenbaum. "Bayesian theory of mind: Modeling joint belief-desire attribution." In *Proceedings of The 33rd Annual Conference of the Cognitive Science Society*, 2011.

[BT11] William Brendel and Sinisa Todorovic. "Learning Spatiotemporal Graphs of Human Activities." In *ICCV*, 2011.

[CCP14] W. Choi, Y. W. Chao, C. Pantofaru, and S. Savarese. "Discovering Groups of People in Images." In *ECCV*, 2014.

[Cho15] François Chollet. "Keras." https://github.com/fchollet/keras, 2015.

[CS14] Wongun Choi and SIlvio Savarese. "Understanding Collective Activities of People from Videos." *IEEE TPAMI*, **36**(6):1242–1257, 2014.

[CSS09] Wongun Choi, Khuram Shahid, and Silvio Savarese. "What are they doing? : Collective Activity Classification using Spatio-Temporal Relationship among People." In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 1282–1289, 2009.

[CSY15] Liang-Chieh Chen, Alexander G. Schwing, Alan L. Yuille, and Raquel Urtasun. "Learning deep structured models." In *International Conference on Machine Learning (ICML)*, 2015.

[DDS09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database." In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

[DHK14] Martin Danelljan, Gustav Häger, Fahad Khan, , and Michael Felsberg. "Accurate scale estimation for robust visual tracking." In *British Machine Vision Conference (BMVC)*, 2014.

[DL94] W. H. Dittrich and S. E. Lea. "Visual perception of intentional motion." *Perception*, **23**(3):253–268, 1994.

[DRC05] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. "Behavior recognition via sparse spatio-temporal features." In *In proceedings of IEEE International Conference on Computer Vision Workshops*, pp. 65–72, 2005.

[DVH16] Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori. "Structure Inference Machines: Recurrent Neural Networks for Analyzing Relations in Group Activity Recognition." In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4772–4781, 2016.

[EB15] Marco Gabiccini Edoardo Farnioli and Antonio Bicchi. "Optimal contact force distribution for compliant humanoid robots in whole-body loco-manipulation tasks." In *IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[FHR12] Alireza Fathi, Jessica K. Hodgins, and James M. Rehg. "Social interactions: A first-person perspective." In *CVPR*, 2012.

[Fis50] R. A. Fisher. *Statistical Methods for Research Workers*. London: Oliver and Boyd, 11 edition, 1950.

[FLP15] Katerina Fragkiadaki, Sergey Levine, Felsen Panna, and Jitendra Malik. "Recurrent Network Models for Human Dynamics." In *ICCV*, 2015.

[GCR12] W. Ge, T. R. Collins, and R. B. Ruback. "Vision-based analysis of small groups in pedestrian crowds." *IEEE TPAMI*, **34**(5):1003–1016, 2012.

[Gib79] James J. Gibson. *The ecological approach to visual perception*. Boston: Houghton Mifflin, 1979.

[GMS10] T. Gao, G. McCarthy, and B. J. Scholl. "The wolfpack effect: Perception of animacy irresistibly influences interactive behavior." *Psychological Science*, **21**:1845–1853, 2010.

[GNS09] T. Gao, G. E. Newman, and B. J. Scholl. "The psychophysics of chasing: A case study in the perception of animacy." *Cognitive Psychology*, **59**(2):154–179, 2009.

[GSE11] Abhinav Gupta, Scott Satkin, Alexei A. Efros, and Martial Hebert. "From 3D Scene Geometry to Human Workspace." In *CVPR*, 2011.

[GSS09] Abhinav Gupta, Praveen Srinivasan, Jianbo Shi, and Larry S. Davis. "Understanding Videos, Constructing Plots Learning a Visually Grounded Storyline Model from Annotated Videos." In *CVPR*, 2009.

[HA04] Andrea S. Heberlein and Ralph Adolphs. "Impaired spontaneous anthropomorphizing despite intact perception and social knowledge." *Proceedings of the National Academy of Sciences*, **101**(19):7487–7491, 2004.

[HGD17] Kaiming He, Georgia Gkioxari, Piotr Dollr, and Ross Girshick. "Mask R-CNN." In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[HK14] De-An Huang and Kris M. Kitani. "Action-Reaction: Forecasting the Dynamics of Human Interaction." In *ECCV*, 2014.

[HLV17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. "Densely connected convolutional networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[HQX18] Siyuan Huang, Siyuan Qi, Yinxue Xiao, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. "Cooperative Holistic Scene Understanding: Unifying 3D Object, Layout, and Camera Pose Estimation." In *Advances in Neural Information Processing Systems*, pp. 207–218, 2018.

[HQZ18] Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, and Song-Chun Zhu. "Holistic 3D scene parsing and reconstruction from a single RGB image." In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 187–203, 2018.

[HS44] F. Heider and M. Simmel. "An experimental study of apparent behavior." *American Journal of Psychology*, **57**(2):243–259, 1944.

[HS97] Sepp Hochreiter and Jurgen Schmidhuber. "Long Short-Term Memory." *Neural Computation*, **9**(8):1735–1780, 1997.

[HT18] Wenjia Huang and Demetri Terzopoulos. "Door and doorway etiquette for virtual humans." *IEEE transactions on visualization and computer graphics*, 2018.

[HYV15] Hossein Hajimirsadeghi, Wang Yan, Arash Vahdat, and Greg Mori. "Visual Recognition by Counting Instances: A Multi-Instance Cardinality Potential Kernel." In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2596–2605, 2015.

[HZC13] Christian Hane, Christopher Zach, Andrea Cohen, Roland Angst, and Marc Pollefeys. "Joint 3D scene reconstruction and class segmentation." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 97–104, 2013.

[HZR16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016.

[IKB17] Leyla Isik, Kami Koldewyn, David Beeler, and Nancy Kanwisher. "Perceiving social interactions in the posterior superior temporal sulcus." *Proceedings of the National Academy of Sciences*, **114**(43), 2017.

[IMD16] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. "Hierarchical Deep Temporal Models for Group Activity Recognition." *arXiv preprint*, **arXiv:1607.02643**, 2016.

[IRF13] Yumi Iwashita, M.S. Ryoo, Thomas J. Fuchs, and Curtis Padgett. "Recognizing Humans in Motion: Trajectory-based Aerial Video Analysis." In *BMVC*, 2013.

[JHM17] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017.

[JKF16] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. "DenseCap: Fully Convolutional Localization Networks for Dense Captioning." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[JKS13] Yun Jiang, Hema Koppula, and Ashutosh Saxena. "Hallucinated Humans as the Hidden Context for Labeling 3D Scenes." In *CVPR*, 2013.

[Joh73] Gunnar Johansson. "Visual perception of biological motion and a model for its analysis." *Perception & psychophysics*, **14**(2):201–211, 1973.

[JZS16] Ashesh Jain, Amir R. Zamir, Silvio Savarese, and Ashutosh Saxena. "Structural-RNN: Deep Learning on Spatio-Temporal Graphs." In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5308–5317, 2016.

[Kas81] S. Kassin. "Heider and simmel revisited: Causal attribution and the animated film technique." *Review of Personality and Social Psychology*, **3**:145–169, 1981.

[KC10] Wesley Kerr and Paul Cohen. "Recognizing Behaviors and the Internal State of the Participants." In *Proceedings of IEEE 9th International Conference on Development and Learning*, 2010.

[KGS13] M. Keck, L. Galup, and C. Stauffer. "Real-time tracking of low-resolution vehicles for wide-area persistent surveillance." In *WACV*, 2013.

[KGV83] Scott Kirkpatrick, C. Daniel Gelatt, and Mario P. Vecchi. "Optimization by simulated annealing." *Science*, **220**(4598):671–680, 1983.

[KHH13] Suha Kwak, Bohyung Han, and Joon Hee Han. "Multi-Agent Event Detection: Localization and Role Assignment." In *CVPR*, 2013.

[KHL17] James R Kubricht, Keith J Holyoak, and Hongjing Lu. "Intuitive physics: Current research and controversies." *Trends in cognitive sciences*, **21**(10):749–759, 2017.

[Kin09] Davis E. King. "Dlib-ml: A machine learning toolkit." *Journal of Machine Learning Research*, **10**:1755–1758, 2009.

[KL00] James J. Kuffner and Steven M. LaValle. "RRT-Connect: An Efficient Approach to Single-Query Path Planning." In *ICRA*, 2000.

[KL12] J. Kwon and K. M. Lee. "Wang-Landau Monte Carlo-based Tracking Methods for Abrupt Motions." *IEEE TPAMI*, **35**(4):1011–1024, 2012.

[KLD14] Abhijit Kundu, Yin Li, Frank Dellaert, Fuxin Li, and James M Rehg. "Joint semantic segmentation and 3d reconstruction from monocular video." In *European Conference on Computer Vision*, pp. 703–718. Springer, 2014.

[KRK11] Hedvig Kjellström, Javier Romero, and Danica Kragic. "Visual object-action recognition: Inferring object affordances from human demonstration." *Computer Vision and Image Understanding*, **115**(1):81–90, 2011.

[KS14] Hema Koppula and Ashutosh Saxena. "Physically-grounded spatio-temporal object affordances." In *ECCV*, 2014.

[LCH06] Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, , and Fu Jie Huang. "A Tutorial on Energy-Based Learning." In G. Bakir, T. Hofman, B. Scholkopf, A. Smola, and B. Taskar, editors, *Predicting Structured Data*, pp. 191–246. MIT Press, 2006.

[LCS14] Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. "A Hierarchical Representation for Future Action Prediction." In *ECCV*, 2014.

[LGF16] Adam Lerer, Sam Gross, and Rob Fergus. "Learning physical intuition of block towers by example." In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, 2016.

[LGL09] Liang Lin, Haifeng Gong, Li Li, and Liang Wang. "Semantic event representation and recognition using syntactic attribute graph grammar." *PRL*, **30**(2):180–186, 2009.

[LH05] Yann LeCun and Fu Jie Huang. "Loss Functions for Discriminative Training of Energy-Based Models." In *Artificial Intelligence and Statistics Conference (AISTATS)*, 2005.

[LPZ13] Ruonan Li, Parker Porfilio, and Todd Zickler. "Finding group interactions in social clutter." In *CVPR*, 2013.

[LSF16] Xiaodan Liang, Xiaohui Shen, Jiashi Feng, Liang Lin, and Shuicheng Yan. "Semantic Object Parsing with Graph LSTM." In *European Conference on Computer Vision (ECCV)*, pp. 125–143, 2016.

[LSK13] Kyuhwa Lee, Yanyu Su, Tae-Kyun Kim, and Yiannis Demiris. "A syntactic approach to robot imitation learning using probabilistic activity grammars." *Robotics and Autonomous Systems*, **61**(12):1323–1334, 2013.

[LSM12] Tian Lan, Leonid Sigal, and Greg Mori. "Social Roles in Hierarchical Models for Human Activity Recognition." In *CVPR*, 2012.

[LSS12] Matthias Luber, Luciano Spinello, Jens Silva, and Kai O Arras. "Socially-aware robot navigation: A learning approach Socially-aware robot navigation: A learning approach Socially-aware robot navigaion: A learning approach." In *IROS*, 2012.

[LWS02] Yan Li, Tianshu Wang, and Heung-Yeung Shum. "Motion texture: a two-level statistical model for character motion synthesis." In *SIGGRAPH*, 2002.

[LWY12] Tian Lan, Yang Wang, Weilong Yang, Stephen N. Robinovitch, and Greg Mori. "Discriminative latent models for recognizing contextual group activities." *IEEE TPAMI*, **34**(8):1549–1562, 2012.

[LXG12] Chen Change Loy, Tao Xiang, and Shaogang Gong. "Incremental Activity Modelling in Multiple Disjoint Cameras." *IEEE TPAMI*, **34**(9):1799–1813, 2012.

[LZZ15] Tian Lan, Yuke Zhu, Amir Roshan Zamir, and Silvio Savarese. "Action Recognition by Hierarchical Mid-level Action Elements." In *ICCV*, 2015.

[LZZ17] Xiaobai Liu, Yibiao Zhao, and Song-Chun Zhu. "Single-view 3d scene reconstruction and parsing by attribute grammar." *IEEE transactions on pattern analysis and machine intelligence*, **40**(3):710–725, 2017.

[MBM16] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. "Asynchronous methods for deep reinforcement learning." In *International Conference on Machine Learning (ICML)*, 2016.

[MGK19] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. "The Neuro-Symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision." In *International Conference on Learning Representations*, 2019.

[Mic63] A. E. Michotte. *The perception of causality (T. R. Miles, Trans.)*. London, England: Methuen & Co. (Original work published 1946), 1963.

[MLB08] Luis Montesano, Manuel Lopes, Alexandre Bernardino, and Jose Santos-Victor. "Learning Object Affordances: From Sensory-Motor Coordination to Imitation." *IEEE Transactions on Robotics*, **24**(1):15–26, 2008.

[MMO12] Bogdan Moldovan, Plinio Moreno, Martijn van Otterlo, Jose Santos-Victor, and Luc De Raedt. "Learning Relational Affordance Models for Robots in Multi-Object Manipulation Tasks." In *ICRA*, 2012.

[MSI09] Eric Meisner, Selma Šabanovic, Volkan Isler, Linnda R. Caporael, and Jeff Trinkle. "ShadowPlay: A Generative Model for Nonverbal Human-Robot Interaction." In *HRI*, 2009.

[NZH03] R. Nevatia, T. Zhao, and S. Hongeng. "Hierarchical Language-based Representation of Events in Video Streams." In *IEEE Workshop on Event Mining*, 2003.

[OA16] Billy Okal and Kai O. Arras. "Learning Socially Normative Robot Navigation Behaviors with Bayesian Inverse Reinforcement Learning." In *ICRA*, 2016.

[Oh11] Sangmin Oh et al. "A large-scale benchmark dataset for event recognition in surveillance video." In *CVPR*, 2011.

[OMS10] Omar Oreifej, Ramin Mehran, and Mubarak Shah. "Human identity recognition in aerial images." In *CVPR*, 2010.

[OY85] Keith Oatley and Nicola Yuill. "Perception of personal and interpersonal action in a cartoon film." *British journal of social psychology*, **24**(2):115–124, 1985.

[PA12] T. Pollard and M. Antone. "Detecting and tracking all moving objects in wide-area aerial video." In *CVPR Workshops*, 2012.

[PBC14] Peter C. Pantelis, Chris L. Baker, Steven A. Cholewiak, Kevin Sanik, Ari Weinstein, Chia-Chien Wu, Joshua B. Tenenbaum, and Jacob Feldman. "Inferring the intentional states of autonomous virtual agents." *Cognition*, **130**:360379, 2014.

[PEK14] Alessandro Pieropan, Carl Henrik Ek, and Hedvig Kjellström. "Recognizing object affordances in terms of spatio-temporal object-object relationships." In *Humanoids*, 2014.

[PEK15] Alessandro Pieropan, Carl Henrik Ek, and Hedvig Kjellström. "Functional descriptors for object affordances." In *IROS 2015 Workshop*, 2015.

[PES09] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. "You'll Never Walk Alone: Modeling Social Behavior for Multi-target Tracking." In *ICCV*, 2009.

[PG89] Dennis R. Proffitt and David L. Gilden. "Understanding Natural Dynamics." *Journal of Experimental Psychology: Human Perception and Performance*, **15**(2):384–393, 1989.

[PG14] Jan Prokaj and Medioni Gerard. "Persistent Tracking for Wide Area Aerial Surveillance." In *CVPR*, 2014.

[PR14] Hamed Pirsiavash and Deva Ramanan. "Parsing videos of actions with segmental grammars." In *CVPR*, 2014.

[PSY13] Mingtao Pei, Zhangzhang Si, Benjamin Yao, and Song-Chun Zhu. "Video Event Parsing and Learning with Goal and Intent Prediction." *CVIU*, **117**(10):1369–1383, 2013.

[PTL17] Yujia Peng, Steven Thurman, and Hongjing Lu. "Causal Action: A Fundamental Constraint on Perception and Inference About Body Movements." *Psychological Science*, p. 0956797617697739, 2017.

[PWZ10] Jake Porway, Kristy Wang, and Song-Chun Zhu. "A Hierarchical and Contextual Model for Aerial Image Parsing." *IJCV*, **88**(2):254–283, 2010.

[QZ18] Siyuan Qi and Song-Chun Zhu. "Intent-aware multi-agent reinforcement learning." In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7533–7540. IEEE, 2018.

[RA11] M. S. Ryoo and J. K. Aggarwal. "Stochastic representation and recognition of high-level group activities." *IJCV*, **93**(2):183–200, 2011.

[RBL85] Bernard Rimé, Bernadette Boulanger, Philippe Laubin, Marc Richir, and Kathleen Stroobants. "The perception of interpersonal emotions originated by patterns of movement." *Motivation and emotion*, **9**(3):241–260, 1985.

[RHA16] Vignesh Ramanathan, Jonathan Huang, Sami Abu-El-Haija, Alexander Gorban, Kevin Murphy, and Li Fei-Fei. "Detecting events and key actors in multi-person videos." In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3043–3053, 2016.

[RPS18] Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S.M. Ali Eslami, and Matthew Botvinick. "Machine Theory of Mind." *arXiv preprint arXiv:1802.07740*, 2018.

[RPZ13] Brandon Rothrock, Seyoung Park, and Song-Chun Zhu. "Integrating Grammar and Segmentation for Human Pose Estimation." In *CVPR*, 2013.

[RYF13] Vignesh Ramananthan, Bangpeng Yao, and Li Fei-Fei. "Social Role Discovery in Human Events." In *CVPR*, 2013.

[SAL14] Lei Sun, Haizhou Ai, and Shihong Lao. "Activity Group Localization by Modeling the Relations among Participants." In *ECCV*, 2014.

[SBL16] Junzhu Su, Jeroen JA van Boxtel, and Hongjing Lu. "Social interactions receive priority to conscious perception." *PloS one*, **11**(8):e0160468, 2016.

[SDC15] Weihua Sheng, Jianhao Du, Qi Cheng, Gang Li, Chun Zhu, Meiqin Liu, and Guoqing Xu. "Robot semantic mapping through human activity recognition: A wearable sensing and computing approach." *Robotics and Autonomous Systems*, **68**:47–58, 2015.

[SGS13] Sunando Sengupta, Eric Greveson, Ali Shahrokni, and Philip HS Torr. "Urban 3d semantic modelling using stereo vision." In *2013 IEEE International Conference on robotics and Automation*, pp. 580–585. IEEE, 2013.

[SHJ14] Eran Swears, Anthony Hoogs, Qiang Ji, and Kim Boyer. "Complex Activity Recognition using Granger Constrained DBN (GCDBN) in Sports and Surveillance Video." In *CVPR*, 2014.

[SK12] Adam Sadilek and Henry Kautz. "Location-based reasoning about complex multi-agent behavior." *Journal of Artificial Intelligence Research*, **43**:87–133, 2012.

[SKT18] DJ Strouse, Max Kleiman-Weiner, Josh Tenenbaum, Matt Botvinick, and David J Schwab. "Learning to share and hide intentions using information regularization." In *Advances in Neural Information Processing Systems*, pp. 10249–10259, 2018.

[SMB96] Ken Springer, Jo A Meier, and Diane S Berry. "Nonverbal bases of social perception: Developmental change in sensitivity to patterns of motion that reveal interpersonal events." *Journal of Nonverbal Behavior*, **20**(4):199–211, 1996.

[Sob13] Andrews Sobral. "BGSLibrary: An OpenCV C++ Background Subtraction Library." In *IX Workshop de Visão Computacional (WVC'2013)*, 2013.

[SPF17] Tianmin Shu, Yujia Peng, Lifeng Fan, Hongjing Lu, and Song-Chun Zhu. "Inferring Human Interaction from Motion Trajectories in Aerial Videos." In *39th Annual Meeting of the Cognitive Science Society (CogSci)*, 2017.

[SPF18] Tianmin Shu, Yujia Peng, Lifeng Fan, Hongjing Lu, and Song-Chun Zhu. "Perception of Human Interaction Based on Motion Trajectories: From Aerial Videos to Decontextualized Animations." *Topics in Cognitive Science*, **10**(1):225–241, 2018.

[SRZ16] Tianmin Shu, M. S. Ryoo, and Song-Chun Zhu. "Learning Social Affordance for Human-Robot Interaction." In *IJCAI*, 2016.

[SSS17] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George Van Den Driessche, Thore Graepel, and Demis Hassabis. "Mastering the game of Go without human knowledge." *Nature*, **550**(7676):354–359, 2017.

[ST00] B. J. Scholl and R. D. Tremoulet. "Perceptual causality and animacy." *Trends in Cognitive Sciences*, **4**(8):299–309, 2000.

[ST19] Tianmin Shu and Yuandong Tian. "M^{3}RL: Mind-aware Multi-agent Management Reinforcement Learning." In *7th International Conference on Learning Representations (ICLR)*, 2019.

[STC16] Tianmin Shu, Steven Thurman, Dawn Chen, Song-Chun Zhu, and Hongjing Lu. "Critical Features of Joint Actions that Signal Human Interaction." In *CogSci*, 2016.

[SV08] Glenn Shafer and Vladimir Vovk. "A tutorial on conformal prediction." *Journal of Machine Learning Research*, **9**:371–421, 2008.

[SXR15] Tianmin Shu, Dan Xie, Brandon Rothrock, Sinisa Todorovic, and Song-Chun Zhu. "Joint inference of groups, events and human roles in aerial videos." In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4576–4584, 2015.

[SXW18] Tianmin Shu, Caiming Xiong, Ying Nian Wu, and Song-Chun Zhu. "Interactive Agent Modeling by Learning to Probe." *arXiv preprint arXiv:1810.00510*, 2018.

[SZ14] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *arXiv preprint*, **arXiv:1409.1556**, 2014.

[TF00] P. D. Tremoulet and J. Feldman. "Perception of animacy from the motion of a single object." *Perception*, **29**(8):943–951, 2000.

[TF06] P. D. Tremoulet and J. Feldman. "The influence of spatial context and the role of intentionality in the interpretation of animacy from motion." *Perception & Pyschophysics*, **68**(6):1047–1058, 2006.

[TH] Tijmen Tieleman and Geoffrey Hinton. "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude." Coursera: Neural networks for machine learning.

[The16] Theano Development Team. "Theano: A Python framework for fast computation of mathematical expressions." *arXiv preprint*, **abs:1605.02688**, May 2016.

[THR06] Graham W. Taylor, Geoffrey E. Hinton, and Sam Roweis. "Modeling Human Motion Using Binary Latent Variables." In *NIPS*, 2006.

[TL13] Steven M Thurman and Hongjing Lu. "Physical and biological constraints govern perceived animacy of scrambled human forms." *Psychological science*, **24**(7):1133–1141, 2013.

[TL14] Steven M Thurman and Hongjing Lu. "Perception of social interactions for spatially scrambled biological motion." *PloS one*, **9**(11):e112539, 2014.

[TML14] K. Tu, M. Meng, M. W. Lee, T. E. Choi, and Song-Chun. Zhu. "Joint Video and Text Parsing for Understanding Events and Answering Queires." *IEEE MultiMedia*, **21**(2):42–70, 2014.

[UBM10] T. D. Ullman, C. L. Baker, O. Macindoe, O. Evans, N. Goodman, and J. B. Tenenbaum. "Help or hinder: Bayesian models of social goal inference." In *Proceedings of Advances in Neural Information Processing Systems*, pp. 1874–1882, 2010.

[Ull15] Tomer D. Ullman. *On the Nature and Origin of Intuitive Theories: Learning, Physics and Psychology*. PhD thesis, Massachusetts Institute of Technology, 2015.

[VPR13] Carl Vondrick, Donald Patterson, and Deva Ramanan. "Efficiently scaling up crowdsourced video annotation." *IJCV*, **101**(1):184–204, 2013.

[WFH08] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. "Gaussian Process Dynamical Models for Human Motion." *IEEE TPAMI*, **30**(2):283–298, 2008.

[WKO15] Qifei Wang, Gregorij Kurillo, Ferda Ofli, and Ruzena Bajcsy. "Evaluation of Pose Tracking Accuracy in the First and Second Generations of Microsoft Kinect." In *ICHI*, 2015.

[WLK17] Jiajun Wu, Erika Lu, Pushmeet Kohli, William T. Freeman, and Joshua B. Tenenbaum. "Learning to See Physics via Visual De-animation." In *Advances in neural information processing systems*, 2017.

[WLS18] Ping Wei, Yang Liu, Tianmin Shu, Nanning Zheng, and Song-Chun Zhu. "Where and Why Are They Looking? Jointly Inferring Human Attention and Intentions in Complex Tasks." In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[WLT16] Zhirong Wu, Dahua Lin, and Xiaoou Tang. "Deep Markov Random Field for Image Modeling." In *European Conference on Computer Vision (ECCV)*, pp. 295–312, 2016.

[WN11] Esteban Walker and Amy S. Nowacki. "Understanding Equivalence and Noninferiority Testing." *Journal of General Internal Medicine*, **26**(2):192, 196 2011.

[WYL15] Jiajun Wu, Ilker Yildirim, Joseph J. Lim, Bill Freeman, and Josh Tenenbaum. "Galileo: Perceiving physical object properties by integrating a physics engine with deep learning." In *Advances in neural information processing systems*, 2015.

[WZS15] Chenxia Wu, Jiemi Zhang, Silvio Savarese, and Ashutosh Saxena. "Watch-n-Patch: Unsupervised Understanding of Actions and Relations." In *CVPR*, 2015.

[XCS10] Jiangjian Xiao, Hui Cheng, Harpreet Sawhney, and Feng Han. "Vehicle Detection and Tracking in Wide Field-of-View Aerial Video." In *CVPR*, 2010.

[XHZ15] Jianwen Xie, Wenze Hu, Song-Chun Zhu, and Ying Nian Wu. "Learning Sparse FRAME Models for Natural Image Patterns." *International Journal of Computer Vision*, **114**(2-3):91–112, 2015.

[XSX16] Caiming Xiong, Nishant Shukla, Wenlong Xiong, and Song-Chun Zhu. "Robot Learning with a Spatial, Temporal, and Causal And-Or Graph." In *ICRA*, 2016.

[XTZ13] Dan Xie, Sinisa Todorovic, and Song-Chun Zhu. "Inferring "Dark Matter" and "Dark Energy" from Videos." In *ICCV*, 2013.

[XZC17] Danfei Xu, Yuke Zhu, Christopher Choy, and Li Fei-Fei. "Scene Graph Generation by Iterative Message Passing." In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[YHC12] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L. Berg, and Dimitris Samaras. "Two-person Interaction Detection Using Body-Pose Features and Multiple Instance Learning." In *CVPR Workshops*, 2012.

[YLC15] Yezhou Yang, Yi Li, Cornelia, Cornelia Fermuller, and Yiannis Aloimonos. "Robot Learning Manipulation Action Plans by "Watching" Unconstrained Videos from the World Wide Web." In *AAAI*, 2015.

[YO07] Jian Yao and Jean-Marc Odobez. "Multi-Layer Background Subtraction Based on Color and Texture." In *CVPR Workshops*, 2007.

[ZCS18] Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. "LayoutNet: Reconstructing the 3D Room Layout from a Single RGB Image." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[ZFF14] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. "Reasoning About Object Affordances in a Knowledge Base Representation." In *ECCV*, 2014.

[ZHY11] Jiangen Zhang, Wenze Hu, Benjamin Z. Yao, Yongtian Wang, and Song-Chun Zhu. "Inferring social roles in long timespan video sequence." In *ICCV Workshops*, 2011.

[ZJR15] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. "Conditional random fields as recurrent neural networks." In *IEEE International Conference on Computer Vision (ICCV)*, pp. 1529–1537, 2015.

[ZLH17] Chuhang Zou, Zhizhong Li, and Derek Hoiem. "Complete 3D scene parsing from single RGBD image." *arXiv preprint arXiv:1710.09490*, 2017.

[ZML16] Junbo Zhao, Michael Mathieu, and Yann LeCun. "Energy-based Generative Adversarial Network." *arXiv preprint*, **arXiv:1609.03126**, 2016.

[ZWM98] Song-Chun Zhu, Yingnian Wu, and David Mumford. "Filters, Random Fields and Maximum Entropy (FRAME): Towards a Unified Theory for Texture Modeling." *International Journal of Computer Vision*, **27**(2):107126, 1998.

[ZZ13] Yibiao Zhao and Song-Chun Zhu. "Scene parsing by integrating function, geometry and appearance models." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3119–3126, 2013.

[ZZZ15] Yixin Zhu, Yibiao Zhao, and Song-Chun Zhu. "Understanding Tools: Task-Oriented Object Modeling, Learning and Recognition." In *CVPR*, 2015.