ANALYZING HANDS WITH FIRST-PERSON COMPUTER VISION

Sven Bambach

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements for the degree

Doctor of Philosophy

in the School of Informatics and Computing

and the Cognitive Science Program,

Indiana University

September 2016

Accepted by the Graduate Faculty, Indiana University,

in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____

David J. Crandall, Ph.D.

_____

Chen Yu, Ph.D.

_____

Michael S. Ryoo, Ph.D.

_____

Linda B. Smith, Ph.D.

August 22, 2016

Für Mama, Papa†, Lars, und Kim.

ACKNOWLEDGMENTS

When I arrived in the United States five years ago, I did not really know what I was doing and what I had to expect. But I had visited Indiana University before, and I knew that I really wanted to be a graduate student here. Five years later, and I still feel humbled by the trust that this incredible institution has invested in me. I know that not many people get the chance to pursue a doctoral degree and I am beyond grateful that I am given the opportunity to fulfill this dream of mine at this wonderful place. A lot of people helped me get to this point, and I am very grateful to all of them.

First and foremost, I want to thank both of my advisors Prof. David Crandall and Prof. Chen Yu. Their guidance has made such a big impact on my life that it is really hard to imagine that I did not know either of them when I first arrived at IU. I met David when taking his computer vision class during my second semester, and luckily could convince him that it was a good idea to let me work with him. Through a collaboration I then also met Chen and his amazing lab, and over time it became clear where my place was and what I wanted to work on as a graduate student. And these guys let me do it! I cannot imagine advisors that would be more supportive, responsive, and helpful in any aspect of life.

Thank you to Mohammed, Kun, Haipeng, Jingya, Chenyou, Mingze, Eman, Jangwon, and everyone else that I spent time with at the IU Computer Vision Lab. Stefan, of course, deserves a special mention. Thank you for being a great classmate, roommate, colleague, co-author, travel buddy, friend. I always feel inspired by your superb analytical mind and I am really glad that I got to share so many of the adventures of graduate school with you.

Thank you also to everyone in Chen's Computational Cognition and Learning Lab who always had an open mind for my crazy computer vision projects: John, Umay, Linger, Alexa, Lauren, Esther, Yayun, Catalina. A special thanks goes to Seth, Steven, and Charlene for running the lab and providing massive amounts of unique data on a silver platter.

There are many other faculty members at IU that I want to express my gratitude to. Thank you Prof. Linda Smith and Prof. Michael Ryoo for always being supportive of my research and for serving on my committee. Thank you Prof. Paul Purdom for entrusting me with a great fellowship. Thank you Prof. Predrag "Pedja" Radivojac, Prof. Andrew Hanson, Prof. Colin Allen, Prof. Dirk Van Gucht, and many others for teaching great courses that I thoroughly enjoyed. And a giant thank you to the staff at the SOIC, the OIS, the graduate school, and at IU as a whole. You people are outstanding!

Finally, I am very grateful for the financial support I have received. This thesis is based upon work supported in part by the National Science Foundation (NSF) under grants CAREER IIS-1253549 and CNS-0521433, and the National Institutes of Health (NIH) under grants R01 HD074601 and R21 EY017843. Any opinions, findings, and conclusions or recommendations expressed are those of the author and do not necessarily reflect the views of the sponsoring institution. Further support came from the Indiana University Vice President for Research through an IU Collaborative Research Grant. A lot of the work used computing facilities provided by NVidia, the Lilly Endowment through support of the IU Pervasive Technology Institute, and the Indiana METACyt Initiative. Thank you for providing these great resources. It sure is fun to run code on supercomputers!

Sven Bambach

ANALYZING HANDS WITH FIRST-PERSON COMPUTER VISION

Egocentric cameras aim to approximate a person's field of view, which provides insight into how people interact with the world. Consequently, many cognitive researchers are interested in using wearable camera systems as tools to study attention, perception, and learning. These systems typically capture vast amounts of image data, so to fully harness the potential of this novel observational paradigm, sophisticated techniques to automatically annotate and understand the data are needed. However, analyzing first-person imagery introduces many unique challenges, as it is usually recorded passively without artistic intent and therefore lacks many of the clean characteristics of traditional photography.

This thesis presents novel computer vision approaches to automatically analyze first-person imaging data. The focus of these approaches lies in extracting and understanding hands in the egocentric field of view. Hands are almost omnipresent and constitute our primary channel of interaction with the physical world. To that end, we argue that analyzing hands is an important factor towards the goal of automatically understanding human behavior from egocentric images and videos. We propose three different approaches that aim to extract meaningful and useful information about hands in the context of social interactions. First, we consider laboratory videos of joint toy play between infants and parents, and develop a method to track and, importantly, distinguish hands based on spatial constraints imposed by the egocentric paradigm. This method allows us to collect fine-grained hand appearance statistics that contribute new evidence towards how infants and their parents coordinate attention through eye-hand coordination. Next, we build upon this approach to develop a general, probabilistic framework that jointly models temporal and spatial biases of hand locations. We demonstrate that this approach achieves notable

results in disambiguating hands even when combined with noisy initial detections that may occur in naturalistic videos. Finally, we ask to what extent we can identify hand types and poses directly based on visual appearances. We collect a large-scale egocentric video dataset with pixel-level hand annotations to permit the training of data-driven recognition models like convolutional neural networks. Results indicate that not only can we distinguish hands, but also infer activities from hand poses.

David J. Crandall, Ph.D.

Chen Yu, Ph.D.

Michael S. Ryoo, Ph.D.

Linda B. Smith, Ph.D.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 THESIS OVERVIEW

Wearable technology is becoming part of everyday life, from smart watches to activity trackers and even head-mounted displays. These devices are often equipped with advanced sensors that gather data about the wearer and the environment in the context of everyday life, which make them interesting to researchers across many domains [74, 109, 116]. A key part of this wearable technology trend is first-person (egocentric) camera technology that aims to approximate a person's field of view (see Figure 1.1 for examples), and thereby provides dynamic insight into how people visually perceive the world while naturally interacting with it. The embodied nature of this paradigm is of special interest to a rising number of cognitive researchers who use wearable camera systems as tools to study human attention [61], perception [39, 101], and learning [121].

Typically these systems capture vast amounts of image or video data, so in order to fully harness the potential of this novel data collection paradigm, sophisticated techniques to automatically annotate, segment, and generally understand the data are needed. However, analyzing first-person imagery is often complicated by many unique challenges from a computer vision perspective. Egocentric data is usually recorded passively or unintentionally (i.e. without artistic intent) and therefore lacks many of the clean characteristics of traditional photography. Recently, the computer vision community has started to acknowl-

edge these new, domain-specific challenges with many workshops at top tier conferences encouraging research in this new domain of "egocentric vision."

This dissertation presents multiple approaches to automatically analyze imaging data recorded with first-person camera devices. While these approaches constitute novel contributions in the area of computer vision and are thus interesting in their own right, they are particularly motivated by the desire to enable and enhance cognitive research in areas like visual attention and perception. The distinct focus of the presented research lies in extracting and understanding hands in the field of view. Hands are almost omnipresent in our view and we use our hands as our primary channel of interaction with the physical world, for manipulating objects, sensing the environment, and expressing ourselves to other people. Hands may even play a special role in our cognitive development. Some studies indicate that the visual information that toddlers perceive primarily depends on their own manual actions, and propose the idea of "visual attention through the hands" [123, 124]. Thus, from a cognitive perspective, we argue that focusing efforts on the analysis of hands in first-person image and video data is a worthwhile endeavor as it could benefit many studies that aim to collect data with head-mounted camera systems.

A large amount of well-known work within the egocentric computer vision community recognizes the importance of hands by explicitly modeling them to help with various goals such as first-person activity recognition [30, 33, 67] or gaze prediction [66]. Yet, somewhat surprisingly, relatively little attention has been paid to developing methods that can robustly extract hands in the context of first-person video. Even the small corpus of existing work that directly addresses hand detection does so only in relatively constrained, static scenarios [64,65], most notably lacking the presence of any other people in view. In contrast, the work presented here considers analyzing hands in the context of social interactions, where the egocentric observer actively interacts with a partner. We argue that collecting naturalistic

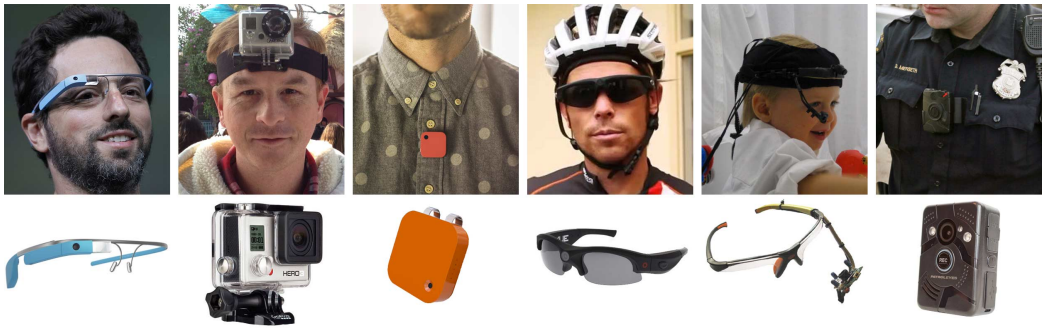


Figure 1.1: *Examples of different first-person camera devices.* From left to right: smart-glasses (e.g. *Google Glass*), compact outdoor cameras (e.g. *GoPro*), life-logging cameras (e.g. *Narrative Clip*), video glasses (e.g. *iVUE*), head-mounted eye tracking systems (e.g. *Positive Science*), and police body cameras (e.g. *PatrolEyes*).

and dynamic data is crucial to foster the development of computer vision methods that work well in many diverse scenarios. Moreover, extracting hands in general may not be sufficient for many applications, and we are the first to address the novel problem of distinguishing hands on a semantic level, i.e. telling apart the observer's left and right hands from the partner's left and right hands.

In the remainder of the introductory chapter, we further motivate the importance of analyzing hands from a cognitive perspective by reviewing the role of hands in guiding attention and modulating perception, both in humans' early cognitive development and as adults. We then review related computer vision work in the domains of first-person vision and hand analysis. Importantly, we highlight key differences between the vast amount of literature on three-dimensional hand pose estimation and our hand analysis in the novel context of egocentric cameras. Next, we provide a brief overview of some of the computer vision models that we build upon later in the thesis, namely probabilistic graphical models (PGMs) and convolutional neural networks (CNNs). Finally, we summarize the motivation and main contributions of the thesis before outlining each of the following chapters individually.

1.2 MOTIVATION FROM A COGNITIVE PERSPECTIVE

Hands are almost omnipresent in our field of view. As they are also our primary tool of physical interaction with the world around us, many interesting actions and object manipulations that are captured by egocentric cameras will prominently include hands. Thus, focusing computer vision efforts on analyzing hands in first-person video could be seen as a purely pragmatic or practical decision. However, a more thorough look at hands and how they relate to human cognition reveals that there is even more virtue to analyzing hands than one might assume. As we will discuss in this section, hands play a special role in early cognitive development by both leading and stabilizing our visual attention [18, 19]. Even as adults, our own hands affect our attention and perception in many interesting ways (e.g., [22, 24]). We argue that considering wearable camera systems as tools that approximate human vision and attention also provides a unique embodied perspective of one's own hands and how they are perceived.

1.2.1 THE ROLE OF HANDS IN EARLY DEVELOPMENT

The human body is a complex system with many degrees of freedom. As infants discover new motor skills within this system, stability and coordination pose profound problems [111]. In early motor development, infants partly solve this problem by using their hands to hold objects near their body's midline, thereby stabilizing trunk and head, and limiting degrees of freedom [8]. For example, Claxton *et al.* [19] found that infants who were just beginning to sit exhibited less postural sway when holding a toy compared to infants that did not hold a toy. Similarly, when just beginning to stand, infants show a lower magnitude of postural sway and more complex sway patterns when holding a toy, suggesting that they adapt postural sway in a manner that facilitates stabilizing and visually fixating on the toy in their hand [18].

Seminal studies that have actually used head-mounted cameras to study visual attention in toddlers (e.g. Smith *et al.* [101], Yu *et al.* [124]) show that the way toddlers perceive objects during free toy play with their parents differs significantly from the parents' perception. While adults tend to have a broad and stable view of multiple objects at the same time, toddlers are more likely to bring single objects very close to their eyes such that they visually dominate the field of view. Thus, much of the visual information that toddlers perceive primarily depends on, and is structured by, their own manual actions. Toddlers use their hands to actively select objects of interest and filter out others, which might be foundational with respect to their visual learning process. A follow-up study by Yu and Smith [122] investigates the role of hands with respect to establishing joint attention between one-year-old infants and their parents. One commonly accepted pathway towards the coordination of looking behavior between social partners is gaze following [37], where one partner follows the other's eye gaze onto the same, joint target. However, the data derived from Yu and Smith's head-mounted eye tracking experiments provide evidence for an alternative pathway, through the coordination of hands and eyes. During joint toy play, infants rarely look to the parent's face and eyes. Instead, infants and parents seem to coordinate looking behavior by following hands to attend to objects held by oneself or the social partner.

Much of the motivation for the work presented in this thesis is based on this series of experiments. Given egocentric videos of toddlers along with eye gaze data (i.e. an estimate of which region of the field of view was overtly attended), locating and identifying the toddler hands and the parent hands in the video could yield fine-grained information on how these joint attention pathways unfold. We address this idea in Chapter 2.

1.2.2 HOW HANDS AFFECT OUR COGNITION

There is a growing body of research that suggests that people's perception of the world depends on their own interactions with it. This means the world within our reach may critically differ from the world beyond our reach [14]. Here, we briefly review some of the work that has investigated the effects of our own hands on visual attention and perception.

Attention

The presence of hands in peripheral vision seems to affect people's attentional prioritization of the space around hands [87]. There are various studies based on static eye tracking with stimuli presented on a computer monitor that show that participants are faster to fixate on target objects when such objects appear near their hand (e.g. Reed *et al.* [86]). Moreover, hands may also shield attention from visual interference. A recent study by Davoli and Brockmole [24] had subjects identify a target letter surrounded by distractor letters, and found that subjects did so more quickly if they flanked the target with their hands (although distractors were still in clear view). Perhaps most interestingly, the special role of hands in guiding attention seems to extend to more dynamic real-world behavior. For example, Li *et al.* [66] used a head-mounted eye tracking system to collect egocentric video data from subjects preparing various meals in a kitchen environment. They built a computational model to predict each subject's gaze location over time based on the video stream, and found that various hand-related features (such as hand motion vectors or hand manipulation points) served as very reliable cues for gaze prediction.

Perception

One's own body in general, and hands in particular, may also play an important scaling role with respect to size perception of nearby objects. Linkenauger *et al.* [69] had subjects

wear size-manipulating goggles while estimating the size of different objects, and found that estimates were more accurate when subjects placed their own hands in their view. Interestingly, this effect did not hold when placing another person's hand in the subject's view instead. Vishton *et al.* [114] found that the effect of the Ebbinghaus illusion (objects appearing bigger/smaller when surrounded by smaller/bigger objects) decreased when subjects were instructed to judge the size of an object before reaching for it. However, observed perceptual changes are not only related to size. For example, Cosman and Vecera [22] found that subjects were more likely to assign regions of an abstract binary visual stimulus as foreground when they placed their hands near them. Such findings are commonly interpreted as evidence that hands may cause a shift from the perception-oriented magnocellular pathway of the visual system to the more action-oriented parvocellular pathway [41, 42].

1.3 RELATED COMPUTER VISION WORK

With the exception of a few pioneering papers from the wearable or ubiquitous computing community [17, 73, 103], the idea of analyzing visual data from first-person cameras is fairly young and arguably closely tied to the recent commercial success of wearable camera devices. The first major attempt within the computer vision community to acknowledge and consolidate efforts in this domain was made at *IEEE Conference on Computer Vision and Pattern Recognition* 2009, which hosted the first workshop on *egocentric (first-person) vision*, establishing the term in the process. Since then, other workshops followed and multiple papers related to egocentric vision appeared in the main proceedings of major computer vision conferences. We review some of the most relevant work in Section 1.3.1. For a near complete survey (until 2014) on first-person vision, including more general work that combines cameras with other body-worn sensors, we refer the interested reader to [11].

Analyzing hands, on the contrary, has a longer computer vision research history. Early

efforts in this domain were driven by the desire to understand hand gestures in the context of human computer interaction (HCI) [28], and the recent abundance of consumer depth cameras (e.g. Kinect) has sparked various interesting advances in depth-based hand pose estimation [107]. As hands are among the most common objects in a person's field of view, researchers are now also turning towards analyzing hands in egocentric images and videos. We put our contributions into perspective by reviewing and contrasting some of this work in Section 1.3.2.

1.3.1 FIRST-PERSON (EGOCENTRIC) VISION

From a computer vision perspective, first-person data has many unique qualities that can be viewed as both challenges and advantages. For example, by approximating the camera wearer's field of view, first-person cameras often implicitly capture the most relevant objects in a scene. At the same time, the lack of artistic control innate to this form of passive (or unintentional) photography often results in low-quality images that suffer from poor illumination and heavy occlusions. Another example is camera motion; in traditional photography and videography, special effort is typically put into minimizing or constraining the motion of the camera. In contrast, body-worn cameras are prone to experience more drastic motion changes, undermining the assumptions of many existing methods of analysis. However, camera motion is directly tied to the body motion of the observer, which in itself could be a useful signal for many objectives.

The objectives that researchers commonly explore in the egocentric domain aim to either take advantage of the domain-specific benefits, or to overcome the domain-specific challenges. Here, we broadly divide these objectives into four categories: *object recognition*, *activity recognition*, *lifelogging analysis*, and *other*.

Object Recognition

Motivated by the idea that recognizing handled objects can provide essential information about the observer's activity, Ren *et al.* [89,90] were the first to explicitly explore the topic of object recognition for egocentric video and provide a benchmark dataset. In particular, they propose a figure-ground segmentation method based on optical flow that isolates the handled object from background clutter [89] and subsequently improves recognition performance. Fathi *et al.* [33] expand this idea by exploring egocentric activities that prominently involve multiple objects (such as making a peanut butter and jelly sandwich). Using multiple instance learning, they take advantage of the frequent co-occurrence of objects in specific activities, and learn how to recognize those objects based on weakly supervised training data that lacks annotations of individual objects.

Activity Recognition

Indeed, one major approach for egocentric activity recognition is object-based recognition, which assumes that the activity of the observer can be characterized by the (handled) objects in view. Consequently, this approach is often explored for common indoor or household activities, or "activities of daily living" [80]. Prominent examples include the work of Pirsiavash and Ramanan [80], who collect a large dataset of 18 daily indoor activities (e.g. washing dishes, watching television), in which 42 different object classes were labeled with bounding boxes, and activities were recognized based on object detection scores. Another line of work is that of Fathi *et al.* [30, 32], who, as mentioned before, consider more fine-grained activities (e.g. preparing a sandwich) that involve multiple objects and can be further divided into a set of labeled actions (e.g. cutting bread). They explore the semantic relationship between objects, activities, and actions to improve recognition among any of these dimensions. Li *et al.* [67] continue work on this dataset and predict actions based on

a set of "egocentric features," including hand locations and head motion.

Analyzing head or body motion of the observer is another major approach for egocentric activity recognition, typically applied when activities or actions are characterized by motion rather than objects (e.g. walking or jumping). Motion features are typically based on the optical flow [45] between adjacent video frames. For example, Kitani *et al.* [54] propose an unsupervised method to segment first-person sport videos (e.g. biking, skiing) into different actions (e.g. hop down, turn left) based on optical flow histograms. Poleg *et al.* [81] use motion cues to segment videos into more general, predefined activities like walking, standing or driving. Ryoo and Matthies [96] explore interaction-level activities from a first-person view, in which other humans directly interact with the egocentric observer. Interactions varied from friendly (e.g. shaking hands) to hostile (e.g. punching or throwing objects), and were classified based on the ego-motion of the observer.

Life Logging Analysis

Egocentric cameras do not necessarily record video. So-called "life logging" cameras are worn by the observer throughout the whole day, but only capture pictures once every few seconds. The purpose of these systems is to record a visual diary that can, for example, serve as a retrospective memory aid for people with memory loss problems [44]. Consequently, the objective of most computer vision work in this domain is either to summarize the captured data in a semantically meaningful way, to retrieve important events, or to detect novelties. Doherty *et al.* [26] were the first to investigate the problem of keyframe selection of life logging data, where they segment one day's worth of images (around 1,900) into events based on different similarity measures, and investigate various quality measures to extract the best keyframe for each event. Lee *et al.* [63] try to go beyond traditional keyframe selection techniques, and summarize videos by regressing importance scores for different

objects and people that the observer interacts with. Lu and Grauman [72] extend this work by developing a story-driven (rather than just object-driven) approach to summarize life logging images, where object scores are combined with temporal coherence scores. Finally, Aghazadeh *et al.* [1] devise a method to detect novelties (e.g. running into a friend) within the context of life logging data captured over multiple weeks that presents some degree of repeating patterns (e.g. a daily commute to work).

Other

There are other egocentric vision objectives that researchers have delved into. For example, Templeman *et al.* [110] address privacy concerns innate to the passive data capturing nature of first-person cameras by automatically detecting blacklisted spaces (e.g. bathrooms, bedrooms) in the captured images, and preventing them from further propagation. Yonetani *et al.* [120] imagine a world where many people passively capture each other with first-person cameras, and develop a method to "ego-surf" such videos, i.e. to automatically find yourself in the videos of others.

Another popular research area is time-lapse videos, which pose a problem for first-person videos, particularly sports videos, as the erratic camera motion gets amplified by the speed-up. Kopf *et al.* [58] overcome this problem by reconstructing the 3D path of the camera, and then rendering a smoother virtual path by blending selected source frames. Poleg *et al.* [82] propose a method which preferably samples frames that are oriented towards the direction of the camera wearer's movement.

Finally, researchers have investigated the problem of localization. For instance, Bettadapura *et al.* [12] registers images (or videos) from a first-person camera to Google Street View data, while Wang *et al.* [115] help orient people in large shopping malls by registering first-person photos to the mall's floor plan.

1.3.2 ANALYZING HANDS

There is a rich and diverse history of computer vision research dedicated to the analysis of hands. Much of this research has been driven by the desire to use hand gestures as a means of human computer interaction [28]. For example, rather than using hands and fingers to operate a mouse, a computer could receive its command input by visually interpreting the movement and pose of hands and fingers directly. As human hands are quite dexterous, hand gestures could indeed be a rich means of communication. However, fully capturing this richness requires visually extracting a very fine-grained, probably 3-dimensional representation of the hand. In contrast, one can imagine simpler applications (like recognizing a waving gesture) that might only require a bounding box detection of the whole hand, or intermediate applications like recognizing the number of raised fingers (to indicate a count), that might require a pixel-wise segmentation of the hand, but not necessarily a 3D representation. These different levels of abstraction have led to a very diverse set of existing literature.

In this section, we briefly survey this literature, starting with seminal research on vision-based hand pose estimation, going over the significant advances in gesture recognition sparked by the emergence of commodity depth sensors, and finally moving towards hand analysis in the egocentric domain. It is worth noting that, despite the large amount of literature, the vast majority of work is still motivated by the human computer interaction paradigm, meaning that many approaches assume a setup where the camera's (or the depth sensor's) sole purpose is to capture a hand. Considering wearable, egocentric cameras instead is therefore not just a minor deviation from existing work, but rather a novel and original perspective on the problem.

Vision-Based Hand Pose Estimation

One can think of hand pose estimation as either a discriminative or generative modeling problem. While discriminative approaches try to classify or regress the observed hand into a finite set of predefined hand pose configurations, generative approaches have an explicit model of a hand (usually with restricted degrees of freedom for each joint) that they try to match to the observation. Many early approaches were discriminative and dealt with the problem of generating enough labeled training exemplars for different hand poses. For example, Athitsos and Sclaroff [5] extract edge features from an input hand image and use Chamfer distance matching to retrieve the closest match from a database of synthetic hand pose images. Wu and Huang [117] use a variation of expectation maximization to train a classifier from a large dataset of 14 different hand poses, of which only a small portion is labeled. Kölsch and Turk [57] experiment with hand detection and hand pose classification by adapting the integral-image based object detector of Viola and Jones [50] to six different hand poses. Some seminal generative modeling approaches include the work of Stenger *et al.* [104], as well as Wu *et al.* [118], but are restricted to modeling hands from one canonical viewpoint [118] or in front of a clean and dark background [104, 118].

Depth-Based Hand Pose Estimation

The availability of commodity depth camera systems (e.g. the Kinect) has spurred many advancements in the area of hand pose estimation. In addition to the 2D image signal, such systems provide a depth map, i.e. a per-pixel estimate of the distance between object point and camera. This is usually achieved by projecting and analyzing an infrared laser grid onto the scene. Considering the per-pixel depth when analyzing images naturally adds robustness with respect to illumination changes for the task of object-background segmentation, and also provides valuable information to fit more fine-grained three-dimensional hand models to

the observed data. Consequently, there is a vast and diverse amount of literature on depth-based hand pose estimation, the taxonomy of which we aim to outline here by selecting some of the most influential recent approaches. We refer to [107] for a more complete survey.

Various data-driven, model-driven (generative), and mixed approaches have been explored for depth-based hand pose estimation. Model-driven approaches are usually applied in the tracking domain, as tracking allows for an initialization pose that helps to constrain the large, non-convex search space of hand joint configurations. For example, Qian *et al.* [84] derive a fast method for detecting fingertips based on local minima in the depth map, and use it to (re-)initialize an iterated closest point optimization method [9] that iteratively updates a joint-based hand model for each subsequent frame.

In contrast, methods that aim at single-image hand pose estimation do not necessarily require real-time performance and can prefer accuracy over processing time. In this context, Oberwerger *et al.* [77] recently proposed a deep learning-based feedback loop system to predict joint-based hand poses from a single depth image. A first neural network is trained to regress hand poses from the input image, while a second network is trained to synthesize the depth image based on the regressed pose. The feedback loop is completed by a third network that compares the input and synthesized image, and produces an updated, refined pose estimate for the synthesizer network. Many approaches for single-image predictions use random decision forests (RDF), which can produce good results at a relatively small computational cost. For example, Keskin *et al.* [53] propose a multi-layer RDF framework to classify single depth images of hands into the American sign language alphabet. Their approach assigns each input depth pixel to hand shape classes, and directs them to corresponding hand pose estimators trained specifically for that hand shape.

Finally, current state-of-the-art real-time hand tracking systems, such as Sharp *et al.* [97], use RDF-based methods to provide (re-)initialization poses for subsequent tracking efforts.

In particular, they aim to propose a set of plausible hand poses with possibly different global hand orientations to allow accurate pose tracking even for camera setups where the global orientation between hand and camera is unconstrained.

Egocentric Approaches

Hands are almost omnipresent in a person's field of view and the first-person perspective creates a very functional and embodied perspective of one's own hands. It is therefore perhaps somewhat surprising that there has been relatively little work on analyzing hands in this domain. One possible reason is that most prevalent first-person camera devices do not include depth sensors and thus implicitly exclude many of the approaches mentioned in the previous section. A notable exception is the work of Rogez *et al.* [92], who recently were the first to transfer the problem of depth-based hand pose estimation to egocentric data, and observe that it is aggravated significantly by the moving depth sensors and self-occlusion of hands and fingers. Consequently, the bulk of existing work and the work presented in this thesis tries to extract hand information directly and only based on 2D image data.

In this context, Ren and Gu [89] and Fathi *et al.* [33] were the first to implicitly consider hands in their attempts to recognize objects held by the egocentric observer. Ren and Gu [89] pose this as a figure-ground segmentation problem, analyzing dense optical flow to partition frames into hands (or held objects) with irregular flow patterns and background with coherent flow. Fathi *et al.* [33] additionally segment between hand and object areas based on superpixels and color-histogram features. Like much pioneering work in a new domain, these approaches make several assumptions to simplify the problem. For instance, both consider video data that includes only one person who carefully manipulates objects in front of a static and rigid scene. Further, as the hand segmentation problem is considered more as a means to an end, the proposed methods make no explicit effort towards

generalizability to different people or scenes.

Li and Kitani [64, 65] were the first to explicitly consider egocentric hand segmentation as a main objective. In particular, they identify drastic changes in scene illumination as the main challenge to overcome. For wearable cameras, such changes are likely to occur while the camera wearer moves between different locations. Li and Kitani propose different models based on a combination of color, texture and gradient features [65], as well as a model recommendation approach [64] that tries to infer the scene illumination and choose the best model accordingly. Some follow-up work aims to make the problem of hand segmentation more suitable for possible consumer applications. Kumar *et al.* [60] learn a color model for each person on-the-fly using a calibration gesture, and focus subsequent action recognition efforts on small image regions around hands to reduce computational costs. Betancourt *et al.* [10] emphasize the difference between hand detection (in the sense of a binary classification for the presence or absence of hands in view) and hand segmentation (in the sense of pixel-level classification), and propose a sequential classifier that only applies computationally expensive segmentation for frames in which hands are assumed to be present in view.

Relationship between this Thesis and Existing Work

The computer vision problems studied in this thesis are novel and distinct from the surveyed work in several important aspects. The main corpus of existing work related to analyzing hands is motivated by the fine-grained, three-dimensional hand pose analysis required for sophisticated human computer interaction. Thus, this work often assumes that locating or identifying hands is trivial due to the hand being the main object in the camera's view, or the depth sensor significantly simplifying the segmentation problem.

In contrast, most wearable cameras are not primarily set up for such fine-grained hand

analysis. However, the egocentric perspective implicitly renders hands as important objects to analyze, as hands are much more likely to be prominently captured by a first-person camera than by other fixed camera setups. The existing work on egocentric hand analysis [64, 65] is primarily concerned with pixel-level segmentation of hands in the face of unconstrained and visually noisy first-person video. While this is an important problem which we also address, the work presented here goes further and aims at a semantic understanding of hands in the egocentric context. For example, as we consider video data that captures interactions with other people, we introduce the problem of hand type classification, where hands must be semantically distinguished not only between left and right, but also between the observer's hands and any other hands in view. Lastly, while we do not perform three-dimensional pose analysis, we do investigate how segmented hand poses in the first-person view relate to the higher-level activity of the first-person observer.

1.4 RELEVANT COMPUTER VISION MODELS

Some of the computer vision models proposed in this thesis are based on more general learning frameworks, such as probabilistic graphical models in Chapter 3, or convolutional neural networks in Chapters 4 and 5. This section aims to briefly introduce the main ideas behind these frameworks.

1.4.1 PROBABILISTIC GRAPHICAL MODELS

Probabilistic graphical models (PGMs) offer convenient frameworks to deal with the inherent uncertainties of noisy observations and have thus been used quite extensively in the field of computer vision. Some of the traditionally most popular examples include stereo vision [35,56], image restoration [35], image segmentation [34], and activity recognition [78]. This section provides a very brief overview about the idea behind graphical models, and

then shows how they can practically be applied to some example vision problems.

Background on Probabilistic Graphical Models

The general idea behind PGMs is to probabilistically model complex systems with many dependent state variables, only some of which may be observable or partly observable. This is achieved by expressing these states as a set of random variables $\mathcal{X}$, where each variable $X \in \mathcal{X}$ is expressed as a node in a graph. Edges between nodes model dependencies between variables such that the whole graph expresses a joint distribution over the space of possible values of all variables in $\mathcal{X}$. If edges are directed and acyclic, such graphs are called Bayesian or belief networks. If edges are undirected (and possibly cyclic), such graphs are referred to as Markov random fields. The goal in either case is usually to ask for the posterior probability distribution of an unobserved variable X_j, given the observation that a set of other variables $\mathcal{X}_i$ takes on values x_i. A simple example of a Bayesian network is given in Figure 1.2. In this case, a doctor might be interested how likely a patient is to have the flu, given that it is spring and he/she has a sinus congestion, but no muscle pain, i.e. $P(Flu = true|Season = spring, Congestion = true, Muscle\_pain = false)$. Notice that the graph breaks up the joint distribution into smaller factors with a smaller space of possibilities by implicitly encoding independencies between variables (e.g. having a sinus congestion is independent from the season given flu and hay fever). This quality allows to answer queries such as $P(Flu = true|\mathcal{X}_i)$ using inference algorithms that work directly on the graph structure and are generally much faster and more practical than manipulating the joint distribution explicitly.

Model Inference

Exact inference on a graphical model (e.g. the MAP estimate for Flu from above) can in principle be done using the variable elimination algorithm. However, in general this

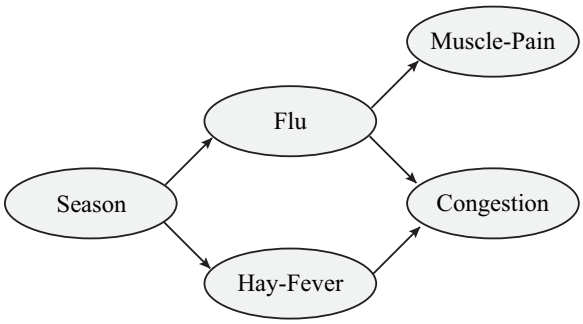


Figure 1.2: A simple Bayesian belief network to model a medical diagnosis setting, showing the dependencies between different variables. The example is borrowed from [55].

algorithm requires exponential time, with the exception of some special cases (e.g. the graph has a polytree structure, meaning there are no loops if we assume all edges are undirected, or the graph is rather small and has a low tree-width). Unfortunately, for many practical cases in computer vision, most graphs have structures that allow only approximate inference. There are a variety of approximate inference algorithms, often customized to be more efficient (or more exact) for certain types of graph structures. One class of algorithms is based on sum-product message passing (also called belief propagation), where the basic idea is that neighboring nodes in the graph iteratively pass messages to each other about their most likely marginal distributions. Another broad class of approximate inference algorithms is based on Monte Carlo methods. Such methods aim to approximate marginal or joint distributions of variables by repeated sampling, which can often be efficient as samples are only required from a small subset of the graph (called the Markov blanket).

Applications in Computer Vision

A very common graph structure for many computer vision problems is a grid, where each pixel of an image is modeled as a node in a grid that connects neighboring pixels. Such a structure could, for instance, enforce that neighboring pixels should likely have a similar depth (e.g. for stereo vision) or belong to the same object class (e.g. for object segmentation). Graphical models are also used in part-based object detection models, where objects

are modeled as a set of parts whose spatial arrangement can vary according to the model. Finally, PGMs can model temporal relationships in order to predict activities shown in a video by enforcing that a certain action a is more likely to be followed or preceded by another action b.

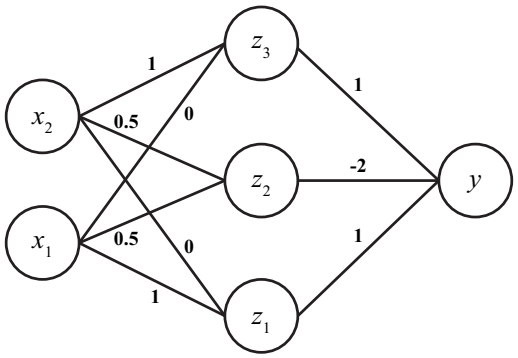
1.4.2 CONVOLUTIONAL NEURAL NETWORKS

In recent years, deep learning, and particularly convolutional neural networks (CNNs), have received great attention and popularity within the computer vision research community. One can find state-of-the-art solutions that rely on such CNNs for almost every interesting vision problem, such as general object recognition [59] and detection [40, 85], semantic segmentation [70], edge or contour detection [119], activity recognition in video [52, 100], optical flow [27], image colorization [16], image captioning [51], etc. Even the famous AlphaGo team that recently developed the first AI to beat a professional human player at the game of Go uses CNNs to analyze the status of the game [99]. In this section we give a brief overview of the background, design, and practical qualities of CNNs.

Background on Neural Networks

CNNs are special types of multi-layer, feed-forward neural networks. Feed-forward neural networks are in essence mathematical models that try to learn a function that maps real-valued n-dimensional input $\mathbf{x}$ to a k-dimensional output $\mathbf{y}$. Input and output are represented as layers of $n+1$ and k neurons respectively, where the additional input neuron can represent a bias term. Inbetween input and output layers can be any number of hidden layers. Every neuron $i \in \{0, \ldots, M\}$ in layer l is connected to every neuron j in layer $l+1$ with a weighted connection $w_{ij}^{(l)}$. The activation at neuron $(l+1)_j$ is determined by the sum of its input and a non-linear activation function $\sigma(.)$, i.e. $a_{(l+1)_j} = \sigma(\sum_{i=0}^{M} w_{ij}^{(l)} a_{l_i})$.

Figure 1.3 shows a simple example of a neural network with one hidden layer, $\mathbf{z}$, that



XOR

x_1	x_2	y
0	0	0
1	0	1
0	1	1
1	1	0

Figure 1.3: A hidden-layer neural network for the XOR function. In this case, the activation function for the z neurons is $\sigma(x) = \{1 \text{ if } x \geq 1,\ 0 \text{ otherwise}\}$. The biases are 0 and therefore not shown.

implements the XOR function (see truth table). In this case, the activation function is $\sigma(x) = \{1 \text{ if } x \geq 1,\ 0 \text{ otherwise}\}$. For example, if $x_1 = 1$ and $x_2 = 1$, then $y = \sigma_1(1 + 0) - 2\sigma_2(0.5 + 0.5) + \sigma_3(1 + 0) = 1 - 2 + 1 = 0$.

It has been shown that, given a sufficient number of hidden neurons and under mild assumptions on the activation function, feed-forward neural networks with only one hidden layer can be universal function approximators [46]. Given this intriguing quality, the question becomes how to define the network parameters (i.e. the weighted connections w) such that the network best approximates the desired function. This is commonly done via the backpropagation method [95]. The core idea is to define a loss function that expresses the error between the current network output and the desired network output, calculate the loss function's gradient with respect to all network weights, and finally update weights in an attempt to minimize the loss. In practice, weights are initialized randomly and then training exemplars are iteratively pushed (forward) through the network. The error of each exemplar is propagated backwards through the network (thus the name backpropagation), gradients are computed at each layer using the chain rule, and weights are updated using gradient descent. This process is repeated until the network's performance is satisfactory.

Convolutional Network Architecture

While a standard network with one hidden layer has the theoretical properties to approximate any function, in practice different network architectures have proven favorable for specific problems. The key idea is to explicitly encode properties of the input data into the structure of the network. For images, we know that nearby pixels are more strongly correlated with each other than more distance ones, and thus we would like to extract features that only rely on small subregions of the image. At the same time, the network should be invariant under small translations of the image content. Both of these qualities are modeled in convolutional neural networks (CNNs). If each pixel in the input image corresponds to one input neuron, rather than fully-connecting each of those neurons to every neuron in the next layer, convolutional networks only connect pixels in a local neighborhood, say 10×10 pixels, to the next layer neuron. The weights of those local connections are shared across the whole image, effectively implementing a convolution for which the filter weights can be learned. Such convolutional layers are usually repeated multiple times, and often combined with sampling/pooling layers to further increase translation invariance.

The idea behind CNNs can be traced back to the late 1980s, when such networks were successfully used for handwritten zip code recognition [62]. However, due to the need for extensive training data and powerful computational infrastructure to train larger networks, it was not until very recently that CNNs found mainstream success. Figure 1.4 shows the well-known network architecture proposed by Krizhevsky *et al.* [59] that in 2012 famously won the ImageNet [25] challenge for large-scale visual image classification. In this case, the input image is resized to a fixed size of 224×224 pixels ($\times$ 3 color channels), and then 48 different kernels of size $11 \times 11 \times 3$ filter the image to create a block of 48 filter responses. Those responses are then subjected to another set of filters, and so on, for a total of five convolutional layers. Intuitively, these layers learn a diverse set of image filters, from low-

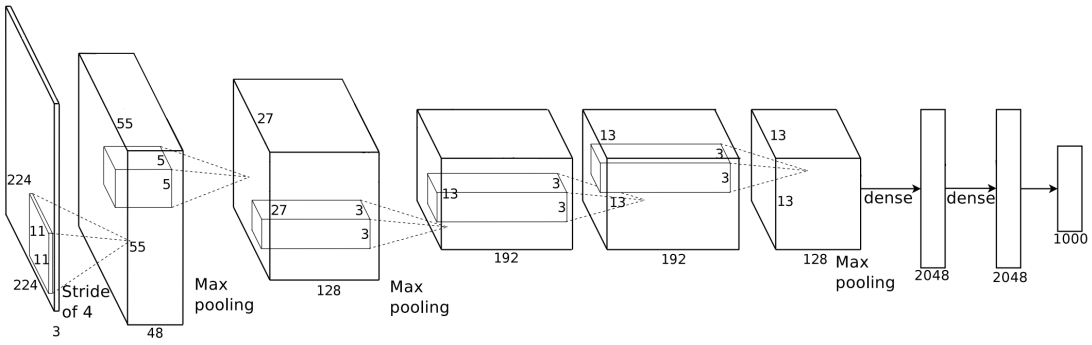


Figure 1.4: A convolutional neural network with the (slightly modified for ease of visualization) architecture of Krizhevsky *et al.* [59].

level edge filters in early layers to high-level semantic filters in later, deeper layers. The convolutional layers are followed by two fully-connected layers that act as a classifier. Here, they map the filter responses of the last convolutional layer to the 1000 visual object classes (e.g. leopard, mushroom, cherry) of the ImageNet [25] challenge.

There are a few other important design decisions that were crucial for the success of this network. For instance, the activation function for the network's neurons is a rectified linear unit (ReLU), i.e. $\sigma(x) = max(0, x)$, which propagates the gradient more directly than other more traditional functions (e.g. $tanh(x)$). Also, as indicated in Figure 1.4, three of the convolutional filters were followed by a max-pooling operation. Max-pooling summarizes the output of neighboring neurons by propagating only the maximum activation within the neighborhood to the next layer, thus further increasing translational invariance.

Since many deep learning software packages (e.g. the Caffe framework [49]) implement each operation as "layers," one often finds CNNs architectures in the literature described in terms of a series of convolutional, ReLU, and pooling layers of certain sizes.

Properties of Deep Networks

For a computer vision researcher, deep neural networks and CNNs in particular have interesting qualities from both a theoretical and practical point of view. From a theoretical

perspective, visual classification problems have traditionally been approached as a two-step process. The first step is to carefully handcraft a visual feature that captures the desired concept (e.g. SIFT [71]), and the second step is to use a machine learning classifier (e.g. support vector machines [21]) within the feature space to distinguish among concepts. In contrast, a CNN such as that shown in Figure 1.4 is trained end-to-end in order to learn the best visual features for the required classification task automatically. It is worth noting, however, that in practice sometimes better classification results are achieved by cutting off the network (after training) before the fully connected layers, and using the network output at the cut as "deep features" in combination with other classifiers [40].

Another interesting practical aspect is the idea of transfer learning. Large networks require large amounts of training data that may not always be available for some specific tasks. However, many low-level filters in coarse network layers are arguably useful for many different high-level tasks, and do not need to be re-learned from scratch each time. Thus, when training a network for a specific task (e.g. classifying between left and right hands), it is common practice to initialize the network with pre-trained (rather than random) weights. Those pre-trained weights are usually the result of longer training on a larger dataset, even if that dataset was collected for a different objective (e.g. ImageNet [25]).

1.5 SUMMARY AND THESIS OUTLINE

Thus far, we introduced the idea that first-person cameras can provide insight into how people visually perceive the world in dynamic, everyday contexts, and that a growing number of cognitive researchers use wearable camera systems as tools to study visual attention, perception, and object learning. Those systems often capture vast amounts of image or video data, such that sophisticated automated techniques are needed to help with data analysis. In this dissertation, we present different computer vision approaches that aim to

analyze hands in the context of first-person cameras. Hands are among the most frequent objects in our field of view, and we argue that the first-person perspective creates a very functional and embodied perspective of one's own hands. Investigating hands in our visual field is particularly interesting from a cognitive perspective, as hands help guide visual attention and modulate visual perception, both in humans' early cognitive development and as adults. In the context of computer vision research, most work on hand analysis is aimed at three-dimensional pose reconstruction with depth cameras, motivated by human computer interaction applications. We argue that considering wearable cameras is not just a minor deviation from existing this work, but rather a novel and original perspective on the problem that sprouts new and different research questions. For example, how can we distinguish our left hand from our right hand? In a social context, can we distinguish our own hands from the hands of partners that interact with us? Are visual features sufficient to distinguish hands or are there other helpful biases? Can hands be detected robustly in naturalistic, unconstrained video? Can we infer what the observer is doing based on the position and pose of hands in the field of view?

To answer these kinds of questions, we propose and investigate three different approaches that all aim to extract information about hands in the context of dynamic social interactions. We demonstrate that this information is meaningful by directly using hand annotations to answer cognitively motivated research questions (Chapter 2) and by showing that hand poses can be used as features to infer high level information about the egocentric observer (Chapter 5). The remainder of this thesis is structured as follows:

– In **Chapter 2** we consider laboratory video data of joint toy play between toddlers and parents, and design a method to track and distinguish hands in the toddler view based on simple spatial constraints imposed by the egocentric paradigm. Using this method, we collect fine-grained hand statistics that contribute new evidence on how infants and

their parents coordinate visual attention towards objects through eye-hand coordination.

– In **Chapter 3** we build upon the ideas of our initial approach to develop a more general, probabilistic graphical model framework that combines temporal and spatial biases of hand locations, as well as head motion of the first-person observer. We demonstrate that this approach can achieve notable results in distinguishing hand types even in situations where initial hand detections are extremely noisy, as often occurs in videos of unconstrained, natural environments.

– **Chapter 4** asks to what extent we can identify different hand types and hand poses directly based on their visual appearance. We collect a novel, large-scale dataset with pixel-level annotations of hand poses in first-person video, and use it to train data-driven, state-of-the-art visual recognition models such as convolutional neural networks (CNNs). We show that CNNs are able to robustly identify hand types based on visual information alone, while we can still use spatial biases of hand locations in first-person video to efficiently locate hands. Moreover, we demonstrate that we can utilize our robust detections to segment hands with state-of-the-art accuracy.

– In **Chapter 5** we begin with the extracted hand poses from the previous chapter, and explore the extent to which poses and locations of hands in the first-person view can inform recognizing the high-level interaction of the video (e.g. playing chess). We demonstrate that CNNs can recognize interactions with accuracy far above baseline based on hand poses extracted from a single frame, and can further improve by combining evidence across time or across the different viewpoints of partners.

– Finally, **Chapter 6** summarizes our contributions and provides an outlook into possible future work.

CHAPTER 2

ANALYZING HANDS IN INFANTS' EGOCENTRIC VIEWS

2.1 INTRODUCTION: VISUAL ATTENTION THROUGH HANDS

The visual world is inherently cluttered and dynamically changes over time. To efficiently process such a complex visual world, our perceptual and cognitive systems must selectively attend to a subset of this information. Humans have the greatest visual acuity around the center of their eye fixation (foveal vision), with acuity roughly declining inversely-linearly towards the edges of their field of view (peripheral vision) [105]. Thus, visual attention is often viewed as a spatial spotlight [83] around the center of a fixation. Although adults can attend to locations outside the area targeted by eye gaze [98], in many situations attention is tied to the body and sensory-motor behaviors, such that adults typically orient gaze direction to coincide with the focus of the attentional spotlight. For example, studies that investigate the coordination of eye, head, and hands of adults engaged in complex sensory-motor tasks (e.g. preparing sandwiches or copying *LEGO* block patterns) suggest that the momentary disposition of the body in space serves as a deictic (pointing) reference for binding sensory objects to internal computations [6, 43].

Visual attention and information selection are also critical factors during early cognitive development since an aptitude for early sustained attention [75] can be predictive of later developmental outcomes [94]. Many traditional studies of the development of attention employ highly-controlled experimental tasks in the laboratory, using remote eye tracking

systems to measure looking behaviors when toddlers passively examine visual stimuli displayed on a computer screen. While these paradigms can be very powerful, we also know that they are very different from young children's everyday learning experiences: active toddlers do not just passively perceive visual information but instead generate manual actions to objects, thereby creating self-selection of object views [124]. Compared with adults, young children's attentional systems may be even more tied to bodily actions. Thus, more recent studies started using head-mounted eye tracking systems to study visual attention in freely-moving toddlers when they are engaged in everyday tasks [39]. Though complex, these are arguably the contexts in which real-world learning for our visual system occurs.

The overarching goal of the study in this chapter is to understand how sensory-motor behavior supports effective visual attention in toddlers. Towards this goal, we developed an experimental paradigm in which a child and parent wear head-mounted eye trackers while freely engaging with a set of toys. Each eye tracking system captures egocentric video as well as the gaze direction within the captured first-person view. In this way, we precisely measure the visual attention of both the parent and child.

Recent findings using the same paradigm show that in toy play, both children and parents follow hands to visually attend not only the objects held by oneself but also the objects held by the social partner [122, 124]; in doing so, they create and maintain coordinated visual attention by looking at the same object at the same time. Similarly, other work has shown that by holding objects, parents increase the likelihood that infants will look at parents' hands [39]. These results suggest the important role of hands and hand activities (of both children and parents) in toddlers' visual attention.

Given these findings, the work presented in this chapter focuses on providing new evidence towards how eye and hand actions interact to support effective visual attention to objects in toddlers. We first describe a new method to automatically detect and distinguish

hands in our video data, allowing us to locate (at a pixel level) both the camera wearer's own hands and the social partner's hands in the first person view. Then, we report a series of results that link hands and hand actions with visual attention, to show how the child's and parent's hands contribute to visual information selection in the child's view.

2.2 RECORDING FREE-FLOWING CHILD-PARENT TOY PLAY

In order to study visual attention in toddlers, we developed a multi-model sensing system that allows free-flowing toy play between a child-parent dyad, while passively recording each participant's view and gaze, as well as other sensory modalities.

2.2.1 MULTI-MODAL SENSING SYSTEM

Our multi-modal sensing environment allows us to monitor parents and children as they engage in free-playing interaction with toy objects, as shown in Figure 2.1. A child and parent sit at a table in a white laboratory environment and face one another. Each wears a lightweight, head-mounted eye tracking system (*Positive Science LLC*) consisting of two cameras: a wide-angle outward-facing camera (100° diagonal angle of view) capturing the egocentric field of view of the participant, and an inward-facing infrared camera pointed at the participant's left eye, which tracks the pupil in order to measure eye gaze position (shown by green cross-hairs in Figure 2.1). While we know that the human visual field is much broader (around 190° for adults), previous studies (e.g. Franchak *et al.* [39]) have demonstrated that well-calibrated first-person video and eye movement data are still reliable approximations of people's (and toddlers') visual fields and overt attention. The interaction environment in our lab setting is arguably less cluttered than the real world since we cover the background with white curtains (see Figure 2.1). We do this to occlude task-irrelevant distractors so that participants focus on free-play, attending solely to the toys and or each

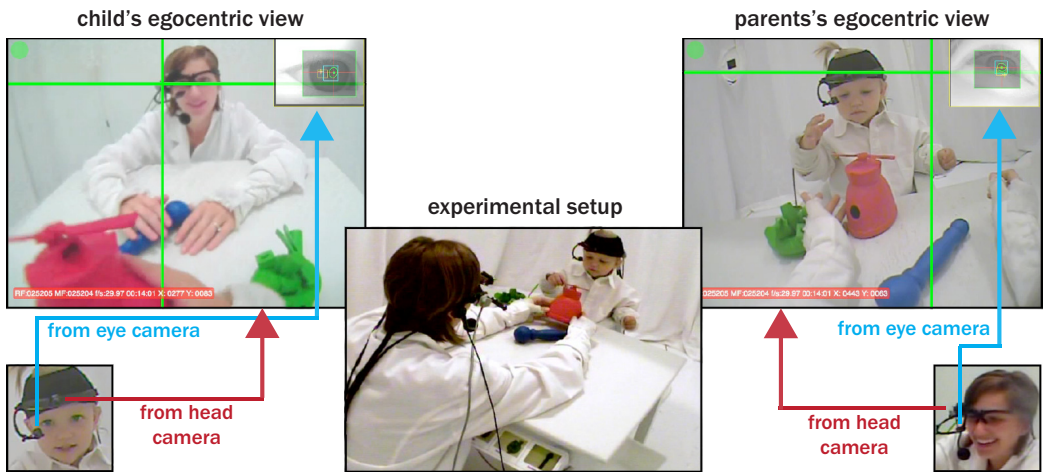


Figure 2.1: *Experimental setup of our multi-modal sensing system.* We use 4 cameras to record joint play between a child and parent. The head-mounted eye tracking systems (worn by both) each consist of a head camera to capture its wearer's egocentric view and an eye camera that tracks the eye's pupil. All cameras work with a temporal resolution of 30Hz and a spatial resolution of 480×720px.

other. Importantly, the setup allows free-flowing interaction with active exploring of toys, resulting in natural unfolding of visual attention.

In addition to the eye tracking camera systems, the setup also includes two scene cameras, two microphones, as well as head- and wrist-mounted motion sensors in order to support analysis on multiple modalities. As the purpose of this study is to investigate the role of hands in the toddler's view, the focus of analysis will be solely on the child's egocentric video and eye gaze data.

2.2.2 SUBJECTS, PROCEDURE, AND DATA COLLECTION

We considered six child-parent dyads for this study. The infants' mean age was 19 months ($SD = 2.56$ months). A team of two experimenters placed the eye tracking system on the infant and performed a calibration procedure (see [122] for details). Parents were told to engage their child with toys (three possible toys were on the table) and otherwise interact as naturally as possible, leading to a free-flowing interaction with no constraints on where

parents or children looked or what they should do or say. Each experiment consisted of four trials and each trial lasted about 1.5 minutes. In between trials, the toy sets were replaced to keep the children interested, and, if necessary, the eye tracking system was re-calibrated.

We collected a total around 68,000 frames (or 38 minutes) of video data from the six children. Of those frames, 54,367 contained valid gaze data (i.e. gaze located within the camera's field of view) in the form of an xy-coordinate, indicating the gaze center.

2.3 DETECTING AND LABELING HANDS

As this chapter is driven by the goal of detecting hands to provide empirical data for the exploration of visual attention, we employ a hand detection method that is tuned for high accuracy in the context of the multi-model sensing system and our experimental setup (see Figure 2.1). Thus, this section does not follow traditional computer vision (or general machine learning) paradigms, such as a clear distinction between training or testing data. Nonetheless, as we will see later, some key ideas of the method presented here will also be incorporated in the more general vision frameworks proposed in Chapters 3 and 4.

We take advantage of the constraints of our lab environment in the following ways: We know there are at most two people in each frame, that the child's hands are closer to the head-mounted camera than the adult's hands, that children and parents are facing one another, and that the participants' clothing is white. Our goal is to identify which of the four hands (child's hands and parent's hands) are visible in each frame, and then to identify the position of the visible ones. We achieve this goal in four major steps: (1) identifying potential skin pixels based on color; (2) clustering these pixels into candidate hand and face regions; (3) tracking these regions over time; and (4) labeling each region with its body type (child's left or right hand, parent's left or right hand, parent's face). Each of these steps are described in more detail in the following sections and summarized in Figure 2.2.

2.3.1 STEP 1: SKIN DETECTION

To look for faces and hands, we first identify pixels that have skin-like colors. Although human skin colors are surprisingly consistent across people when represented in an appropriate color space (we use YUV here), pixel-level skin classification is still a difficult problem because illumination can dramatically alter skin appearance and because many common objects and surfaces often have skin tones. We thus tuned our skin classifier for each individual subject pair, by sampling 20 frames at random and having a human label the skin regions in each frame. We then used these labeled pixels as training exemplars to learn a simple Gaussian classifier, in which each pixel is encoded as a 2D feature vector consisting of the two color dimensions (U and V). To detect skin in unlabeled frames, we evaluate the likelihood of each pixel under this model, threshold to find candidate skin pixels, and use an erosion filter to eliminate isolated pixels.

2.3.2 STEP 2: SKIN CLUSTERING

Given the detected skin pixels from Step 1, we apply mean shift clustering [20] to each frame to group skin pixels into candidate skin regions. Mean shift places a kernel (in our implementation we use a circular disk with a fixed radius) at a random location of the image, calculates the mean position of all skin pixels that it covers, and then shifts the kernel such that it is centered around the mean. This procedure is iteratively repeated until the kernel converges on a cluster center. Then, a new disk is added and the whole procedure is repeated until all skin pixels are covered. Mean shift does not require knowing the number of clusters ahead of time (as k-means does), which is beneficial as we do not know the total number of hands in each frame. The only free parameter is the size of the radius of the disk, for which we chose 75 pixels, which roughly corresponds to the expected size of a hand in our data.

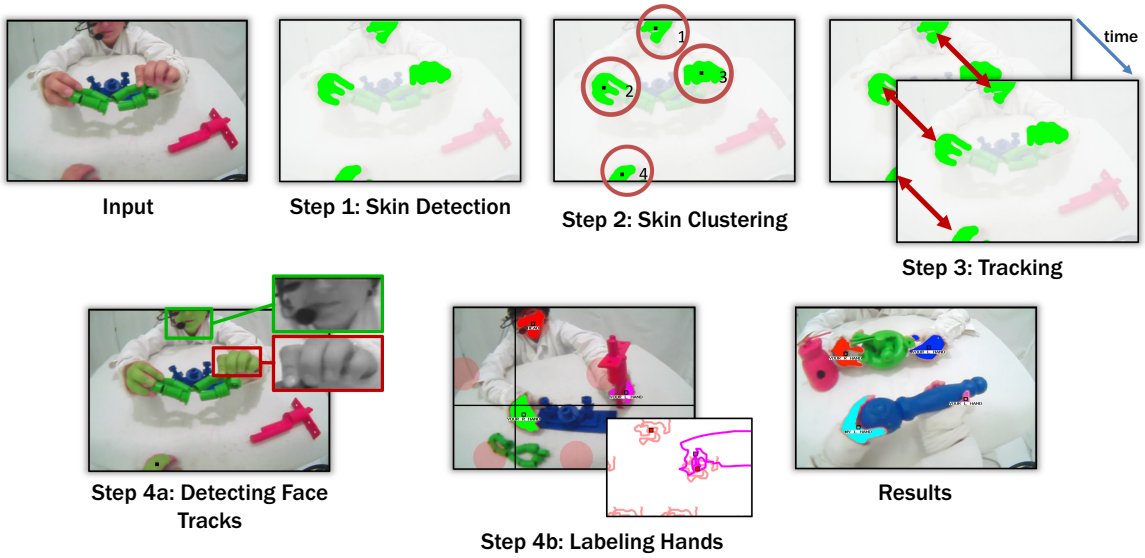


Figure 2.2: *Summary of our hand detection method for the egocentric toddler data.* The approach consists of four major steps: (1) identifying potential skin pixels based on color; (2) clustering these pixels into candidate hand and face regions; (3) tracking these regions over time; and (4) labeling each region with its body type (face, child's left or right hand, parent's left or right hand).

2.3.3 STEP 3: TRACKING

Next, we attempt to find correspondences between skin clusters across temporally-adjacent frames, in order to create *tracks* of skin regions over time. To do this, we scan the frames of a video in sequence. For each frame i, we assign each skin region to the same track as the closest region in frame $i-1$ as long as the Euclidean distance between the region centroids is below a threshold (we use 50 pixels), and otherwise we start a new track. Each track thus consists of a starting frame number indicating when the region appears, an ending frame number indicating when it disappears, and an (x, y) position of the region (together with all pixels belonging to the cluster) within each intervening frame.

2.3.4 STEP 4: LABELING SKIN REGIONS

Finally, we need to label each of the tracks from Step 3 with one of five possible body parts (i.e. child's left or right hand, parent's left or right hand, parent's face). We experimented

with various strategies and settled on a relatively simple approach that uses the relative spatial location of tracks within the video (and in particular the observation that the parent's head is usually above and between the parent's hands, which are in turn above the child's hands). We thus first try to find tracks corresponding to the parent's face, and then check the relative position of other tracks to find and label the hands.

Detecting Face Tracks

We tried off-the-shelf face detectors (such as [50]), but found them unreliable in our context because the parent's face is often not fully visible (e.g. in the input frame of Figure 2.2). Instead we built a very simple face detector that uses the fact that the parents in our experiments wear a black head-mounted camera. Thus, regions with the parent face should have a higher portion of dark pixels (due to the camera) in comparison to other hand regions. We trained a linear support vector machine classifier [21] on manually-labeled face regions (using the same 20 per-subject frames that we used to learn skin color, with the remaining skin regions serving as negative exemplars), where the features consist of a 256-bin grayscale histogram over the pixels in the skin region. We then identify faces by finding tracks for which the trained SVM classifies over half of the regions in the track as faces.

Labeling Hands

Once face tracks have been found, we mark potential hand tracks based on their relative position with respect to the face. Besides taking advantage of the special constraints of our setup, anchoring the expected spatial locations of hands to the parent's head also helps to compensate for view changes due to the child's head motion. In particular, we create a configuration of four points ("hotspots") that roughly correspond to the expected (mean) position of the four hands relative to the face, illustrated as red circles in Figure 2.2 (Step 4b). For each non-face candidate track generated by Step 3, we compute the centroid of its

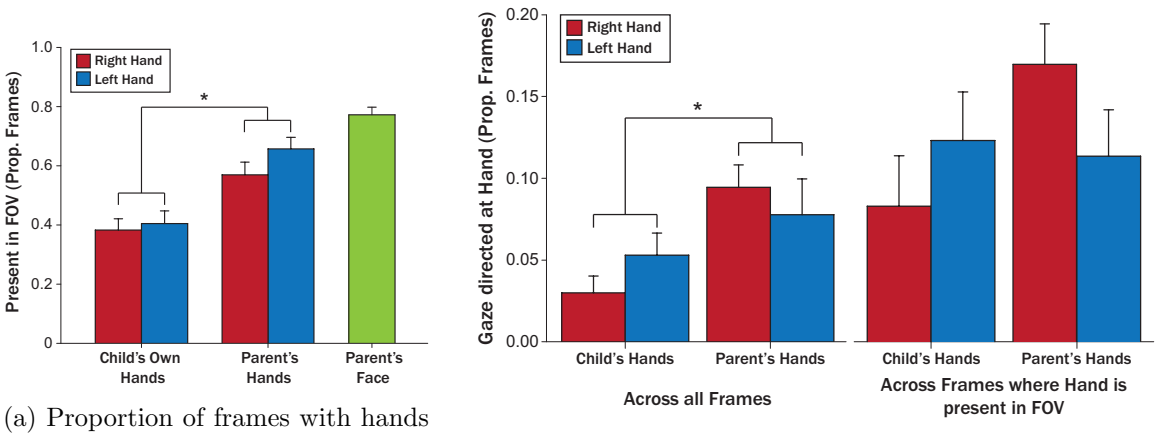
location across the frames in which it is visible, find the hotspot closest to the centroid, and assign the track to the corresponding hand. When no face is detected, the hotspots take a default position that assumes the face is in the center-top region of the frame.

2.3.5 EVALUATION

We manually tested the accuracy of our hand tracking algorithm on 600 randomly-selected frames (100 frames for each of 6 subjects), and counted the proportion of correctly-labeled regions. We found that the overall accuracy was 71%, ranging from 67% to 75% across the subjects. In comparison, a baseline method that randomly assigns labels to skin regions (and assuming that the skin segmentation and clustering perform correctly) achieves 20% accuracy. Labeling errors are caused by a variety of factors, but the two most common are: (1) When hands are close together and the clustering algorithm incorrectly combines them into a single body part, and (2) when hands spend a significant amount of time away from their expected location relative to the head.

2.4 RESULTS: HOW INFANTS PERCEIVE HANDS

The proposed hand detection scheme provides frame-by-frame data about the position, size, shape, and label of each hand in the child's field of view. This data allows a fine-grained analysis of hand appearances both in terms of frequency (How often are hands in the toddler's view?) and spatial distribution (Where in the view are hands?). In combination with the recorded eye gaze, the collected data additionally allows investigation of where and when each hand was the target of the child's overt visual attention.



(a) Proportion of frames with hands present in the field of view

(b) Proportion of frames with gaze directed at hands

Figure 2.3: *Frequencies of hands in view and of hands targeted by gaze.* (a) Bar graphs showing the proportions of frames in which each class of hands was detected (error bars show 1 SE). For comparison, the value for the parent's face is also shown. (b) Bar graphs showing the proportions of frames in which each class of hands was looked at (based on a 10° gaze hot spot). *Left:* Fractions based on all frames with valid eye gaze ($N = 54,367$). *Right:* Fractions based on all frames where the corresponding hand was in the field of view.

2.4.1 HANDS IN THE INFANT'S FIELD OF VIEW

To determine how often children had the opportunity to view their own hands and their parents' hands, we first calculated how often hands are present in the field of view.

Frequency of Hands in View

Figure 2.3a shows the proportion of frames in which each hand class was detected. As the hand tracking algorithm needs to distinguish hands from faces and thus implicitly tracks the parent's face as well, we include results for the face for reference. Overall, hands were frequently in view, although the child's own hands (right hand = 38% and left hand = 40%) are in view less frequently than the parent's hands (right hand = 57% and left hand = 66%). A 2 (agent: child, parent) × 2 (hand: left, right) repeated-measures ANOVA confirmed a main effect of agent, $F(1, 5) = 21.74$, $p = .006$. The main effect of agent × hand interaction did not reach significance.

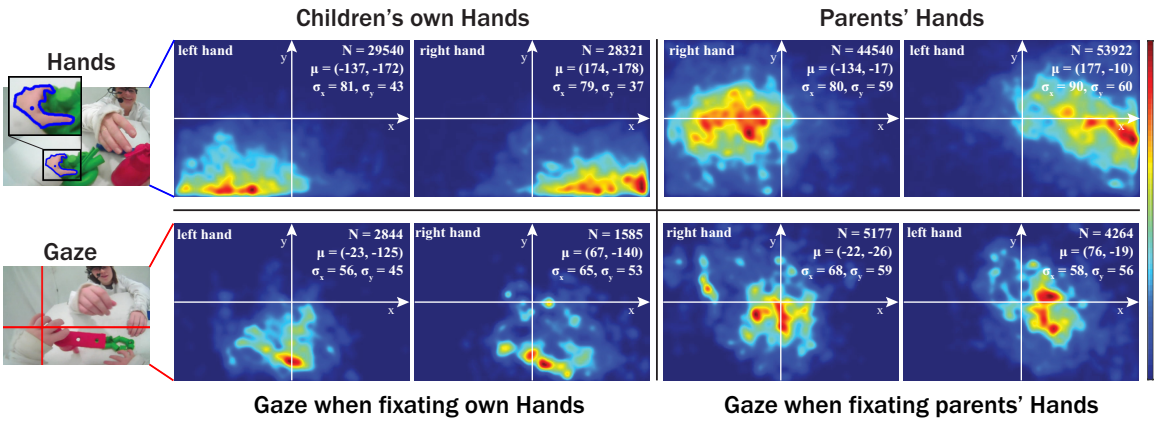


Figure 2.4: *Spatial distributions of hands and eye gaze.* The top row (left) shows the spatial distributions of the children's own hands (based on hand centroids) within their field of view. Similarly, the right side of the top row shows the distributions of the parents' hands in the children's field of view. The bottom row shows the spatial distributions of the children's eye gaze while looking at their own hands (left) or their parents' hands (right). Also shown are robust (60% trimmed) estimates of mean (μ) and standard deviation (σ) of the distributions as well as the number of data points (N). Heat maps are 480×720px and a small Gaussian blur ($\sigma_G = 10$px) was applied for better visualization.

Spatial Distribution of Hands in View

Spatial asymmetries might account for the different frequencies with which children's and parents' hands were visible. Next, we present spatial distributions of hands in the children's field of view in the form of heat maps. The top row of Figure 2.4 shows the distributions of the child's left hand, the child's right hand, the parent's right hand and the parent's left hand, respectively. Each data point in the heat map corresponds to the centroid (mean of the hand area) of the detected hand. The distributions are accumulated across all six subjects where N depicts the total number of frames with the hand in view. To allow quantitative comparison, we calculated robust (60% trimmed) statistics in the form of horizontal and vertical mean (μ) as well as horizontal and vertical standard deviation (σ) of the distributions (off-diagonal co-variances are not shown).

Children's left and right hands had very similar distributions in terms of variance with distributions that expanded more horizontally than vertically: σ_x was roughly twice as much

as σ_y for each hand. Parents' left and right hands also have similar distributions in terms of variance. A 2 (agent: child, parent) $\times$ 2 (hand: left, right) $\times$ 2 (direction: horizontal, vertical) ANOVA confirmed the main effect of direction, $F(1,5) = 36.4$, $p = .002$. However, a significant agent $\times$ direction interaction, $F(1,5) = 10.5$, $p = .023$ and follow-up pairwise comparisons show that parents' hands occupy a larger vertical space (right hand $\sigma_y = 59$, left hand $\sigma_y = 60$) compared to children's hands (right hand $\sigma_y = 37$, left hand $\sigma_y = 43$, $p = .009$). Horizontal variance terms did not differ between the hands of children and parents, and no other effects approached significance.

Children's and parents' hands were spatially segregated in visual space. Overall, children's hands were lower in the visual field compared to parents' hands and were often seen towards the lower boundary of the field of view ($\mu_y = -172$ for the left hand and $\mu_y = -178$ for the right hand). A 2 (agent: child, parent) $\times$ 2 (hand: left, right) ANOVA on μ_y revealed that parents' hands were significantly higher than children's hands (main effect of agent, $F(1,5) = 184.6$, $p < .001$). In the horizontal dimension, the child's right hand and parents' left hand tended to reside in the right half of the visual field, while the child's left hand and parents' right hand tended to reside in the left half of the visual field. A 2 (agent: child, parent) $\times$ 2 (hand: left, right) ANOVA on μ_x confirmed a significant agent $\times$ hand interaction, $F(1,5) = 1377.7$, $p < .001$.

Since our automatic hand labeling is not perfect and makes spatial assumptions, these results could potentially be biased by our algorithm. Thus, we manually labeled the location of hands in a random subset of the data (2,800 frames) and repeated our analyses. A 2 (agent: child, parent) $\times$ 2 (hand: left, right) ANOVA on the μ_y's of manually labeled frames confirmed that parents' hands were located higher than those of children (main effect of agent, $F(1,5) = 111.1$, $p < .001$). In addition, a 2 (agent: child, parent) $\times$ 2 (hand: left, right) ANOVA on the μ_x's in hand labeled frames showed a significant agent $\times$ hand

interaction as in frames labeled by our algorithm, $F(1, 5) = 529.4$, $p < .001$). We conclude that our results on spatial locations of hands in the field of view are likely not an artifact of the algorithm.

Different spatial distributions of hands may account for different frequencies of hands being visible. Most likely, children's hands were not as frequent as parents' hands because they occupied locations towards the lower boundary of the field of view. If children moved their hands down or tilted their heads up, their own hands would leave the field of view.

2.4.2 HANDS AS TARGETS OF THE INFANT'S OVERT ATTENTION

Next we examined how often and where hands were targeted by children's gaze. We counted a gaze fixation on the hand whenever a 10° hot spot (corresponding to a circle with radius of 32 pixels) around the gaze center overlapped with the area of a detected hand.

Frequency of Hands Being Targeted by Gaze

Figure 2.3b (left) shows mean values for the overall proportion of frames in which children's gaze overlapped with each hand. Children spent about twice as long looking at parent's hands (about 9.5% for the right hand and 7.8% for the left hand) than they did looking at their own hands (3.0% right hand and 5.3% left hand). A 2 (agent: child, parent) × 2 (hand: left, right) on proportion of frames targeting hands confirmed a main effect of agent, $F(1, 5) = 8.52$, $p = .03$, and found no other significant effects.

Higher rates of looking to parents' hands may be the result of parents' hands being in view more often. Thus, we recalculated the proportion of looking to hands based on the number of frames where each hand was present in the field of view (right side of Figure 2.3b). This normalization increased the proportion of looking for both the child's own hands and the parent's hands. Furthermore, the difference between the time spent looking at parent's

hands and looking at their own hands is no longer significant when taking the availability of hands into account (no effects found in a 2 (agent: child, parent) $\times$ 2 (hand: left, right) on normalized proportions of frames targeting hands).

Spatial Distribution of Gaze when Targeting Hands

Prior work has shown that gaze allocation in natural environments tends to be biased towards the center of the field of view [38]. The overall gaze distribution across all six toddlers in our experiment confirms this bias with a mean near the center ($\mu_{xy} = (10, -22)$) and similar variances in horizontal and vertical direction ($\sigma_{xy} = 79, 82$)). In the bottom row of Figure 2.4, we present the spatial distributions of children's eye gaze when viewing hands. The gaze heat maps are composed similarly to the hand heat maps (top row of Figure 2.4), except that each data point now corresponds to the eye gaze center as opposed to a hand centroid. Across children's and parents' hands, we observed that distributions of gaze targeting hands were more centrally located compared to the overall distributions of hands in the field of view (FOV). To verify this statistically, we calculated the distances from the FOV center to the per-subject means of the distributions of the hands, and to the means of all gaze locations when hands were fixated. A 2 (agent: child, parent) $\times$ 2 (hand: left, right) $\times$ 2 (distribution: hands overall, gaze-targeted) revealed a main effect of agent, $F(1, 5) = 66.1$, $p < .001$, distribution, $F(1, 5) = 13.7$, $p = .014$, and a significant 3-way interaction, $F(1, 5) = 17.7$, $p = .008$. Overall, parents' hands ($M = 131.4$ pixels) were closer to the center of the field of view compared to children's hands ($M = 195.5$ pixels). Follow-up tests on the 3-way interaction showed that child's left hand, child's right hand, and parent's left hand were more centrally located when targeted by gaze compared to their overall distributions ($p < .05$), while the parent's right hand location did not change when targeted by gaze ($p = .48$).

2.4.3 DISCUSSION

Hands are an important visual stimulus. One's own hands are relevant for guiding reaching actions and manipulating objects [39, 43], while the hands of others can convey information about the attention and goals of social partners [79, 113]. But for toddlers to learn from hands, they must be able to see them. Here, we demonstrate that for toddlers playing with adults, hands are frequently in view. However, what infants see depends on where they actively point their heads: the resulting spatial constraints (e.g. child's hands being low in the field of view) mean that children's own hands are in view less often than their parents' hands. Consequently, children overtly attend to parents' hands more often than their own hands. Moreover, we show that when children fixate on hands, they do so more often when hands are centrally located in their fields of view, suggesting that children move their heads to bring visual targets into the center of their visual fields. Most likely, children coordinate their eyes and heads to focus on areas relevant to the task at hand, looking down towards their own hands when reaching and looking up towards their parent's hands when parents present objects [122].

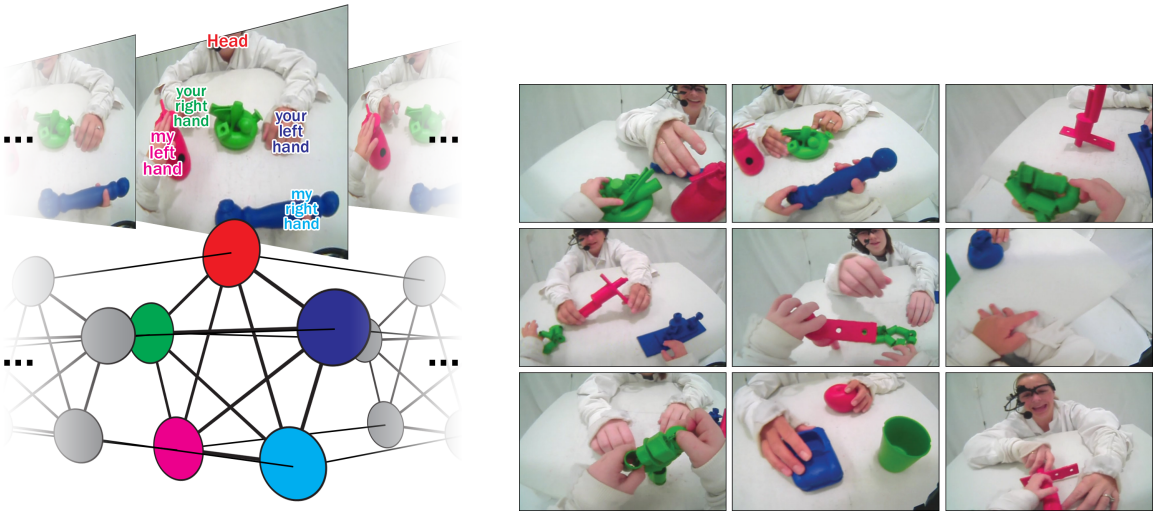
CHAPTER 3

A PROBABILISTIC FRAMEWORK TO LOCATE AND DISTINGUISH HANDS

3.1 INTRODUCTION: SPATIAL BIASES OF HANDS

The work discussed in the previous chapter demonstrated that it is possible to locate different hand types in dynamic egocentric interactions, such as joint child-parent toy play, with sufficient quality to generate useful hand annotations. While the presented approach took advantage of some of the clean characteristics of the laboratory video data (e.g. exploiting the white background by using a simple color-based skin detector), it also proposed an idea that is arguably true for characterizing first-person interactions more generally: spatial biases of hands in the first-person view.

In this chapter, we build upon such biases to overcome some of the challenges inherent to the dynamics of the first-person view. Continuing with the egocentric toddler data as a motivating example, Figure 3.1b shows a set of random example frames to illustrate these challenges. Parent and child play together and constantly use their hands to point to, reach for, and exchange toys. The toddler's view can potentially be very close to the action such that perceived hand sizes can vary from small to very large. In addition, the child's view is very dynamic over time: the hands of both the child and the parent frequently disappear, reappear, and overlap with each other. However, the egocentric context imposes important spatial constraints on where to expect each type of hand. These constraints are quite



(a) Hands as part of a graphical model (b) Sample frames

Figure 3.1: *Model overview and sample frames.* (a) We propose a probabilistic model that incorporates appearance, spatial, and temporal cues to locate and distinguish hands. (b) Some sample images from the egocentric videos that were used in our experiments. The child's view is very dynamic; hands come in and out of view and overlap very frequently.

intuitive; for example, given that we see the scene through the eyes of the child, we expect the child's left hand to enter the child's view predominantly from the lower left and the right hand to enter from the lower right. In addition to these absolute spatial assumptions, one can also think of relative spatial biases. We generally expect the child's left hand to be to the left of the right hand, and vice-versa. As most social interactions have both participants facing each other, similar assumptions can be made for the social partner (in this case the parent). The parent's right hand usually occupies the left side of the child's field of view and the parent's left hand is on the right.

We propose a probabilistic graphical model framework (sketched in Figure 3.1a) that jointly models various spatial and temporal biases of hands, as well as the head motion of the observer, in order to overcome the challenges of the dynamic first-person perspective. While again motivated by the toddler data from Figure 3.1b, the proposed framework generally models the dynamics between hands in egocentric interactions. Thus, we evaluate our approach not only on a collection of 20 parent-child videos, but also on a small set of

videos of interacting adults in a naturalistic environment.

The following sections formally introduce our proposed egocentric hand modeling framework, show how to apply it with the child-parent data as an example, and evaluate its performance compared to a set of baselines. As we will demonstrate, our approach helps to significantly decrease errors resulting from confusing different types of hands while also improving the overall detection rate.

3.2 MODELING EGOCENTRIC INTERACTIONS

Given a first-person video containing an interaction between the observer and a social partner, our goal is to estimate which of the (up to four) hands are visible to the observer at any given time, and where in the view those hands are. As we expect to often also have our partner's face prominently in view, and as the location of our partner's hands are naturally tied to the rest of his or her body, we also try to estimate the location of the partner's face. We will jointly refer to these five objects (four hands and the face) as body parts or simply parts.

3.2.1 HANDS AS LATENT RANDOM VARIABLES

More formally, given a video sequence of n frames, each with $r \times c$ pixels, our goal is to estimate the position of each of a set of parts $\mathcal{P}$ in each frame. We specifically consider five parts, $\mathcal{P} = \{\mathsf{ml}, \mathsf{mr}, \mathsf{yl}, \mathsf{yr}, \mathsf{yf}\}$, referring to the observer's hands ('my left' and 'my right'), the partner's hands ('your left' and 'your right'), and the partner's face ('your face'), respectively. We denote the latent 2D position of part $p \in \mathcal{P}$ in frame i as L_p^i and define L^i to be the full configuration of parts within the frame, $L^i = \{L_p^i\}_{p \in \mathcal{P}}$. At any moment in time any given part can be anywhere in view or not in view at all. To address the possible absence of parts, we augment the domain of L_p^i with an additional state $\emptyset$

indicating that the part is not visible in the frame, i.e. $L_p^i \in \{\emptyset\} \cup ([1, r] \times [1, c])$.

In addition to the position of each body part, we also want to estimate the global motion of the observer's view. If the observer decides to turn his or her head, all of the visible parts should change their position accordingly. Thus, intuitively, an estimate for the direction and amplitude of the observer's head's motion should help to estimate more likely positions of body parts in view. We model global head motion by introducing a set of random variables $G = (G^1, \ldots, G^{n-1})$, where G^i is an estimate of the two-dimensional global coordinate shift between frame i and frame $i + 1$. The domain of G^i is bounded by some maximum shift in horizontal and vertical direction that we allow to observe, i.e. $G^i \in ([0, x_{max}] \times [0, y_{max}])$. In this way, we assume the world has uniform depth such that a change of viewing angle would have the same effect on all points in the 2D projection of the environment. This assumption is reasonable given that the distances involved in a paired interaction are relatively small.

3.2.2 BUILDING A GRAPHICAL MODEL

We now consider all of these variables as nodes in one big graphical model framework, which will allow us to estimate the locations of all parts jointly and probabilistically across the whole video. For each observed frame i of the video, we can apply any object detection model to get a (noisy) estimate for likely positions of each part L_p^i. In addition to those individual observations, the graphical model also incorporates three types of constraints: (1) absolute constraints on where each body part should be, (2) intra-frame constraints on spatial relationships between body parts, and (3) inter-frame constraints between body parts, which enforce temporal smoothness on part positions. These constraints model intuitions such as: (1) my own left hand should likely be on the left side of my field of view, (2) my own right hand should likely be to the right of my left hand, and (3) a hand in frame i

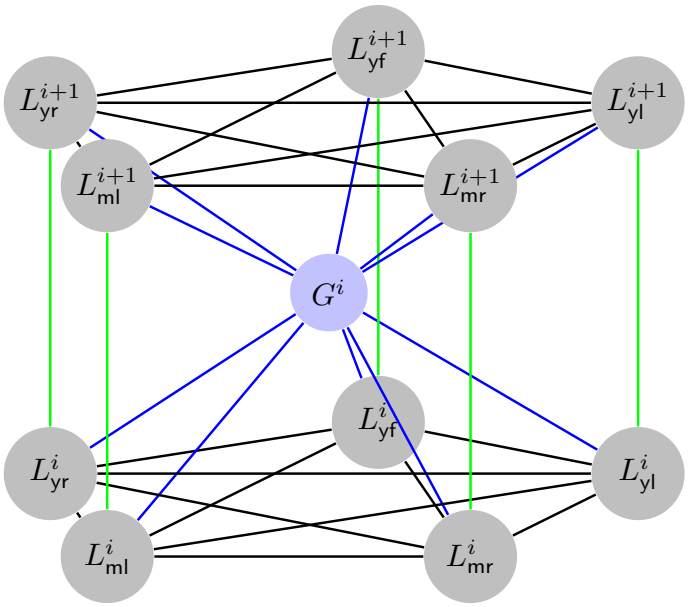


Figure 3.2: *Graphical depiction of our PGM for a 2-frame video,* where the bottom five nodes represent the locations of face and hands in one frame, and the top five nodes represent the locations in the next frame. Between-frame links (green) enforce temporal smoothness, shift links (blue) model global shifts in the field of view, and in-frame links (black) constrain the spatial configuration of the body parts.

should be close to where it was in frame $i-1$.

A visualization of the resulting graph for a 2-frame video is presented in Figure 3.2. The connections within a frame (in black) form a complete graph over the five part nodes and capture the pairwise correlations between spatial locations of the parts. The green edges between each part and its corresponding variable in the next frame enforce the temporal smoothness constraint. Finally, the global shift variable is influenced by all pairs of corresponding parts such that a similar motion in all part pairs is likely to indicate a global shift, and conversely, an observed global shift is likely to influence all parts.

Following these constraints yields a joint distribution over all the latent variables $L = (L^1, \ldots, L^n)$ and $G = (G^1, \ldots, G^{n-1})$, conditioned on all video frames $I = (I^1, ..., I^n)$,

$$P(L, G|I) \propto \prod_{i=1}^{n} \left[P(I^i, I^{i+1}|G^i) \prod_{(p,q)\in\mathcal{E}} P(L^i_p|L^i_q) \prod_{p\in\mathcal{P}} P(I^i|L^i_p) P(L^{i+1}_p|L^i_p, G^i) P(L^i_p) \right] \tag{3.1}$$

where $\mathcal{E} \subset \mathcal{P}^2$ is the set of undirected edges in the complete graph over $\mathcal{P}$.

We can solve the part-tracking problem for an entire video I by maximizing Equation (3.1). Unfortunately, as discussed in Section 1.4.1, finding the global maximum is intractable. Instead, we settle for approximate inference using Gibbs sampling [15]. As we will discuss in Section 3.2.7, this avoids the need to compute or store the full joint distribution because the sampling involves only small neighborhoods of the graph.

3.2.3 SPATIAL DISTRIBUTIONS AS ISOTROPIC GAUSSIANS

The next step in defining the PGM is to decide how to model the spatial relationships of and among parts. As is common in part-based object detection models [23,36], we model all spatial distributions as 2D Gaussians. For simplicity and computational efficiency, we also assume that these Gaussians are isotropic, which is also common [23]. Thus, the probability distribution of a part being at any location (x, y) in the frame is given as

$$f_{\mu,\Sigma}(x, y) = \mathcal{N}(x; \mu_1, \Sigma_{11}) * \mathcal{N}(y; \mu_2, \Sigma_{22}), \tag{3.2}$$

parameterized by $\mu = [\mu_1 \ \mu_2]^T$ and $\Sigma = \text{diag}(\Sigma_{11}, \Sigma_{22})$. With that assumption, we can calculate the probability of a part falling into a pixel bin (x_p, y_p) as

$$\begin{aligned} F_{\mu,\Sigma}(x_p, y_p) &= \int_{x_p-\frac{1}{2}}^{x_p+\frac{1}{2}} \int_{y_p-\frac{1}{2}}^{y_p+\frac{1}{2}} f_{\mu,\Sigma}(x, y)\, dy\, dx \\ &= \left[\Phi(x_p + \frac{1}{2}; \mu_1, \Sigma_{11}) - \Phi(x_p - \frac{1}{2}; \mu_1, \Sigma_{11})\right] \\ &\quad * \left[\Phi(y_p + \frac{1}{2}; \mu_2, \Sigma_{22}) - \Phi(y_p - \frac{1}{2}; \mu_2, \Sigma_{22})\right], \end{aligned} \tag{3.3}$$

where $\Phi(\cdot)$ is the normal cumulative density function and can be precomputed for efficient computation.

One extra complication for our problem is that we need to explicitly model the possibility of a body part being outside of the field of view (the $\emptyset$ state introduced in Section 3.2.1). We assume that calculating the probability that a part is 'out' of a frame is equal to one minus the probability of being anywhere within the frame,

$$F^{\emptyset}_{\mu,\Sigma} = 1 - \int_{1}^{c}\int_{1}^{r} f_{\mu,\Sigma}(x,y)\,dy\,dx, \tag{3.4}$$

where the integral can again be computed efficiently using $\Phi(\cdot)$ similarly to Equation 3.3.

3.2.4 ABSOLUTE SPATIAL PRIORS

Given the above parameterization, we can now define the spatial constraints of the model, starting with (1) absolute spatial priors on part locations as

$$P(L^i_p) = \begin{cases} F^{\emptyset}_{\mu_{pp},\Sigma_{pp}} : L^i_p = \emptyset \\ F_{\mu_{pp},\Sigma_{pp}}(L^i_{p,x}, L^i_{p,y}) : L^i_p \neq \emptyset, \end{cases} \tag{3.5}$$

where the mean absolute position μ_{pp} and the diagonal covariance matrix Σ_{pp} are learned for each part p based on a set of ground truth training frames in which part locations were manually annotated.

3.2.5 PAIRWISE SPATIAL PRIORS

Next, we define the pairwise spatial priors that model (2) intra-frame constraints on spatial relationships between body parts, and (3) inter-frame constraints to enforce temporal smoothness.

In-Frame Conditionals

Consider a pair of parts $p, q \in \mathcal{P}$ $(p \neq q)$ in frame i having positions L_p^i and L_q^i, respectively. Based on training data, suppose we have an estimate of the relative spatial relationship between these parts such that $L_p^i - L_q^i \sim \mathcal{N}(\mu_{qp}, \Sigma_{qp})$ for diagonal Σ_{qp}. We define the conditional probability distribution between L_p^i and L_q^i as

$$P(L_p^i | L_q^i) = \begin{cases} \beta : L_q^i = \emptyset \\ F^{\emptyset}_{\mu_{qp}+L_q^i, \Sigma_{qp}} : L_p^i = \emptyset, L_q^i \neq \emptyset \\ F_{\mu_{qp}+L_q^i, \Sigma_{qp}}(L_{p,x}^i, L_{p,y}^i) : L_p^i, L_q^i \neq \emptyset, \end{cases} \tag{3.6}$$

where β is a constant. Intuitively, this means that if part q is outside the frame, then it does not constrain part p's location (the conditional probability distribution is uniform), whereas if q is inside the frame, then p is either outside (and the conditional probability is given by one minus the probability of being inside the frame), or it is inside the frame with a probability given by the Gaussian distribution.

Between-Frame Conditionals

The inter-frame conditionals impose temporal smoothness on part locations, connecting together part p's location L_p^{i+1} in frame $i+1$, its location L_p^i in frame i, and the latent global shift G^i between frames i and $i+1$ (caused by head motion). We assume that if the part is within view in both frames i and $i+1$, then L_p^i and L_p^{i+1} are related by a Gaussian distribution with diagonal Σ_p around the location predicted by the global shift, $L_p^{i+1} - L_p^i \sim \mathcal{N}(G^i, \Sigma_p)$. Including the possibility of parts entering or leaving the frame, the

full conditional probability is similar to the above in-frame distribution,

$$P(L_p^{i+1}, G^i | L_p^i) = \begin{cases} \alpha : L_p^i, L_p^{i+1} = \emptyset \\ \frac{1-\alpha}{rc} : L_p^i = \emptyset, L_p^{i+1} \neq \emptyset \\ F^{\emptyset}_{\mu^i, \Sigma_p} : L_p^i \neq \emptyset, L_p^{i+1} = \emptyset \\ F_{\mu^i, \Sigma_p}(L_{p,x}^{i+1}, L_{p,y}^{i+1}) : L_p^i, L_p^{i+1} \neq \emptyset, \end{cases} \tag{3.7}$$

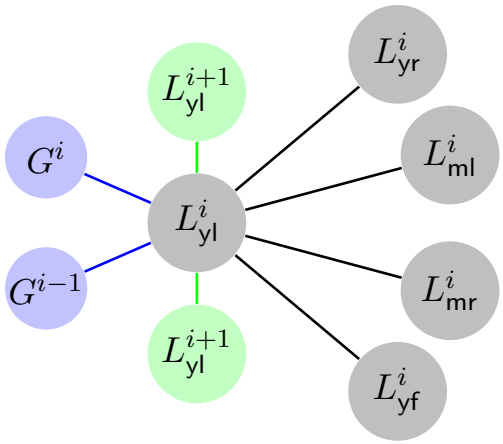
where $\mu^i = L_p^i + G^i$ and α is a constant. This conditional encodes the intuition that if a part is outside the image in one frame, it is outside the next frame with probability α or is uniformly distributed at a pixel in the frame. On the other hand, if a part is in the image in one frame, its probability distribution over pixels in the next frame is Gaussian, or it is outside the frame with probability one minus the integral over all pixel locations. This formulation encourages parts to stay at roughly the same position from one frame to the next, but allows for large jumps due to global motion if the jump is observed for many of the parts.

3.2.6 FULL CONDITIONALS

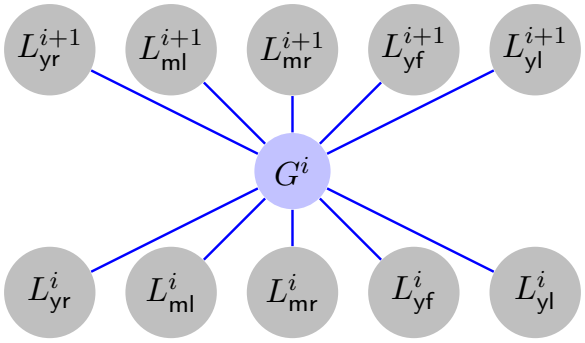
We use Gibbs sampling to perform inference, as we will describe in the next section. To do this, we need to sample each random variable from its full conditional distribution. Fortunately, because of the independence assumptions of our model, the full conditionals can be written and computed easily.

Part Nodes

We begin by deriving the conditional distribution of a part node given the rest of the variables in the graph. From Equation 3.1, we can compute the full conditional up to a



(a) Components for part node L^i_{yl}



(b) Components for shift node G^i

Figure 3.3: *Markov blankets for the full conditionals in our model.* Due to the independence assumptions given by the model, all latent variables only depend on a small set of other variables. (a) The Markov blanket for the part node L^i_{yl}, with the blankets for all other part nodes in L^i looking equivalently. (b) The Markov blanket for each shift node G^i.

proportionality constant,

$$\begin{aligned} P(L^i_p|G,L,I) \propto\ & P(L^{i+1}_p, G^i|L^i_{\text{yl}})P(L^{i-1}_p, G^{i-1}|L^i_p) \\ & * P(I^i|L^i_p)P(L^i_p) \prod_{q\in\mathcal{P}-\{p\}} P(L^i_q|L^i_p), \end{aligned} \tag{3.8}$$

where $P(I^i|L^i_p)$ is produced by an object detection model for p, which we define in Section 3.3. Taking L^i_{yl} as an example, the Markov blanket (defining all variables that L^i_{yl} is directly dependent upon) is shown in Figure 3.3a, such that the conditional is given as

$$\begin{aligned} P(L^i_{\text{yl}}|G,L,I) \propto\ & P(L^{i+1}_{\text{yl}}, G^i|L^i_{\text{yl}})P(L^{i-1}_{\text{yl}}, G^{i-1}|L^i_{\text{yl}}) \\ & * P(I^i|L^i_{\text{yl}})P(L^i_{\text{yl}})P(L^i_{\text{yf}}|L^i_{\text{yl}})P(L^i_{\text{yr}}|L^i_{\text{yl}}) \\ & * P(L^i_{\text{ml}}|L^i_{\text{yl}})P(L^i_{\text{mr}}|L^i_{\text{yl}}). \end{aligned} \tag{3.9}$$

Since the state space is discrete, the normalization constant is not needed for sampling, as it can be computed at runtime.

Shift Nodes

As illustrated in Figure 3.3b, the full conditional of a shift node G^i can also be written as a product of its neighbors in the graph,

$$\begin{aligned} P(G^i|G,L,I) &\propto P(I^i, I^{i+1}|G^i) \prod_{p\in\mathcal{P}} P(L_p^{i+1}, G^i|L_p^i) \\ &= P(I^i, I^{i+1}|G^i) P(L_{\mathsf{yf}}^{i+1}, G^i|L_{\mathsf{yf}}^i) P(L_{\mathsf{yl}}^{i+1}, G^i|L_{\mathsf{yl}}^i) \\ &* P(L_{\mathsf{yr}}^{i+1}, G^i|L_{\mathsf{yr}}^i) P(L_{\mathsf{mr}}^{i+1}, G^i|L_{\mathsf{mr}}^i) P(L_{\mathsf{ml}}^{i+1}, G^i|L_{\mathsf{ml}}^i). \end{aligned} \tag{3.10}$$

This product has several intuitive properties. If there is disagreement between the relative movements of parts, then the overall distribution is diffuse and the likelihood $P(I^i, I^{i+1}|G^i)$ term dominates, meaning the global shift is best estimated based on observing the frames. Here, we use a normal distribution fit to the dense optical flow [106] between frames i and $i+1$ as our estimate. If parts are in agreement, there is a high peak and the global shift is best estimated based on the joint movements of all parts.

3.2.7 INFERENCE

We use Gibbs sampling [15] to perform inference on our model. Gibbs sampling is a Markov-Chain Monte-Carlo method that generates samples from the full joint distribution of our model based on iterative sampling of all conditionals. Thus, this method does not need a (parametric or otherwise) representation of the very large joint distribution of our PGM. In the limit, these samples form an accurate representation of the true joint distribution. We obtain a solution from these samples as follows. If for any given frame, the majority of samples for a given part are in the $\emptyset$ state, we label the part as "out" of the frame. Otherwise, we take the median position over the in-frame samples. In our experiments, just 50 samples of the joint distribution provided good solutions.

3.3 SPECIALIZING TO CHILD-PARENT TOY PLAY

Our hand tracking approach could in principle be applied to any egocentric video data, with the various parameters and distributions set to customize it to a specific application. As mentioned in Section 3.2.6, one can apply any object model to generate the distributions for the per-part image likelihood terms $P(I^i|L_p^i)$ for each part location p in frame i. Since the context of our child-parent toy play data is rather controlled, we use simple (but fast to extract) image features to demonstrate the effectiveness of our methodology. In particular, we first detect skin pixel regions and use a distribution of $P(I^i|L_p^i)$ that is near zero unless L_p^i is on a skin pixel, and otherwise is proportional to the likelihood that an image patch around L_p^i 'looks like' part p, as described below.

3.3.1 SKIN MODEL

As our data only contains indoor footage with controlled lighting, we found that a color-based approach was sufficient for pixel-level skin detection. We learn non-parametric skin and background models in YUV color space (discarding the luminance plane Y). To detect skin in unlabeled images, we compute the log odds of each pixel under these models as $\log\left(\frac{P(U,V|skin)}{P(U,V|background)}\right)$, and threshold the output value to create a binary skin mask. We then apply a median filter to suppress noise. We explicitly modeled the background (in contrast to the skin detection approach in Section 2.3.1) as we found it to produce slightly better results.

3.3.2 FACE MODEL

We apply the Viola & Jones [50] face detector to each frame i to compute the face likelihood distribution, $P(I^i|L_{\mathsf{yf}}^i)$. We used a simple formulation in which pixels inside a detected face box are assigned high likelihoods and pixels outside are assigned a low (non-zero) likelihood.

We trained the detector on a small set of hand-labeled faces from our data.

3.3.3 ARM MODEL

Distinguishing hands based on the extracted color features alone is difficult. We thus implemented a simple model that takes advantage of the clean lab setting to extract arm regions, which provides a noisy estimate of the ownership of adjacent skin regions. The participants' sleeves tend to have a higher edge density than surrounding areas. To find arms, we apply an edge detector to each image, blur the output, apply a threshold to detect arm regions, and find skin patches that are adjacent to these regions. Suppose a skin patch and arm region intersect at a point u. We calculate the longest possible straight line through u intersecting the set of candidate arm pixels (i.e. the diameter of the arm pixel region through u). The direction and length of this line are a measure of the arm direction and length, so we use them to set the 'your hand' likelihoods, $P(I^i|L^i_{\mathsf{yl}})$ and $P(I^i|L^i_{\mathsf{yr}})$, based on thresholding the line length and direction. For instance, a skin patch with a long, upwards line is likely to belong to the partner's hand.

3.4 EXPERIMENTS

We primarily evaluate our method on the child-parent toy play data introduced in Chapter 2. We use video data from five different child-parent dyads, where each of the five play sessions consists of four trials that have an average length of 1.5 minutes, leading to a total of 20 videos containing 56,535 frames (about 31 minutes) of social interaction from the children's perspective. Some example frames of this data are shown in Figure 3.1b.

We also collected a small second dataset that was designed to test our model in more naturalistic settings. We used Google Glass to record three egocentric videos containing two adults engaged in three kinds of social interactions: playing cards, playing tic-tac-toe,

and solving a 3D puzzle. Each video is 90 seconds long, for a total of 4.5 minutes (8,100 frames), and was captured at 30Hz with a resolution of $1280 \times 720px$. Figure 3.5 shows some example frames of this data.

3.4.1 EVALUATION

To evaluate our approach, we manually annotated 2,400 random frames (around 120 per trial) from the lab dataset, and 300 frames (100 per video) from our Google Glass dataset, with bounding boxes. This is about one frame for every second of video. Depending on which body parts are in view, each frame has up to five bounding boxes: two observer's hands, two partner's hands, and one partner face.

Detection Accuracy

For each frame, our system estimates the location of each of the five body parts, by either providing a coordinate or indicating that it is outside the frame. We evaluate the accuracy of our method as the fraction of true positives (i.e. cases where we correctly estimate a position inside the ground truth bounding box) and true negatives (i.e. cases where we correctly predict the part to be outside the frame) over all predictions.

We also evaluate the percentage of "perfect" frames, i.e. the fraction of frames in which all five parts are predicted correctly.

Hand Disambiguation Error Rate

We are particularly interested in errors made when disambiguating the observer's hands from the partner's hands, so we measure this explicitly. We consider a ground truth hand to be a disambiguation error if it is either unlabeled, labeled as the wrong person's hand, or is marked with multiple labels of different people (falsely estimating that hands overlap). The disambiguation error rate is the total number of incorrectly disambiguated hands over

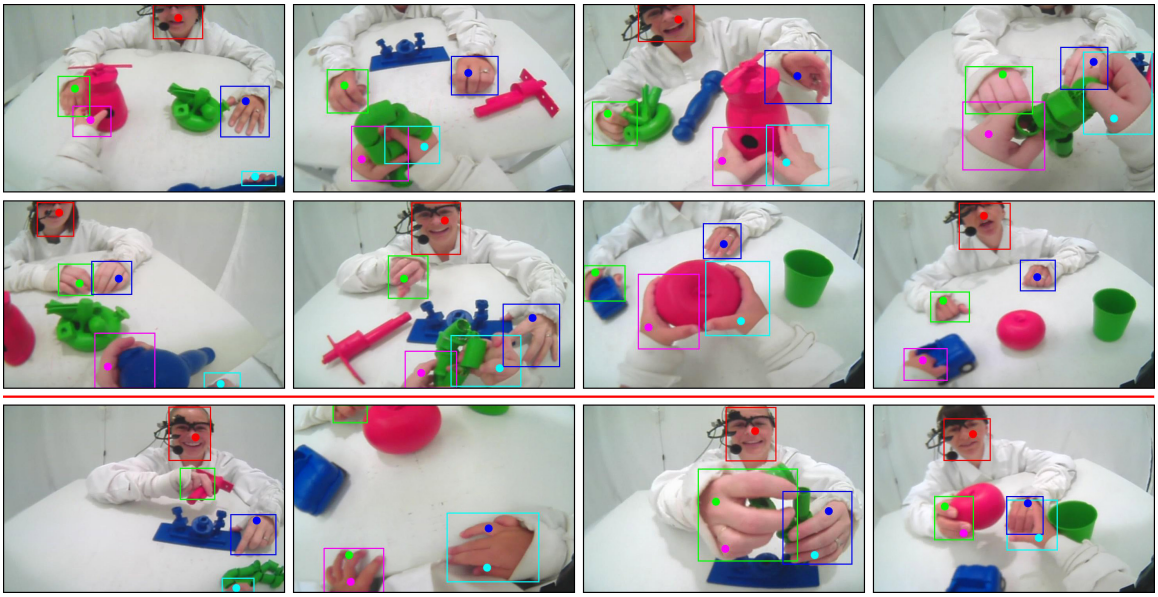


Figure 3.4: *Sample frames from our results,* with rectangles showing ground truth bounding boxes and dots showing predicted part positions (red = your face, blue = your left hand, green = your right hand, magenta = my left hand, cyan = my right hand). The first two rows show frames with perfect detection and demonstrate robustness with respect to partial occlusions and changes in hand configurations, while the bottom row shows failure cases.

the total number of hands in all frames.

3.4.2 RESULTS

We first present qualitative results on the lab dataset. Figure 3.4 shows some sample frames, where rectangles depict the ground truth bounding boxes, and dots mark our predicted position. Part identities are represented by color, so that dots inside boxes of the same color indicate correct estimates. The first two rows show perfect frames, while the last row shows some error cases. Common failures include incorrectly estimating a hand to be out of frame (e.g. the green box in the leftmost image) or falsely estimating overlapping hands. This can be caused by hands that are closer to the observer than expected and thus too big (e.g. in the middle two images), or because one hand is farther away from the other than usual (e.g. wrong prediction for 'my left hand' in the right image).

We also show qualitative results for the naturalistic videos in Figure 3.5.

Figure 3.5: *Sample results for naturalistic video,* in which two people played cards, tic-tac-toe, and puzzles, while one wore Google Glass. (See Fig. 3.4 caption for the color legend.)

Quantitative Evaluation

We present detailed quantitative results in Table 3.1. Our overall detection accuracy across the five child-parent dyads of the lab dataset is 68.4%. The technique generalized well between different dyads, as evidenced by a low standard deviation across videos ($\sigma = 3.0$). Accuracies between different hands are also fairly stable, ranging from 61.2% for 'my left hand' to 70.7% for 'my right hand.' Overall, our approach perfectly predicted 19.1% of frames on average, and for the third dyad even achieved a 24.7% perfect detection rate.

As expected, accuracy was lower for the naturalistic videos, at 50.7% overall. This drop in accuracy is caused by two factors. First, we do not use a model for the partner's hand (the edge-based method described in Section 3.3 does not work well here). Second, the simple color-based skin detection is much noisier in the natural environment compared to the controlled laboratory. To quantify this, we calculated the fraction of detected skin pixels that fall into ground truth bounding boxes of hands and faces. While 97% of detected skin pixels fall into boxes in the lab videos, only 70% do so in the naturalistic videos. Interestingly, we can still retain a relatively low disambiguation error rate in the naturalistic videos (35.6% versus 32.7%), showing that our model can compensate for noisy likelihoods.

Although our main purpose is to detect hands, the temporal and spatial constraints in our model also improve face detection. Table 3.1 compares the head-detection accuracy of our model to that of the raw Viola-Jones detector (column head^{VJ}). We achieve about a 10-percentage-point increase for the lab dataset, and an over 17-percentage-point improvement

	Overall Accuracy	**Observer** right hand	left hand	**Partner** right hand	left hand	head	head^{VJ}	**% Perfect Frames**	**Disambiguation Error Rate**
Dyad 1	64.1	50.3	60.2	68.0	54.2	87.7	86.2	14.8	37.8
Dyad 2	72.6	78.5	63.3	63.8	79.7	77.5	55.5	22.8	27.4
Dyad 3	70.1	64.2	66.7	60.5	68.8	90.0	85.5	24.7	34.5
Dyad 4	67.3	88.0	54.7	59.5	59.3	75.2	66.0	15.5	33.1
Dyad 5	68.1	72.5	61.0	66.2	60.5	80.2	69.0	17.7	30.5
Average	68.4	70.7	61.2	63.6	64.5	82.1	72.4	19.1	32.7
Natural	50.7	54.3	18.7	73.3	49.3	57.7	40.3	9.0	35.6

Table 3.1: *Detection accuracies of our approach,* as well as a breakdown into different hands. We also compare our head-detection accuracy with the accuracy of the raw Viola-Jones detector (head^{VJ}). The second to the last column shows the percentage of frames in which all five predictions were correct and the last column shows the error when differentiating the observer's hands and the partner's hands.

on the Google Glass videos.

Comparing to Baselines

We compared our model to three baselines of increasing complexity. First, we tried a simple random predictor: for every part in every frame, we first flip a coin to decide whether it is in the frame or not, and if it is in the frame, we assign it a random position. Second, we added the skin likelihood by repeating the same process but limiting the space of possible positions to be in skin patches. Finally, we build a more sensible baseline, clustering the detected skin pixels into hand-sized patches using Mean Shift [20]. Then, we greedily assign each part the position of the closest cluster centroid based on distance between centroid and part-wise absolute spatial priors.

The results of these baselines and our method are compared in Table 3.2. The two random baselines perform poorly, with accuracies of 17.0% and 27.3%, respectively. The third method using clustering and distances to centroids performs better at 58.1%, but our approach still beats it with 68.4% accuracy. We also tested a simplified version of our model in which the in-frame and between-frame links were removed, so that only absolute spatial priors and likelihoods are used. This achieved 59.1% accuracy, comparable to the third baseline (which similarly does not incorporate temporal or relative spatial constraints).

	Overall Accuracy	% Perfect Frames	Disambiguation Error Rate
Lab videos:			
random	17.0	0.1	95.1
random (skin)	27.3	4.3	72.0
skin clusters	58.1	14.4	36.0
ours (likelihood + spatial prior)	59.1	9.2	44.5
our method (full)	**68.4**	**19.1**	**32.7**
Naturalistic videos:			
skin clusters	39.2	0.0	65.4
our method	**50.7**	**9.0**	**35.6**

Table 3.2: *Comparison of our model's results to baselines,* in terms of overall accuracy, percentage of perfect frames, and hand disambiguation error rate (see text).

Our full model outperforms all the baseline methods by more then 10 percentage points for accuracy and also performs best in terms of perfect frames and hand disambiguation error.

Finally, we compare the performance of our method and the third baseline on our naturalistic videos. The baseline method suffers drastically from the noisy skin detections and does not predict a single frame perfectly. Our method does much better at overcoming weak object models, perfectly predicting almost 10% of frames, showing that it has the potential to work well in less constrained scenarios.

3.5 SUMMARY

In this section, we proposed a probabilistic graphical model (PGM) that encodes spatial and temporal constraints of hand locations in the view of an egocentric observer who interacts with a partner. We demonstrated that, given noisy initial estimates of hands (e.g. by using a simple skin color classifier), theses constraints can help to better detect hands and to distinguish between different types of hands. Overall, this approach can produce high-quality results for visually controlled video data (such as the child-parent videos introduced in Chapter 2). For more naturalistic data, the success will depend on the quality of the initial estimates. Nonetheless, we showed that our model has the potential to produce very reasonable results even if the initial estimates are extremely noisy.

CHAPTER 4

DETECTING HANDS BASED ON VISUAL APPEARANCE

4.1 INTRODUCTION

In the previous chapters, we established that spatial biases of hand trajectories in the first-person view can be utilized to better detect hands, and, importantly, to distinguish between different types of hands that may occur in egocentric interactions. The discussed approaches all assume that initial (albeit noisy) detections of hands are possible (for example by using simple skin color based models), but classification of those hands into semantic labels such as left/right hands or own/other hands based on visual features is hard. In this chapter, we investigate to what extent we can use strong appearance models to detect, distinguish, and segment different hands in first-person videos of realistic settings directly based on visual appearance. We are particularly motivated by the recent success of convolutional neural networks (CNNs), which have improved the state-of-the-art for visual object recognition by a large margin [59].

To utilize their full potential, CNN models require training on large amounts of data. Thus, to make our experiments possible, we introduce a new dataset of 48 videos featuring different participants interacting in a variety of activities and environments. The dataset includes high-quality ground truth hand segmentation masks for over 15,000 hands. To detect hands efficiently, we present a lightweight hand candidate proposal method that quickly suggests potential hand locations to the CNN and produces better results than existing ap-

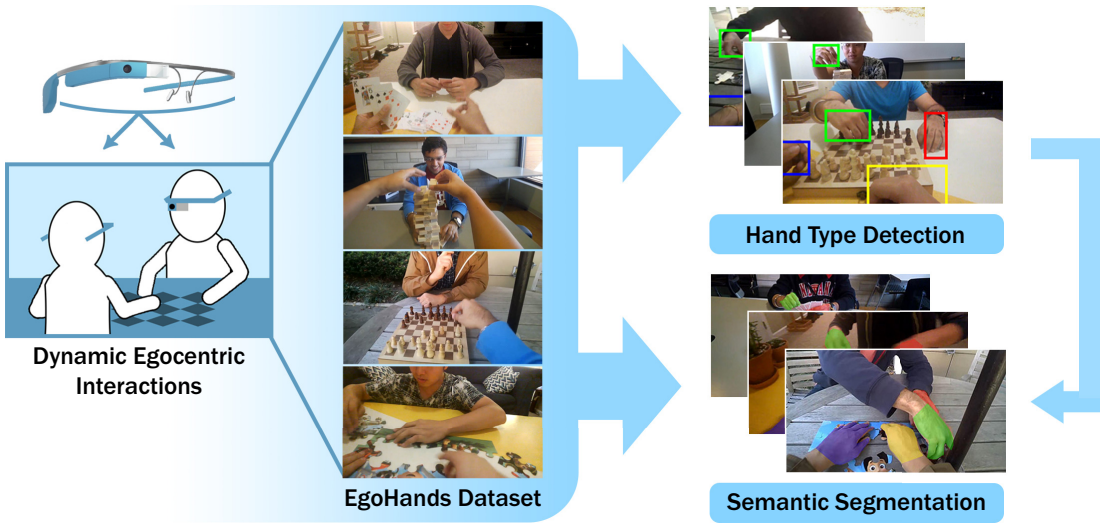


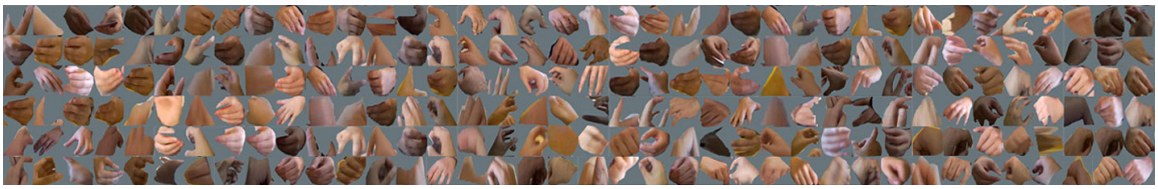
Figure 4.1: *Dataset and method overview.* We present a CNN-based technique for detecting, identifying, and segmenting hands in egocentric videos of multiple people interacting with each other. To make this possible, we introduce a new large-scale dataset with over 15,000 hands with ground truth segmentations.

proaches for general object detection [2,112] at a fraction of the computational cost. Finally, we use the resulting high-quality hand detections to perform pixel-level segmentations that outperform existing first-person hand segmentation approaches [65].

4.2 EGOHANDS: A LARGE-SCALE EGOCENTRIC HAND DATASET

We begin by presenting a new dataset, *EgoHands*, that contains high quality first-person video of interacting people in naturalistic environments. The dataset was collected with the intent to collect many visually unconstrained exemplars of hands as seen from an egocentric perspective, including the observer's hands and hands of the interaction partner. While we briefly experimented with a small set of naturalistic videos in Chapter 3, this dataset is at a scale that (1) can facilitate the training of data-driven models such as CNNs and (2) allows different partitionings into training, testing, and validation sets.

(a) Ground truth hand segmentations superimposed on sample frames



(b) A random subset of cropped hands according to ground truth segmentations

Figure 4.2: *Visualization of the EgoHands dataset.* (a) Ground truth hand segmentation masks superimposed on sample frames from the dataset, where different colors indicate the four different hand types. Each column shows a different activity. (b) A random subset of cropped hands according to ground truth segmentations (resized to square aspect ratios for ease of visualization).

4.2.1 DIFFERENCES TO OTHER DATASETS

Various other first-person imaging datasets have been proposed in the past (e.g. [31, 33, 65, 80, 90, 102]). While many of them are designed to test recognition of activities that include objects being held by the observer [33, 80], most of them do not contain high quality (i.e. pixel-level) annotations of hands. One notable exception is the dataset of Li and Kitani [65], who study pixel-level hand segmentation under varying illumination conditions. However, their dataset contains no social interactions, so that the only hands in the video belong to the camera owner, and there are no labels for hands of different semantic levels (such as left or right hand). They also define a "hand" to include any contiguous skin regions up to the sleeves, so they are really studying skin segmentation as opposed to trying to

cleanly segment only hands. This is an important distinction from our work; we argue that in applications like hand pose classification or activity recognition, segmentations that are invariant to type of clothing are important. Lastly, we provide labeled hand exemplars at a much larger scale than any previous dataset, with over 15,000 exemplars as opposed to just a few hundred [65].

4.2.2 DATA COLLECTION

To create as realistic a dataset as possible while still giving some experimental control, we collected data from different pairs of four participants who sat facing each other while engaged in different activities (see Figure 4.1). After experimenting with various options, we chose four activities that encourage interaction and hand motion: (1) playing cards, specifically a simple version of Mau Mau; (2) playing chess, where for efficiency we encouraged participants to focus on speed rather than strategy; (3) solving a 24- or 48-piece jigsaw puzzle; and (4) playing Jenga, which involves removing wooden building blocks from a tower until it collapses. Each column in Figure 4.2a contains three sample frames for each activity. We further varied context by collecting videos in three different locations: a table in a conference room, a patio table in an outdoor courtyard, and a coffee table in a home. In order to create realistic data, we did not constrain the locations in any way other than that participants had to sit at a table and face each other. We also recorded over multiple days and did not restrict participant clothing, resulting in a significant variety (e.g. both short- and long-sleeved shirts, etc.). We systematically collected data from four actors performing all four activities at all three locations while randomly assigning participants to one another for interaction, resulting in $4 \times 4 \times 3 = 48$ unique combinations of videos. Each participant wore a Google Glass device, which recorded 720×1280px video at 30Hz.

In post-processing, we manually synchronized the video pairs from both participants to

one another and cut them to be exactly 90 seconds (2,700 frames) each. To create ground truth data, we manually annotated a random subset of 100 frames from each video (about one frame per second) with pixel-level hand masks. Each hand pixel was given one of four labels: the camera wearer's left or right hand ("own left" or "own right"), or the social partner's left or right hand ("other left" or "other right"). Figure 4.2a shows examples of our ground truth hand masks, where different colors indicate the different labels. The ground truth was created by six students using a custom-built annotation tool that allowed drawing polygons for each hand. Students were told to label any hand pixels they could see, including very small hand regions caused by occlusion with objects or truncation at frame boundaries. Importantly, we defined the "hand" to stop at the wrist, in contrast to other work [64, 65] which has also included arms up to the participant's sleeves. We believe our definition is more useful and realistic in practice: if the goal is to detect hand pose and activities, for instance, the definition of what is a hand should not change dramatically depending on what a participant is wearing.

4.2.3 DATASET PROPERTIES

In total, our dataset contains 48 videos with a total of 72 minutes (or 129,600 frames) of video, of which 4,800 frames have pixel-level ground truth consisting of 15,053 hands (see Figure 4.2b for examples). The partner's hands appear in the vast majority of frames (95.2% and 94.0% for left and right, respectively), while the observer's hands are seen less often (53.3% and 71.1% for left and right). This is likely because one's own hands are more frequently outside the camera's field of view, but right hands occur more often because people tend to align their attention with their dominant hand (and all our participants were right-handed).

To our knowledge, this is the largest dataset of hands in egocentric video or any other

first-person photo collection. To enable others to also profit from the data and to encourage further research in this domain, we released the entire dataset including documented ground truth annotations online.[1] To facilitate quantitative comparisons with the detection and segmentation approaches that we will present in the following chapters, we defined a benchmark partitioning of videos into training, validation, and test groups. This partitioning has 36 training, 4 validation, and 8 test videos, with actors, activities and locations evenly distributed across groups. This is the default partitioning for our experiments and we will refer to it as "main split."

4.3 CNN-BASED HAND DETECTION

In principle, one could consider finding hands in first-person images as a simple instantiation of one particular object detection task, for which we could apply any general object detection algorithm. However, in practice, detecting hands requires some special considerations. Hands are highly flexible objects whose appearance can vary drastically (e.g. a fist as opposed to an open hand). On top of that, we are interested in detecting semantically different types of hands (i.e., left vs. right hands, and the camera wearer's own hands vs. their social partner's), such that we need visual object models that can distinguish hand variation caused by different types from variation caused by different poses (e.g. a left fist and an open left hand both belong to the same type, but a left fist and a right fist do not). Convolution Neural Networks (CNNs) offer state-of-the-art performance for visual classification tasks [59] and may be suitable for this kind of visual hand type detection.

For CNN-based object detection, one common approach is to divide an image into candidate windows, rescale each window to a fixed size, fine-tune a CNN for window classification [40,108], and then perform non-maximum suppression to combine the output of the

[1] `http://vision.soic.indiana.edu/egohands/`

region-level classifier into object detection results. Of course, the space of possible proposal windows is enormous, so it is important to propose regions that capture as many objects as possible in the fewest number of proposals. Much of the work on region proposals has studied general object detection in consumer photography, where there is typically little prior information on the location or appearance of an object in an image. In the context of detecting hands in egocentric views, however, we have shown in Chapters 2 and 3 that there are strong spatial biases to hand location and size. We thus propose a simple approach to candidate window sampling that combines spatial biases and appearance models in a unified probabilistic framework.

4.3.1 GENERATING PROPOSALS EFFICIENTLY

Our primary motivation is to model the probability that an object O appears in a region R of image I,

$$P(O|R,I) \propto P(I|R,O)P(R|O)P(O), \tag{4.1}$$

where $P(O)$ is the occurrence probability of the object, $P(R|O)$ is the prior distribution over the size, shape, and position of regions containing O, and $P(I|R,O)$ is an appearance model evaluated at R for O. Given a parameterization that allows for sampling, high quality regions can then be drawn from this distribution directly.

Here we assume regions are rectangular, so they are parameterized by an image coordinate, width, and height. Thus, we can learn $P(R|O)$ for each hand based on our training data by fitting a four-dimensional $(x, y, width, height)$ Gaussian kernel density estimator [47]. Similarly, we can estimate $P(O)$ directly from the training data as the fraction of labeled frames that contain each hand. For the appearance model $P(I|R,O)$ we define a simple color model that estimates the probability that the central pixel of R is skin, based

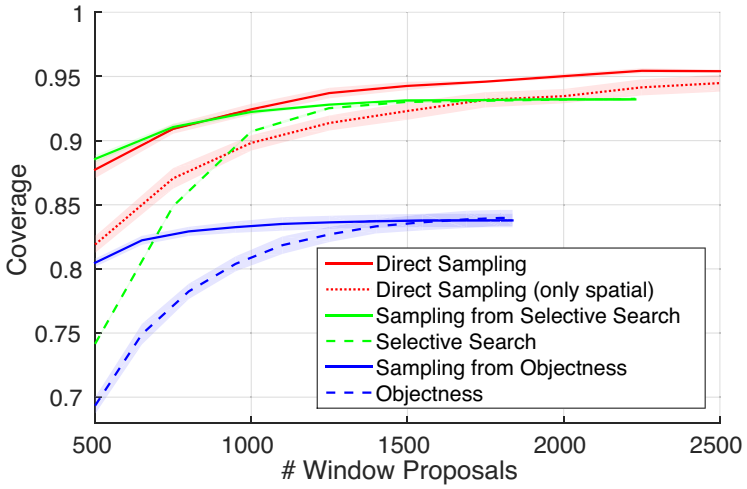


Figure 4.3: *Hand coverage versus number of proposals per frame,* for various proposal methods. Shaded areas indicate the standard deviations across five trials.

on a non-parametric modeling of skin color in YUV color space (disregarding the luminance channel). While simple, this model lets us sample very efficiently, by drawing a hand type O, and then sampling a bounding box from the KDE of $P(R|O)$, with the kernel weights adjusted by $P(I|R,O)$.

Evaluation

To evaluate this candidate generation technique, we measured its *coverage*, i.e. the percentage of ground truth objects that have a high enough overlap with the proposed windows (an intersection over union between hand bounding box and candidate region of at least 50%) to be counted as positives during detection. This is an important measure because it is an upper-bound on recall; an object that is not covered by a candidate can never be detected. Figure 4.3 shows coverage as a function of the number of proposed windows per frame for our method and two other popular window proposal methods: selective search [112] (which is the basis of the popular R-CNN detector [40]) and objectness [2]. The baselines were run using those authors' code, with parameters tuned for best results (for selective search, we used the "fast" settings given by the authors but with k set to 50; for objectness, we

retrained the object-specific weights on our dataset). As shown in the figure, our direct sampling technique (red solid line) significantly outperforms either baseline (dashed green and blue lines) at the same number of candidates per frame. Surprisingly, even our direct sampling without the appearance model (red dotted line) performed significantly better than objectness and about the same as selective search.

To further investigate the strength of the spatial consistencies of egocentric interaction, we also subsampled the baseline proposals biased by our learned model $P(O|R, I)$. For both baselines, incorporating our learned distribution improved results significantly (solid blue and green lines), to the extent that biased sampling from selective search performs as well as our direct sampling for lower numbers of proposals. However, our full technique offers a dramatic speedup, producing 1,500 windows per frame in just 0.078 seconds versus 4.38 and 7.22 seconds for selective search and objectness. All coverage experiments were performed on a machine with a 2.50GHz Intel Xeon processor.

4.3.2 WINDOW CLASSIFICATION USING CNNS

Given our accurate, efficient window proposal technique, we can now use a standard CNN classification framework, as introduced in Section 1.4.2, to classify each proposal (after resizing it to the fixed-sized input of the CNN). We used the CaffeNet architecture from the Caffe software package [49] which is a slightly modified form of AlexNet [59] as shown in Figure 1.4. We also experimented with other common network designs such as GoogLeNet [108], but found that when combined with our window proposal method, detection results were practically identical.

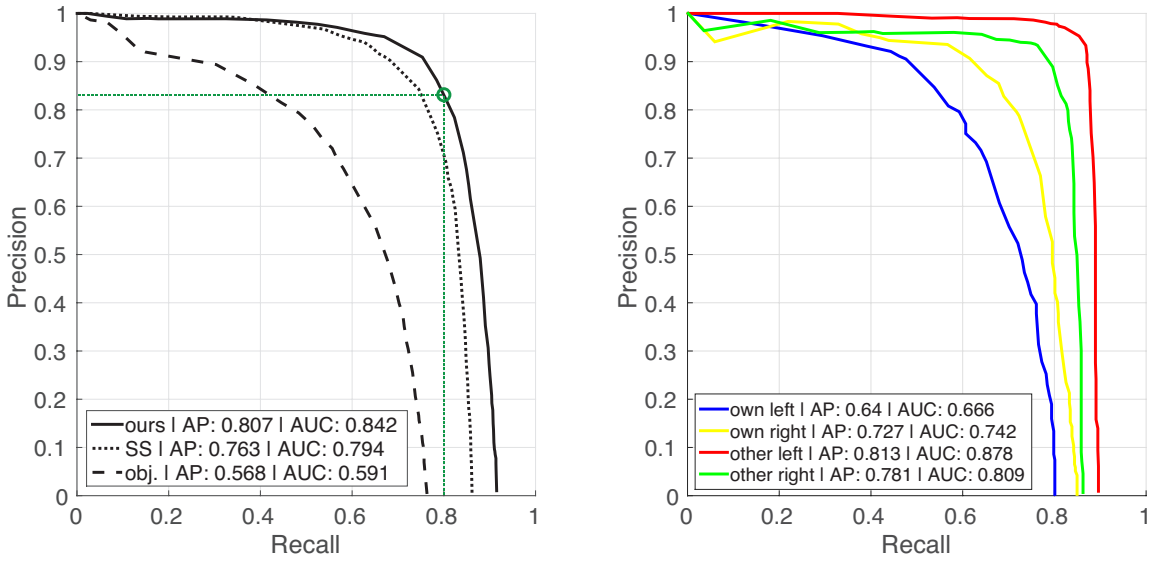
We found that certain adjustments to the default Caffe training procedure were important both to convergence and the performance of our networks. Despite the high coverage, only 3% of our proposed windows are positive so to avoid converging to the trivial major-

ity classifier, we construct each training batch to contain an equal number of samples from each class. Also, we disabled Caffe's default feature that randomly mirrors exemplar images during training. While this is a clever means of data augmentation for most visual object classes, in our case flipping images reduces the classifier's ability to differentiate between left and right hands, for example.

The full detection pipeline consists of generating spatially sampled window proposals, classifying the window crops with the fine-tuned CNN, and performing per-class non-maximum suppression for each test frame. Each of these components has a number of free parameters that must be learned. For our window proposal method, we estimate the spatial and appearance distributions from ground truth annotations in the training set and sample 2,500 windows per frame to provide a high coverage. The CNN weights are initialized from CaffeNet, meaning all network weights are set to pre-trained values from ImageNet [25] excluding the final fully-connected layer, which is set using a zero-mean Gaussian. We then fine-tune the network using stochastic gradient descent with a learning rate of 0.001 and momentum of 0.999. The network was trained until the validation set error converged. Finally, the non-maximum suppression step disregards windows with a high detection score if they are closely overlapping with other windows that have an even higher score. These overlap thresholds were optimized for each class based on average precision on the validation set. We intentionally do not take advantage of the constraint that each hand type appears at most once in a given frame in order to evaluate our technique as generally as possible.

4.3.3 DETECTION RESULTS

We evaluate the effectiveness of our detection pipeline in two contexts: detecting hands of any type, and then detecting hands of specific types ("own left," "own right," etc.). We thus train two different networks, one that only distinguishes between hands and background



(a) General hand detection

(b) Detecting different hand types

Figure 4.4: *Precision-Recall curves for detecting hands.* (a) General hand detection results with other window-proposal methods as baselines. (b) Results for detecting four different hand types.

(i.e. is trained on two classes), and one that also classifies between the four hand types (i.e. is trained on five classes). In both cases, we use the PASCAL VOC [29] criteria for scoring detections (that the intersection over union between the ground truth bounding box and detected bounding box is greater than 0.5). Figure 4.4 shows precision-recall curves for both tasks, based on the "main split" of our dataset as discussed in Section 4.2. Precision-recall curves evaluate the performance of a retrieval system (in this case the hand detector) by comparing the fraction of detected hands over all existing hands (recall) with the fraction of correctly detected hands over all proposed detections (precision), across all possible sensitivity thresholds of the system. The area under the curve (AUC) or the closely related average precision (AP, as defined in [29]) thus quantify the overall retrieval performance of the system.

For the general hand detection task (Figure 4.4a), we obtain an average precision (AP) of 0.807 using our candidate window sampling approach, which is significantly higher than the 0.763 for selective search [112] and 0.568 for objectness [2]. These overall detection

results are quite strong and underline the power of CNNs for visual object classification. For example, we can correctly detect 80% of all hands with only 18% false positives (as indicated by the green circle in the figure).

Figure 4.4b shows precision-recall curves for distinguishing between the four hand types. There is a curious asymmetry in our hand type detections, with our approach achieving significantly better results for the social partner's hands versus the camera owner's. Figure 4.5 gives insight on why this may be, presenting detection results from randomly-chosen frames of the test set. Hands of the camera wearer tend to have more duplicate detections on subparts of the hands (e.g. in row 2, column 2 of the figure). We attribute this tendency to how frequently "own" hands are truncated by the frame boundaries and thus appear as single or only a few fingers in the dataset. Including these partial detections alongside fully visible hands during training encourages the network to model both appearances to minimize error. While this does result in a loss of precision, the system gains the ability to robustly detect hands that are occluded or only partially in the frame (e.g. row 3, column 3) which is often the case for egocentric video, due to the relatively narrow field of view of most cameras compared to that of humans.

Error Analysis

Overall, the average detection performance across hand types in Figure 4.4b is a little lower than general hand detection performance reported in Figure 4.4a (AP 0.740 versus 0.807). An interesting question is whether this difference is primarily caused by failure to detect hands of different types or confusion between hand types once a hand is detected. An analysis of the per-window classifications showed that only 2% of hand windows are mislabelled as other hands. Similarly for detection, 99% of undetected hands at a recall of 70% are due to confusion with the background class. In those rare cases with ambiguous

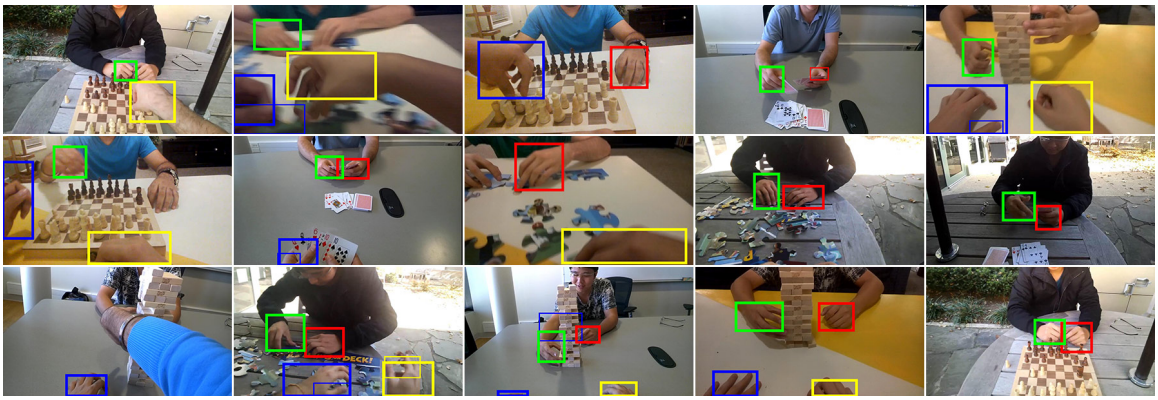


Figure 4.5: *Randomly-chosen frames with hand detection results,* for own left (blue), own right (yellow), other left (red), and other right (green) hands, at a detection threshold where recall was 0.7. Thick and thin rectangles denote true and false positives, respectively.

hand types, the system predicts nearly uniform probabilities for each type, which are then removed by reasonable decision thresholds and non-max suppression, decreasing the overall performance. However, the system almost never confidently predicts the wrong type of hand, which is also evidenced by the randomly-sampled examples shown in Figure 4.5.

Generalizing Across Actors, Activities, and Locations

We next tested how well our hand detectors generalize across different activities, different people, and different locations. To do this, we generated dataset partitionings across each dimension, where each split leaves out all videos containing a specific (first-person) actor, activity, or location during training, and tests only on the held-out videos. We also split on actor pairs and activities jointly, creating 18 divisions (as not all actor pairs did all activities). This stricter task requires our method to detect hands of people it has never seen, doing activities it has never seen.

Table 4.1 summarizes our results, again in terms of average precision (AP), with averages across splits weighted by the number of hand instances. The table shows that detectors generalize robustly across actors, with APs in a tight range from 0.790 to 0.826 no matter which actor was held out. This suggests that our classifier may have learned general

characteristics of human hands instead of specific properties of our particular participants, although our sample size of four people is small and includes limited diversity (representing three different ethnicities but all were male). For locations, the courtyard and office environments were robust, but AP dropped to 0.648 when testing on the home data. A possible explanation is that the viewpoint of participants in this location is significantly different, because they were seated on the floor around a low table instead of sitting in chairs. For activities, three of the four (cards, puzzle, and chess) show about the same precision when held out, but Jenga had significantly lower AP (0.665). The Jenga videos contain frequent partial hand occlusions with the tower, and the tower itself is prone to be mistaken for hands that it occludes (e.g. row 3, column 3 of Figure 4.5). Finally, splitting across actor pairs and activities results in a sharper decrease in AP, although results are still quite reasonable given the much smaller (about 6×) training sets caused by this strict partitioning of the data.

4.4 SEGMENTING HANDS

While simply detecting hands may be sufficient for some applications, pixel-wise segmentation is often more useful, especially for applications like hand pose recognition and in-hand object detection [68]. In Chapters 2 and 3 we did not talk about segmentation explicitly, mostly because the child-parent data considered in these chapters was visually clean and controlled enough to make the problem relatively easy. Pixel-wise segmentation of objects in unconstrained, natural photos and videos is a much harder computer vision problem. Fortunately, as shown in the previous section, we do not have to start from scratch, but instead can utilize the strong performance of our CNN-based hand type detector. Once we have accurately localized hands using this approach, semantic segmentation is relatively straightforward, as we can (1) focus segmentation efforts on local image regions and (2) get

	All hands	Own hands Left	Own hands Right	Other hands Left	Other hands Right
Main split	0.807	0.640	0.727	0.813	0.781
All activities but:					
cards	0.768	0.606	0.776	0.708	0.732
chess	0.851	0.712	0.788	0.821	0.808
Jenga	0.665	0.644	0.693	0.583	0.502
puzzle	0.803	0.747	0.813	0.675	0.681
weighted average	0.772	0.675	0.768	0.699	0.686
All actors but:					
B	0.799	0.669	0.773	0.779	0.796
H	0.816	0.718	0.772	0.756	0.740
S	0.790	0.709	0.798	0.799	0.696
T	0.826	0.689	0.783	0.770	0.789
weighted average	0.807	0.700	0.782	0.776	0.756
All locations but:					
courtyard	0.790	0.702	0.785	0.755	0.755
office	0.772	0.659	0.757	0.794	0.687
home	0.648	0.558	0.703	0.538	0.591
weighted average	0.737	0.639	0.748	0.698	0.678
Split across actor pairs and activities:					
weighted average	0.627	0.492	0.598	0.513	0.542

Table 4.1: *Hand detection accuracy when holding out individual activities, participants, and locations,* in terms of average precision. For example, the training set for *all activities but cards* included all videos *not* containing card playing, while the test set consisted *only* of card playing videos.

the correct hand type label for each segment for free.

4.4.1 REFINING LOCAL SEGMENTATIONS WITH GRABCUT

Our goal is to label each pixel as belonging either to the background or to a specific hand class. We assume that most pixels inside a box produced by our CNN-based detector correspond with a hand, albeit with a significant number of background pixels caused both by detector error and because hands rarely fill a bounding rectangle. This assumption allows us to apply a well-known semi-supervised segmentation algorithm, GrabCut [93], to our problem.

GrabCut was proposed as an interactive figure-ground segmentation tool. Given a photo of a foreground object with relatively distinct colors compared to the background (e.g. a person standing on a green meadow), a user would initialize the algorithm by manually

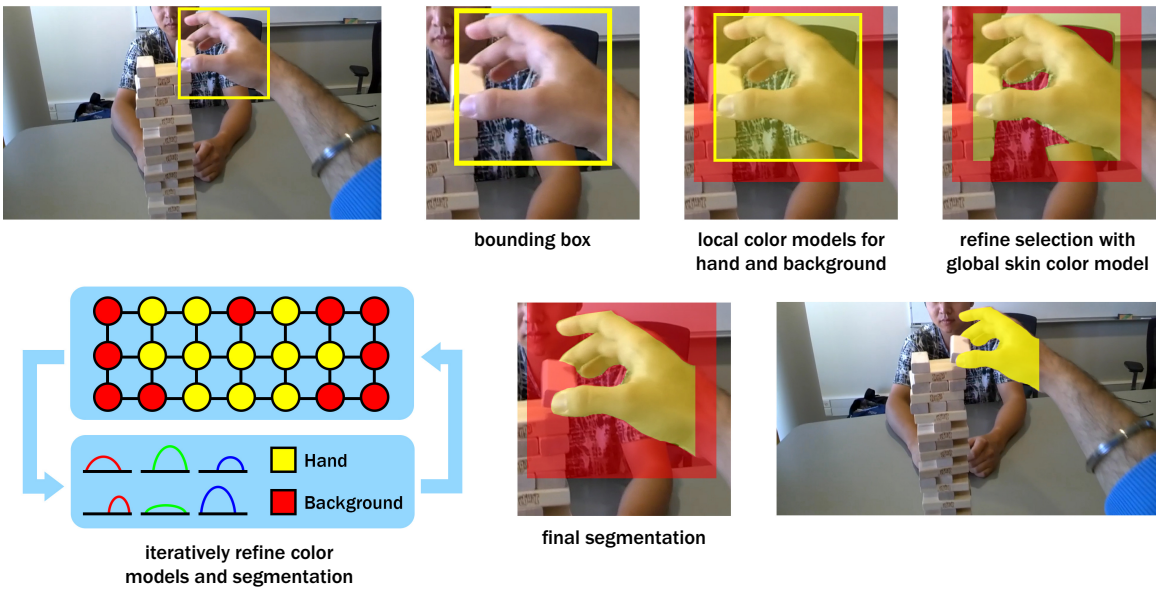


Figure 4.6: *Semantic hand segmentation using GrabCut.* We utilize the CNN-based hand type detector to initialize the GrabCut [93] algorithm with local color models for hands and background, and to provide the right hand type for each segment.

drawing a coarse segmentation (e.g. a bounding box) around the object. The color distributions of foreground (all pixels within the bounding box) and background (all pixels outside the bounding box) are then represented as Gaussian Mixture Models (GMMs) [91] in RGB space. To refine the user segmentation, all pixels within the bounding box are represented as a PGM with a grid structure (a Markov Random Field), and assigned to either foreground or background according to the likelihood given by the color models, subject to the constraint that neighboring pixels should have the same label. Since labels are binary, exact inference on the PGM is possible using the Graph Cut algorithm [13]. The refined segmentation is then used to repeatedly update the color models and relabel foreground and background pixels under the new models, until convergence to the final segmentation.

Instead of having any user interaction, we initialize the GrabCut procedure with the bounding boxes from our hand detector, as summarized in Figure 4.6. For each hand bounding box, we first use the simple global skin color model as described in Section 4.3.1 to estimate an initial foreground mask. We use an aggressive threshold so that all pixels

within the box are marked foreground except those having very low probability of being skin. Importantly, we avoid initializing the background color model with the entire image outside of the bounding box because arms, faces, and other hands would likely lead to confusion with the foreground model. Instead, we use a padded region (marked as red in Figure 4.6) around the bounding box, ensuring that only local background content is modeled. This procedure is done for each detected box separately. If there are multiple detected boxes for hands of same type, we take the union of all output masks as the final segmentation. Should masks of different hand types overlap, we prefer the label of the hand that had a stronger detection score.

4.4.2 SEGMENTATION RESULTS

Using the CNN hand detector (at a recall of 0.7) and the global skin color model trained on the training set of the "main split" of our data, we detected hands and produced segmentations for each frame in our test set. To put our results in context, we ran the publicly-available pixel-wise hand detector of Li et al. [65], which was designed for first-person data. We trained their technique with 900 randomly-sampled frames from our training set. As we mentioned before, Li et al. [65] defines "hand" to include any skin regions connected to a hand, including the entire arm if it is exposed. To enable a direct comparison to our more literal definition of hand detection, we took the intersection between their method's output and our bounding boxes. Finally, as [65] is agnostic with respect to the semantic labels of each hand, we use our inferred labels for the comparison.

Table 4.2 presents segmentation accuracy, in terms of pixel-wise intersection over union between the estimated segmentation mask and the ground truth annotations. Our technique achieves significantly better accuracy than the baseline of [65] (0.556 versus 0.478). A similar trend is present across the stricter actor pair and activity data splits. We attribute

	Own hands		Other hands		
	Left	Right	Left	Right	Average
Main split:					
Ours	0.515	0.579	0.560	0.569	0.556
Li et al. [65]	0.395	0.478	0.534	0.505	0.478
Split across actor pairs and activities:					
Ours	0.357	0.477	0.367	0.398	0.400
Li et al.	0.243	0.420	0.361	0.387	0.353

Table 4.2: *Hand segmentation accuracy* measured in terms of pixel-level intersection over union with ground truth masks.

this success to the fact that our GrabCut-based approach looks only at local image color distributions and leans heavily on the quality of our detections. The baseline method, however, learns classifiers that must perform well across an entire frame, which is complicated by the close visual similarity between hands and other visible skin. Figure 4.7 provides qualitative examples of our segmentations based on some randomly-sampled test frames.

Failure Modes

Our method has two main possible failure modes: failure to properly detect hand bounding boxes, and inaccuracy in distinguishing hand pixels from background within the boxes. To analyze the influence of each, we performed an ablation study based on the ground truth annotations. Applying our segmentation approach to the ground truth hand bounding boxes instead of the output of the hand detector, our average accuracy rose from 0.556 to 0.73. On the contrary, taking the output of our hand detector but using the ground truth segmentation masks (by taking the intersection with the detected boxes) achieved and accuracy of 0.76. Thus, each of the studies improve over our fully automatic approach by roughly 30-35%, indicating that neither detection nor segmentation is individually to blame for the decrease in accuracy, and that there is room for future work to improve upon both.

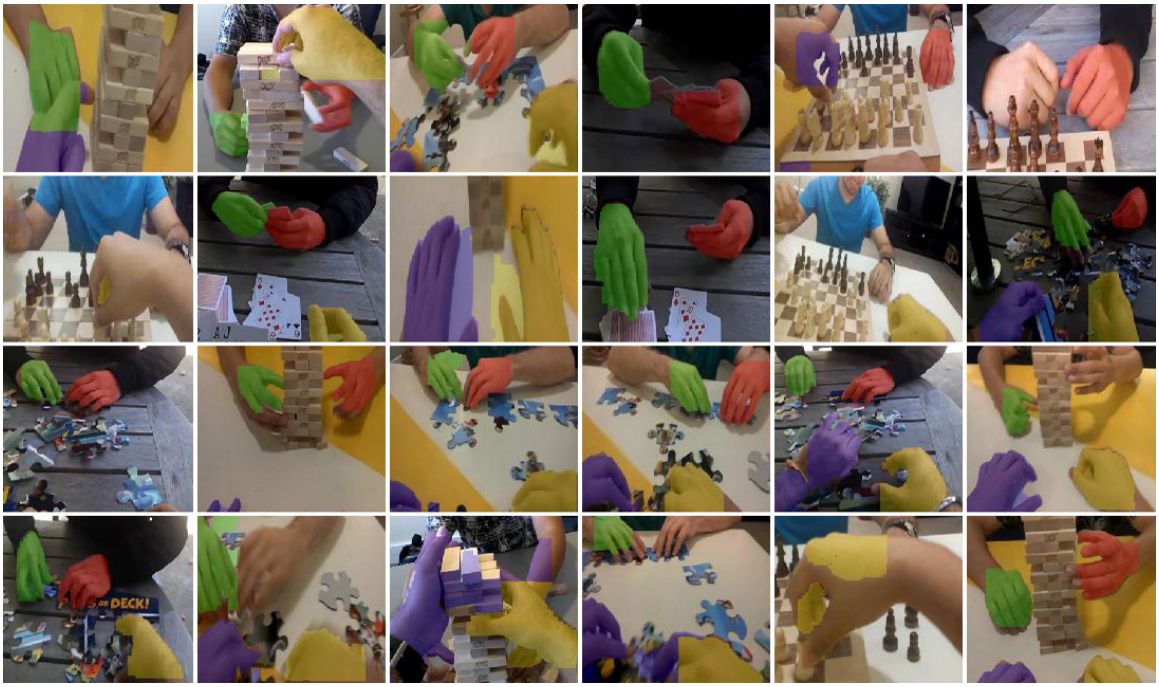


Figure 4.7: *Hand segmentation results on randomly-chosen test frames*, zoomed into areas containing hands for better visualization.

4.5 CONCLUSION

In this chapter we demonstrated how to detect and distinguish hands in natural and dynamic first-person videos, by combining CNN-based classification with a fast candidate region proposal method based on sampling from a joint model of hand appearance and geometry. This approach is significantly more accurate and dramatically faster than existing region proposal techniques based on selective search [112] or objectness [2]. An important difference to the graphical model proposed in Chapter 3 is that instead of relying on spatial biases to infer hand types, here we distinguish hands based on visual information alone, and use spatial biases of hand locations primarily to locate hands more efficiently. We also showed that our hand detections can be used to yield state-of-the-art hand pose segmentations. Finally, we introduced a novel, publicly available first-person dataset with dynamic interactions between people, along with fine-grained ground truth.

CHAPTER 5

USING HANDS TO INFER SOCIAL ACTIVITIES

5.1 INTRODUCTION

In the last chapter, we demonstrated that we can use convolutional neural networks to robustly detect hands in egocentric videos, even in very naturalistic data. Importantly, we showed that our CNN-based technique is able to distinguish high level concepts such as left and right hands, and observer hands and other hands, without taking into consideration any context other than the visual information of the captured hand. As we are also able to segment the shape of each hand and thus extract information about its two-dimensional pose, one interesting question is whether there is other high level information that we can infer from automated visual analysis of hands and hand poses. In this chapter, we explore this question by further experimenting with the *EgoHands* dataset that we introduced in Chapter 4. More precisely, we investigate whether we can infer social activities (specifically the four in the dataset: cards, chess, puzzle, Jenga) by analyzing only the pose and position of hands within the egocentric view. As interacting with different objects affords different types of hand grasps (the taxonomies of which have been thoroughly studied [76]), the 2D pose of a hand should to some extend contain high level information about the type of interaction. Moreover, when multiple people are interacting with each other, it seems likely that the absolute and relative position of all hands within in the field of view should also reveal some evidence about the activity they are engaged in.

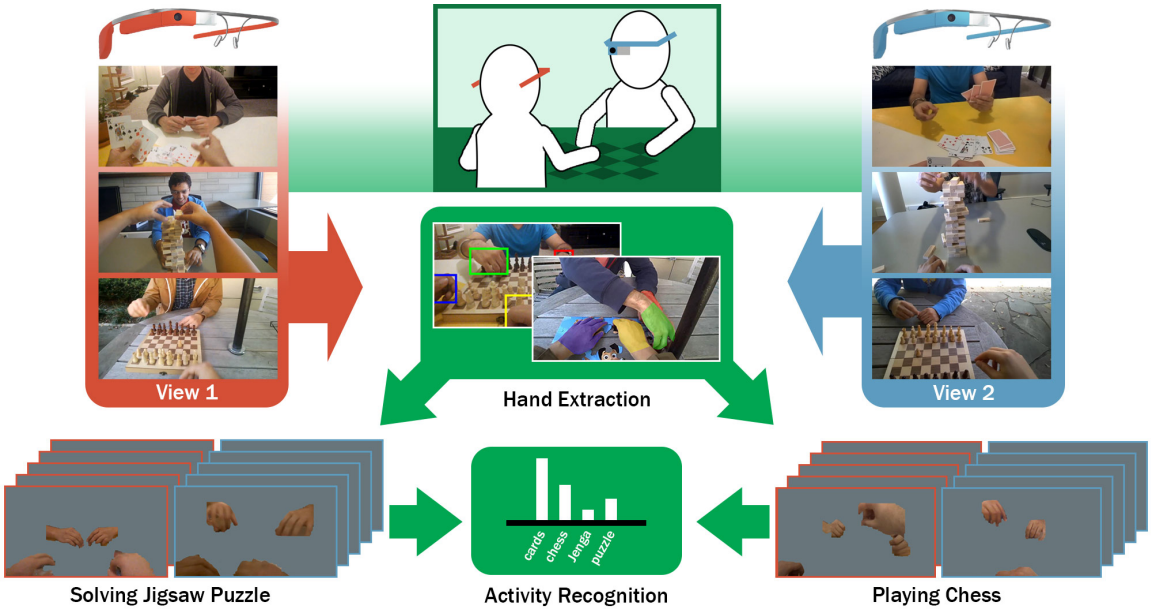


Figure 5.1: *Hand-based activity recognition overview.* Two actors are engaged in different social interactions, while both wear Google Glass to capture video from each field of view. We present a vision-based framework that extracts hands from each view to investigate how well we can estimate the performed activity based on hand pose and position alone.

Our main hypothesis is that hand poses by themselves reveal significant evidence about the objects people are interacting with and the activities they are doing. This would imply that automatic activity recognition systems could focus on accurately recognizing one type of object – the hands – instead of having to model and detect the thousands of possible objects and backgrounds that occur in real-world scenarios. While the hand poses in any given video frame may not necessarily be informative, we hypothesize that integrating hand pose evidence across frames and across viewpoints may significantly improve activity recognition results. We investigate (1) how well activities can be recognized in our dataset based on hand pose information alone, and (2) whether the two first-person viewpoints can be complementary with respect to this task.

5.2 RECOGNIZING FIRST-PERSON HAND POSES WITH CNNS

To explore if hand poses can uniquely identify activities, we create masked frames in which all content except hands is replaced by a gray color. Some examples of masked frames are shown in Figure 5.1 and also Figure 5.2. These frames contain information about the relative position of hands in the frame (and to each other) as well as the 2D pose of each hand, but no other visual context. Since CNNs showed a strong performance in the hand type detection task, we again make use of the AlexNet [59] CNN architecture described in Section 4.3.2. However, instead of classifying cropped-out region proposals into different hands or background, here we feed the entire masked frame to the network and directly classify between the four activities (cards, chess, puzzle, Jenga) it belongs to.

Because we are also interested in aggregating information across both participants' viewpoints, we create a new partitioning of the *EgoHands* dataset that ensures that corresponding viewpoints are grouped together. We split the 48 videos into 16 test videos (8 videos per viewpoint), 24 training videos (12 per viewpoint) and 8 validation videos (4 per viewpoint), such that each set has an equal amount of each activity.

In the training phase, we used ground truth hand segmentations to create masked hands to prevent the classifier from learning any visual bias not related to hands (e.g. portions of other objects that could be visible due to imperfect hand extraction). With 100 annotated frames per video and 24 videos in the training set, this led to a total of 2,400 training images (600 per activity). Network weights were initialized with pre-trained values from ImageNet [25] and all other learning parameters were as described in Section 4.3.2. The network was trained with stochastic gradient descent until the accuracy on the validation videos converged, wich occured after around 12 epochs, i.e. after observing the entire training data about 12 times.

5.3 EXPERIMENTS

To test the performance of the trained CNN, we performed a series of experiments including single frame predictions with ablation studies, as well as predictions based on multiple frames and viewpoints.

5.3.1 SINGLE FRAME PREDICTION AND ABLATION STUDY

One interesting question is how well activities can be inferred from hands in a fully automated system. To test this, we first applied the hand detection and segmentation pipeline as described in Chapter 4 to each frame of all 16 videos in our test dataset, resulting in $16 \times 2{,}700 = 43{,}200$ masked test frames. Classifying each frame individually gave a 53.6% accuracy, nearly twice the random baseline (25.0%). This promising result suggests a strong relationship between hand poses and activities.

This fully automated approach of course suffers from the same types of errors as discussed in Section 4.4.2: incorrect information about the spatial configuration of the hands due to imperfect detection, and incorrect hand pose information due to imperfect segmentation. We once again investigated the relative effect of these errors with an ablation study, and tested the network on the subset of frames with ground truth hand data ($16 \times 100 = 1{,}600$ masked test frames). We found that replacing either detection or segmentation with ground truth increased the fully automatic performance by about nine percentage points. This suggests that capturing the spatial arrangement of hands and correctly predicting their pose are equally important to per-frame activity recognition using only hand information. Finally, a perfect hand extraction system (simulated by using the full ground truth data) could improve the performance by around 16 percentage points.

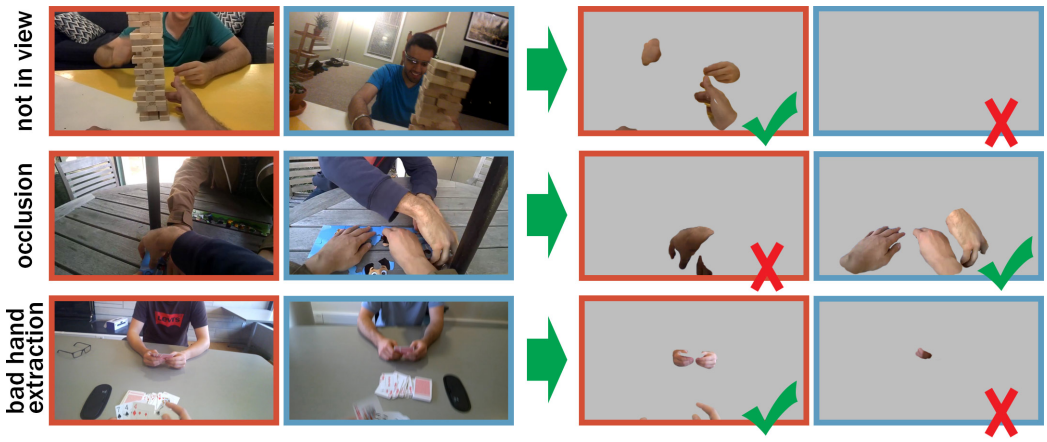


Figure 5.2: *Viewpoint differences.* Different examples of moments where one viewpoint observes much more informative hand poses than the other.

5.3.2 INTEGRATING INFORMATION ACROSS FRAMES

As discussed in the previous section, our automatic hand extraction is not without errors. Even when hand masks are perfect, they may not always be informative: Hands may be occluded or not in view at all (see Figure 5.2 for examples). However, even if the hands are occluded or inaccurately extracted in one frame, it is likely that another frame, either from the other person's view or from a nearby moment in time, yields a more informative estimate of the hand pose.

We integrate evidence across frames using a straightforward late fusion method at the decision level. Suppose we have a set $\mathcal{P}$ of actors, each of whom records a sequence of n frames, i.e. $\mathcal{F}_p = (F_p^1, F_p^2, ..., F_p^n)$ for each $p \in \mathcal{P}$. The frames are synchronized so that for any t and pair of actors $p, q \in \mathcal{P}$, F_p^t and F_q^t were captured at the same moment. Without loss of generality, we consider the specific case of two actors, $\mathcal{P} = \{A, B\}$. Suppose that our goal is to jointly estimate the unknown activity label H from a set of possible activities $\mathcal{H}$. By applying the CNN trained in the last section on any given frame F_p^t, we can estimate (using only the evidence in that single frame) the probability that it belongs to any activity $h \in \mathcal{H}$, $P(H = h|F_p^t)$.

Temporal Integration

We integrate evidence across the temporal dimension, given the evidence in individual frames across a time window from t_i to t_j in a single view p,

$$\begin{aligned}\hat{H}_p^{t_i,t_j} &= \arg\max_{H\in\mathcal{H}} P(H|F_p^{t_i}, F_p^{t_i+1}, ..., F_p^{t_j}) \\ &= \arg\max_{H\in\mathcal{H}} \prod_{k=t_i}^{t_j} P(H|F_p^k),\end{aligned} \tag{5.1}$$

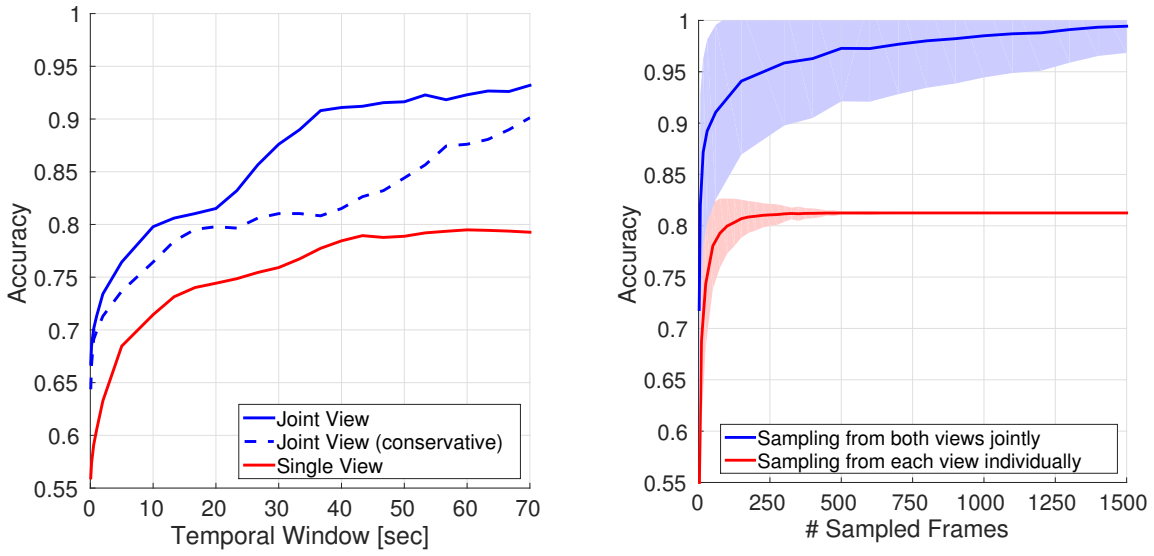
where the latter equation follows from assumptions that frames are conditionally independent given activity, that activities are equally likely *a priori*, and from Bayes' Law. We evaluated this approach by repeatedly testing classification performance on our videos over many different time windows of different lengths (different values of $|t_j - t_i|$). The red line in Figure 5.3a shows that accuracy increases with the number of frames considered. For instance, when observing 20 seconds of interacting hands from a single viewpoint, the system predicts the interaction with 74% accuracy.

Viewpoint Integration

Next we take advantage of the coupled interaction by integrating evidence across viewpoints,

$$\hat{H}^{t_i,t_j} = \arg\max_{H\in\mathcal{H}} \prod_{k=t_i}^{t_j} P(H|F_A^k)P(H|F_B^k), \tag{5.2}$$

which makes the additional assumption that the viewpoints are independent conditioned on activity. We again test over many different temporal windows of different sizes in our videos, but now using frames from both viewpoints. The results are plotted in blue in Figure 5.3a and clearly outperform the single view approach, showing that the two views are indeed complementary. However, this fusion method has the potentially unfair advantage of seeing twice as many frames as the single view method, so we also show a more conservative line



(a) Temporal window

(b) Sampling frames

Figure 5.3: *Comparison of activity recognition accuracies*, using one (red) and both (blue) viewpoints, using (a) a sliding temporal window and (b) sampling nonadjacent frames.

(dashed blue) that considers half the temporal window size (so that for any position on the x-axis, the red line and dashed blue line are seeing the same number of frames). This conservative comparison still outperforms the single view, demonstrating that observing $x/2$ seconds of hand interactions from both viewpoints is more informative than observing x seconds from only one.

Sampling Frames

Temporally adjacent frames are highly correlated, so another interesting question is how classification performance changes when sampling frames across the whole video, thus integrating evidence across wider time periods. A generalization of Equation 5.2 is to use a set $\mathcal{T} \subseteq [1, n]$ of randomly sampled times at which to observe frames,

$$\hat{H}^{\mathcal{T}} = \arg\max_{H \in \mathcal{H}} \prod_{k \in \mathcal{T}} P(H|F_A^k) P(H|F_B^k). \tag{5.3}$$

We tested this by repeatedly sampling different numbers of frames from each video. The red line in Figure 5.3b shows accuracy as a function of number of frames ($|\mathcal{T}|$) for single viewpoints (with the shaded area indicating standard deviation over 2,700 sampling iterations). After a high initial variance, accuracy converges to 81.25%, or 13 of 16 videos, when sampling around 500 frames (about 20% of a video). Of the three incorrect videos, two are of chess (where we predict puzzle and cards) and one is of cards (where we predict Jenga).

Finally, we combine both viewpoints together by sampling sets of corresponding frames from both views. The blue line in Figure 5.3b shows the results (plotted so that at any position on the x-axis, the red line sees x frames while the blue line sees $x/2$ frames from each viewpoint). Even for a small number of samples, this method dramatically outperforms single view, albeit with a large standard deviation, indicating that some paired samples are much more informative than others. More importantly, as the number of samples increases, the joint view method approaches 100% accuracy. This means that using the complementary information from both viewpoints helps correctly predict the three videos that were not correctly predicted with the single view.

5.4 SUMMARY

Our results demonstrate that (1) hands in the field of view can be informative visual cues for inferring higher level semantic information, such as the type of social interaction between two people, and (2) CNNs are powerful enough to interpret these cues. More precisely, we show that we can build a fully automated computer vision system that robustly extracts information about hand pose and position, and use it to distinguish between the four activities present in our first-person video data. Further, we demonstrate that the two viewpoints are complementary and that predicting the interaction based on integrating evidence across viewpoints leads to better results than analyzing them individually.

CHAPTER 6

CONCLUSION

6.1 THESIS SUMMARY

This dissertation began with the idea that wearable cameras are becoming increasingly less intrusive and thus can be used as tools to study visual attention and perception in dynamic real-world contexts, in addition to strictly controlled laboratory environments. However, the downside of this newly gained freedom is that the vast amounts of collected video data are often hard to analyze. Computer vision techniques could potentially help to overcome this problem by annotating data automatically. Reviewing related literature revealed that computer vision researchers have identified many challenges unique to data captured by first-person cameras, most of them due to the fact that egocentric images and videos often lack the clean characteristics of traditional photography.

We identified one particular domain that has a rich history in computer vision but has received relatively little attention within the first-person context: the analysis of hands. Hands are arguably the most frequently present objects in our field of view and the primary tool of physical interaction with the world around us. Thus, first-person cameras can provide a unique embodied perspective of one's own hands and how they are perceived. We further argued that hands are closely tied to visual attention and also motivated the analysis of hands from a cognitive perspective. We reviewed work that, for example, shows that infants' eye gaze tends to follow both their own hands and their caretaker's hands in

order to establish joint attention towards toy objects [122], and that adults are faster to fixate on target objects if they appear near their own hands [86].

We proposed and investigated different methods to extract information about hands from first-person videos, particularly focusing on videos containing interactions between two people. This focus allowed us not only to examine the detection and segmentation of hands, but also address questions concerning higher-level understanding of a detected hand, such as determining whether it belongs to the camera wearer or the other person. We demonstrated that this information can be useful using a dataset of first-person infant videos that also contained infant eye gaze data. Automatically collecting fine-grained statistics of hands within the infant's view revealed where and when the infant's own hands guided visual attention, and where and when it was guided by the parent's hands.

We introduced two major vision-based approaches for automatic hand type detection in naturalistic first-person videos. The first approach was based on the idea of building a probabilistic graphical model (PGM) to encode temporal and spatial constraints of hands in the field of view. For example, a left hand is more likely to appear in the left side of one's view and also more likely to be to the left of the right hand. We demonstrated that encoding this information in our model can help reducing uncertainty in locating hands and distinguishing among different types of hands. The second approach was based on training convolutional neural networks (CNNs) for our task. Such networks are very powerful models for visual recognition, but depend on supervised learning that requires large amounts of labeled data. We collected a large-scale egocentric dataset that included fine-grained annotations of different types of hands (left vs. right, own hands vs. other hands), and demonstrated that a properly trained CNN can detect hands and distinguish among types of hands based on visual information alone.

Finally, we demonstrated that a CNN could also infer other higher-level semantics, such

as the type of interaction displayed in the video, based on the position and shape of hands in view. This result underscores the importance and the informative potential of hands with respect to our overarching goal of automatically analyzing data from first-person cameras.

In the remainder of this chapter, we round off the dissertation with a comparison of the proposed approaches from Chapter 3 and Chapter 4, and also give some practical recommendations on when to prefer one over the other. Finally, we give an outlook into possible future work, both directly tied to this thesis, but also including thoughts that go beyond the work presented here.

6.2 METHOD COMPARISON: DEEP OR SPATIAL?

Chapters 3 and 4 introduced two different approaches to hand type detection in first-person videos. While we described both methods in great detail separately, we have yet to compare them explicitly. This section aims to fill this gap.

The PGM introduced in Chapter 3 aimed to help distinguish between (potentially noisy) detections of different hands by encoding spatial relationships between them. The CNN (from Chapter 4) then demonstrated that this distinction can actually be done robustly without encoding any additional information other than local appearance information. This naturally raises the question if there is still any benefit in modeling spatial constraints at all. There are two ways to approach this question.

The first way is from a purely practical perspective. As presented in this thesis, the PGM-based approach can process an entire video in near real-time on a conventional CPU. The CNN-based approach as presented here requires a few seconds per frame on a high-end GPU. At the same time, the PGM only provides location estimates in the form of an object-center coordinate while the CNN provides a more informative bounding box detection. For applications where coordinate estimates are sufficient (e.g. the toddler data presented in

this thesis), or for applications where time and computational resources are limiting factors, the PGM-based approach might be the preferable solution. However, it is worth noting that there is currently a large amount of research (some of which we will address in Section 6.3) on improving deep network architectures to allow much faster processing speeds, albeit likely at the cost of decreased accuracy and still requiring the processing power of expensive GPUs. Another possible practical benefit of the PGM approach is that it includes implicit tracking of hands (by encoding temporal constraints on hand locations), while the CNN currently treats each frame independently. In scenarios where hands are likely to overlap, the PGM has the potential to infer the position of an occluded hand based on its temporal trajectory, which the CNN as presented here does not.

The second way to approach the question of whether there is any benefit to modeling spatial constraints is from a more theoretical perspective. Can adding spatial information (in whichever way) to the CNN-based method improve detection or decrease errors resulting from confusing hand types? One way to investigate this would be to feed the hand location estimates derived from the CNN into the graphical model. After all, the PGM is agnostic with respect to where the estimates come from as long as they can be expressed as a probability distribution. We performed some preliminary experiments in this direction, but instead of first running the CNN detection pipeline and then adding spatial constraints, we tried to directly inject spatial information into the deep network. As described in Section 4.3.2 of the CNN-based method, different window proposals are cropped from the frame and then fed to the network separately. Thus, the spatial information of each window (its xy-position within the frame, but also its width and height) is implicitly available. We experimented with different network architectures that would receive this information (via additional input neurons) in combination with the image input and thus could potentially learn to utilize spatial biases on top of the visual information. However, preliminary experi-

ments based on training such CNNs using the *EgoHands* dataset from Section 4.2 indicated no improvement in hand type classification over the method described in Chapter 4. At the same time, training a simple two-layer neural network with only the spatial coordinates of each window proposal showed that networks can in principle learn the spatial biases of different hand types in view. Taken together, these results suggest that CNNs are usually confident enough in distinguishing different types of hands based on visual appearance that additional spatial information, although useful on its own, is not a strong enough signal to change the network's decision.

6.3 FUTURE WORK

The work presented in this dissertation opens up many different directions for future work, both on the level of improving upon proposed computer vision algorithms, and on the level of utilizing first-person hand analysis and first-person computer vision in general as tools to aid cognitive and behavioral research. We conclude with some final thoughts on both of these levels.

6.3.1 HANDS

As hinted to in Section 1.4.2, deep learning and convolutional neural networks in particular are currently dominating research efforts in many domains of computer vision, and are quickly driving progress in areas potentially relevant to the type of first-person hand analysis presented here. One interesting area of future work relates to further improving the bounding box-based hand detection described in Section 4.3. Our approach relies on first generating a set of proposal boxes and then processing those boxes with the CNN in a subsequent step. While we demonstrated that our method is faster and more suitable for first-person hand detection than other proposal methods [2, 112], there is an emerging line

of research on potentially even faster methods that do not rely on proposals at all [85, 88]. The core idea among these approaches is to design network architectures and corresponding loss functions that aim to regress window locations which maximize object detection scores during training. For example, Redmon *et al.* [85] divide the input image into an $n \times n$ grid where each grid cell regresses k bounding boxes anchored around the cell. Such a grid design might be particularly interesting for hand detection in first-person images as it has the potential to implicitly encode spatial biases of objects.

Another potential area for future research relates to hand segmentation. Our approach as described in Section 4.4 relies on local segmentations based on prior coarse detections. While similar ideas (e.g. Arbeláez *et al.* [4]) have traditionally been successful on semantic segmentation benchmark challenges such as PASCAL VOC [29], fully convolutional neural networks have recently achieved superior results [70]. Instead of predicting a k-dimensional distribution over k classes like the CNN described in Section 1.4.2, fully convolutional networks are trained end-to-end to directly make dense per-pixel class predictions. It might be interesting to investigate how well hands and hand poses could be segmented without the intermediate detection step.

Finally, there is a compelling future direction for the idea of inferring interactions from hands as introduced in Chapter 5. As demonstrated in Section 5.3.2, the prediction accuracy can improve drastically when integrating frames across time. However, our approach does not combine evidence until after propagating each frame through the CNN separately. Recently proposed network architectures [52, 100] aim to jointly process stacks of frames, sometimes in combination with optical flow [100]. Such networks could explicitly take advantage of motion information or specific temporal sequences of hand gestures to further improve hand pose-based inference from egocentric cameras.

6.3.2 BEYOND HANDS

The work presented in this dissertation underlined the potential of computer vision as an effective tool in the scientific analysis of visual data collected with head-mounted cameras. While we focused on hands as one particularly interesting subject to study in this context, there are arguably many more ways in which the combination of wearable cameras systems and powerful vision algorithms can yield novel insights into the visual systems of humans (or even other animals [48]).

We are particularly interested in extending our work to further investigate the development of visual attention and visual object learning in infants. One paradigm that we are actively studying is to utilize deep neural networks as human proxies to evaluate the learnability of real-world visual input captured from wearable cameras [7]. For example, in a scenario where infants and parents jointly play with a set of toys, CNNs trained on first-person data (from either parents or infants) for the task of toy object recognition could yield interesting differences in the way the two groups visually explore and perceive these objects. Adding eye gaze tracking to simulate the effects of central and peripheral vision in this learning context is an interesting future direction.

Lastly, recent advancements in deep learning combine computer vision with another research area: natural language processing (NLP). For example, researchers have proposed networks that process an image and generate a sentence describing the image (e.g. "A little girl is eating piece of cake.") as an output [51]. This process in generally referred to as image captioning. A related topic is that of visual question answering [3]. Here, the network receives two inputs, an image and a question regarding the image (e.g. "How many horses are in this image?"), and then aims to produce the correct answer. While these systems currently rely on labeled large-scale photo datasets, evaluating pre-trained networks with real-world, first-person data could potentially open up exciting new directions

to jointly study visual learning and language learning. Again considering the scenario of infant-parent toy play, one could image a system that tries to learn the mapping between a novel toy object and its name by interpreting both the infant's visual input as well as the language description of the parent.

Overall, we believe that computer vision methods have matured enough in recent years to provide reliable annotations for many interesting applications related to wearable camera systems. Moreover, as outlined in this section, deep learning-based computer vision systems have the potential to be directly used as models for visual learning, which (particularly in the context of first-person vision) could provide new insights into how we see the world around us.

BIBLIOGRAPHY

[1] Omid Aghazadeh, Josephine Sullivan, and Stefan Carlsson. Novelty detection from an ego-centric perspective. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3297–3304, 2011.

[2] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2189–2202, 2012.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, December 2015.

[4] Pablo Arbeláez, Bharath Hariharan, Chunhui Gu, Saurabh Gupta, Lubomir Bourdev, and Jitendra Malik. Semantic segmentation using regions and parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3378–3385, 2012.

[5] Vassilis Athitsos and Stan Sclaroff. Estimating 3d hand pose from a cluttered image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–432, 2003.

[6] Dana H Ballard, Mary M Hayhoe, Polly K Pook, and Rajesh PN Rao. Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20(04):723–742, 1997.

[7] Sven Bambach, David J Crandall, Linda B Smith, and Chen Yu. Active viewing in toddlers facilitates visual object learning: an egocentric vision approach. In *Proceedings of the 38th annual meeting of the Cognitive Science Society*, 2016.

[8] Sven Bambach, Linda B Smith, David J Crandall, and Chen Yu. Objects in the center: how the infant's body constrains infant scenes. In *6th Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*. IEEE, 2016.

[9] Paul J Besl and Neil D McKay. A method for registration of 3d shapes. *IEEE Trans on PAMI*, 14(2):239–256, 1992.

[10] Alejandro Betancourt, Miriam M. Lopez, Carlo S. Regazzoni, and Matthias Rauterberg. A sequential classifier for hand detection in the framework of egocentric vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2014.

[11] Alejandro Betancourt, Pietro Morerio, Carlo S Regazzoni, and Matthias Rauterberg. The evolution of first person vision methods: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(5):744–760, 2015.

[12] Vinay Bettadapura, Irfan Essa, and Caroline Pantofaru. Egocentric field-of-view localization using first-person point-of-view devices. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 626–633, 2015.

[13] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(11):1222–1239, 2001.

[14] James R Brockmole, Christopher C Davoli, Richard A Abrams, and Jessica K Witt. The world within reach: Effects of hand posture and tool use on visual cognition. *Current Directions in Psychological Science*, 22(1):38–44, 2013.

[15] George Casella and Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[16] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 415–423, 2015.

[17] Brian Clarkson, Alex Pentland, and Kenji Mase. Recognizing user context via wearable sensors. In *IEEE International Symposium on Wearable Computers*, page 69, 2000.

[18] Laura J Claxton, Dawn K Melzer, Joong Hyun Ryu, and Jeffrey M Haddad. The control of posture in newly standing infants is task dependent. *Journal of experimental child psychology*, 113(1):159–165, 2012.

[19] Laura J Claxton, Jennifer M Strasser, Elise J Leung, Joong Hyun Ryu, and Kathleen M O'Brien. Sitting infants alter the magnitude and structure of postural sway when performing a manual goal-directed task. *Developmental psychobiology*, 56(6):1416–1422, 2014.

[20] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[21] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[22] Joshua D Cosman and Shaun P Vecera. Attention affects visual perceptual processing near the hand. *Psychological Science*, 2010.

[23] David Crandall, Pedro Felzenszwalb, and Daniel Huttenlocher. Spatial priors for part-based recognition using statistical models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 10–17, 2005.

[24] Christopher C Davoli and James R Brockmole. The hands shield attention from visual interference. *Attention, Perception, & Psychophysics*, 74(7):1386–1390, 2012.

[25] 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)*, pages 248–255, 2009.

[26] Aiden R Doherty, Daragh Byrne, Alan F Smeaton, Gareth JF Jones, and Mark Hughes. Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs. In *Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*, pages 259–268. ACM, 2008.

[27] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[28] Ali Erol, George Bebis, Mircea Nicolescu, Richard D Boyle, and Xander Twombly. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding*, 108(1):52–73, 2007.

[29] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[30] Alireza Fathi, Ali Farhadi, and James M Rehg. Understanding egocentric activities. In *IEEE International Conference on Computer Vision (ICCV)*, pages 407–414, 2011.

[31] Alireza Fathi, Jessica K Hodgins, and James M Rehg. Social interactions: A first-person perspective. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1226–1233, 2012.

[32] Alireza Fathi and James Rehg. Modeling actions through state changes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2579–2586, 2013.

[33] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3281–3288. IEEE, 2011.

[34] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[35] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient belief propagation for early vision. *International journal of computer vision*, 70(1):41–54, 2006.

[36] Robert Fergus, Pietro Perona, and Andrew Zisserman. Object class recognition by unsupervised scale-invariant learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–264, 2003.

[37] Ross Ed Flom, Kang Ed Lee, and Darwin Ed Muir. *Gaze-following: Its development and significance.* Lawrence Erlbaum Associates Publishers, 2007.

[38] Tom Foulsham, Esther Walker, and Alan Kingstone. The where, what and when of gaze allocation in the lab and the natural environment. *Vision research*, 51(17):1920–1931, 2011.

[39] John M Franchak, Kari S Kretch, Kasey C Soska, and Karen E Adolph. Head-mounted eye tracking: A new method to describe infant looking. *Child development*, 82(6):1738–1750, 2011.

[40] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[41] Stephanie C Goodhew, Davood G Gozli, Susanne Ferber, and Jay Pratt. Reduced temporal fusion in near-hand space. *Psychological Science*, page 0956797612463402, 2013.

[42] Davood G Gozli, Greg L West, and Jay Pratt. Hand position alters vision by biasing processing through different visual pathways. *Cognition*, 124(2):244–250, 2012.

[43] Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4):188–194, 2005.

[44] Steve Hodges, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur, and Ken Wood. Sensecam: A retrospective memory aid. In *UbiComp 2006: Ubiquitous Computing*, pages 177–193. Springer, 2006.

[45] Berthold K Horn and Brian G Schunck. Determining optical flow. In *1981 Technical symposium east*, pages 319–331. International Society for Optics and Photonics, 1981.

[46] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

[47] A. Ihler. Kernel Density Estimation (KDE) Toolbox for Matlab. `http://www.ics.uci.edu/~ihler/code/kde.html`.

[48] Yumi Iwashita, Asamichi Takamine, Ryo Kurazume, and MS Ryoo. First-person animal activity recognition from egocentric videos. In *22nd International Conference on Pattern Recognition (ICPR)*, pages 4310–4315. IEEE, 2014.

[49] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[50] Michael Jones and Paul Viola. Fast multi-view face detection. *Mitsubishi Electric Research Lab TR-20003-96*, 3:14, 2003.

[51] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[52] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1725–1732, 2014.

[53] Cem Keskin, Furkan Kıraç, Yunus Emre Kara, and Lale Akarun. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In *Computer Vision–ECCV 2012*, pages 852–863. Springer, 2012.

[54] Kris M Kitani, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3241–3248, 2011.

[55] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[56] Vladimir Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1568–1583, 2006.

[57] Mathias Kölsch and Matthew Turk. Robust hand detection. In *IEEE International Conference on Automatic Face and Gesture Recognition (FGR'04)*, pages 614–619, 2004.

[58] Johannes Kopf, Michael F Cohen, and Richard Szeliski. First-person hyper-lapse videos. *ACM Transactions on Graphics (TOG)*, 33(4):78, 2014.

[59] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[60] Jayant Kumar, Qun Li, Survi Kyal, Edgar A. Bernal, and Raja Bala. On-the-fly hand detection training with application in egocentric action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2015.

[61] Michael F Land and Mary Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision research*, 41(25):3559–3565, 2001.

[62] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[63] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1346–1353, 2012.

[64] Cheng Li and Kris Kitani. Model recommendation with virtual probes for egocentric hand detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2624–2631, 2013.

[65] Cheng Li and Kris Kitani. Pixel-level hand detection in ego-centric videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3570–3577, 2013.

[66] Yin Li, Alireza Fathi, and James Rehg. Learning to predict gaze in egocentric video. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3216–3223, 2013.

[67] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 287–295, 2015.

[68] Yizhou Lin, Gang Hua, and Philippos Mordohai. Egocentric object recognition leveraging the 3d shape of the grasping hand. In *Computer Vision-ECCV 2014 Workshops*, pages 746–762. Springer, 2014.

[69] Sally A Linkenauger, Veronica Ramenzoni, and Dennis R Proffitt. Illusory shrinkage and growth body-based rescaling affects the perception of size. *Psychological Science*, 21(9):1318–1325, 2010.

[70] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[71] David G Lowe. Object recognition from local scale-invariant features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1150–1157, 1999.

[72] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2714–2721, 2013.

[73] Walterio W Mayol and David W Murray. Wearable hand activity recognition for event summarization. In *IEEE International Symposium on Wearable Computers*, pages 122–129, 2005.

[74] Oliver J Muensterer, Martin Lacher, Christoph Zoeller, Matthew Bronstein, and Joachim Kübler. Google glass in pediatric surgery: an exploratory study. *International Journal of Surgery*, 12(4):281–289, 2014.

[75] Peter Mundy and Lisa Newell. Attention, joint attention, and social cognition. *Current directions in psychological science*, 16(5):269–274, 2007.

[76] John R Napier. The prehensile movements of the human hand. *Bone & Joint Journal*, 38(4):902–913, 1956.

[77] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Training a feedback loop for hand pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[78] Nuria M Oliver, Barbara Rosario, and Alex P Pentland. A bayesian computer vision system for modeling human interactions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):831–843, 2000.

[79] Eric L Olofson and Dare Baldwin. Infants recognize similar goals across dissimilar actions involving object manipulation. *Cognition*, 118(2):258–264, 2011.

[80] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2847–2854, 2012.

[81] Yair Poleg, Chetan Arora, and Shmuel Peleg. Temporal segmentation of egocentric videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[82] Yair Poleg, Tavi Halperin, Chetan Arora, and Shmuel Peleg. Egosampling: Fast-forward and stereo for egocentric videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[83] Michael I Posner. Orienting of attention. *Quarterly journal of experimental psychology*, 32(1):3–25, 1980.

[84] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[85] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[86] Catherine L Reed, Ryan Betz, John P Garza, and Ralph J Roberts. Grab it! biased attention in functional hand and tool space. *Attention, Perception, & Psychophysics*, 72(1):236–245, 2010.

[87] Catherine L Reed, Jefferson D Grubb, and Cleophus Steele. Hands up: attentional prioritization of space near the hand. *Journal of Experimental Psychology: Human Perception and Performance*, 32(1):166, 2006.

[88] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[89] Xiaofeng Ren and Chunhui Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 3137–3144, 2010.

[90] Xiaofeng Ren and Matthai Philipose. Egocentric recognition of handled objects: Benchmark and analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1–8. IEEE, 2009.

[91] Douglas Reynolds. Gaussian mixture models. *Encyclopedia of Biometrics*, pages 827–832, 2015.

[92] Grégory Rogez, Maryam Khademi, JS Supančič III, Jose Maria Martinez Montiel, and Deva Ramanan. 3d hand pose detection in egocentric RGB-D images. In *Computer Vision-ECCV 2014 Workshops*, pages 356–371. Springer, 2014.

[93] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.

[94] Holly Alliger Ruff and Mary Klevjord Rothbart. *Attention in early development: Themes and variations*. Oxford University Press, 2001.

[95] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, DTIC Document, 1985.

[96] Michael Ryoo and Larry Matthies. First-person activity recognition: What are they doing to me? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2730–2737, 2013.

[97] Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, et al. Accurate, robust, and flexible real-time hand tracking. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3633–3642. ACM, 2015.

[98] Martin Shepherd, John M Findlay, and Robert J Hockey. The relationship between eye movements and spatial attention. *The Quarterly Journal of Experimental Psychology*, 38(3):475–491, 1986.

[99] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[100] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[101] Linda B Smith, Chen Yu, and Alfredo F Pereira. Not your mother's view: The dynamics of toddler visual experience. *Developmental science*, 14(1):9–17, 2011.

[102] Ekaterina H Spriggs, Fernando De La Torre, and Martial Hebert. Temporal segmentation and activity classification from first-person sensing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 17–24, 2009.

[103] Thad Starner, Bernt Schiele, and Alex Pentland. Visual contextual awareness in wearable computing. In *IEEE International Symposium on Wearable Computers*, pages 50–57, 1998.

[104] Bjoern Stenger, Paulo RS Mendonça, and Roberto Cipolla. Model-based 3d tracking of an articulated hand. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–310, 2001.

[105] Hans Strasburger, Ingo Rentschler, and Martin Jüttner. Peripheral vision and pattern recognition: A review. *Journal of vision*, 11(5):13–13, 2011.

[106] Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2439, 2010.

[107] James S Supancic, Grégory Rogez, Yi Yang, Jamie Shotton, and Deva Ramanan. Depth-based hand pose estimation: data, methods, and challenges. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1868–1876, 2015.

[108] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[109] Judit Takacs, Courtney L Pollock, Jerrad R Guenther, Mohammadreza Bahar, Christopher Napier, and Michael A Hunt. Validation of the Fitbit One activity monitor device during treadmill walking. *Journal of Science and Medicine in Sport*, 17(5):496–500, 2014.

[110] Robert Templeman, Mohammed Korayem, David J Crandall, and Apu Kapadia. Placeavoider: Steering first-person cameras away from sensitive spaces. In *The Network and Distributed System Security Symposium*, 2014.

[111] Esther Thelen, JA Scott Kelso, and Alan Fogel. Self-organizing systems and infant motor development. *Developmental Review*, 7(1):39–65, 1987.

[112] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[113] Shimon Ullman, Daniel Harari, and Nimrod Dorfman. From simple innate biases to complex visual concepts. *Proceedings of the National Academy of Sciences*, 109(44):18215–18220, 2012.

[114] Peter M Vishton, Nicolette J Stephens, Lauren A Nelson, Sarah E Morra, Kaitlin L Brunick, and Jennifer A Stevens. Planning to reach for an object changes how the reacher perceives it. *Psychological Science*, 18(8):713–719, 2007.

[115] Shenlong Wang, Sanja Fidler, and Raquel Urtasun. Lost shopping! monocular localization in large indoor spaces. In *IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[116] Daryl J Wile, Ranjit Ranawaya, and Zelma HT Kiss. Smart watch accelerometry for analysis and diagnosis of tremor. *Journal of neuroscience methods*, 230:1–4, 2014.

[117] Ying Wu and Thomas S Huang. View-independent recognition of hand postures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 88–94, 2000.

[118] Ying Wu, John Y Lin, and Thomas S Huang. Capturing natural hand articulation. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 426–432, 2001.

[119] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[120] Ryo Yonetani, Kris M Kitani, and Yoichi Sato. Ego-surfing first person videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5445–5454, 2015.

[121] Chen Yu and Linda B Smith. What you learn is what you see: using eye movements to study infant cross-situational word learning. *Developmental Science*, 14(2):165–180, 2011.

[122] Chen Yu and Linda B Smith. Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PloS one*, 8(11):e79659, 2013.

[123] Chen Yu and Linda B Smith. Multiple sensory-motor pathways lead to coordinated visual attention. *Cognitive Science*, 2016.

[124] Chen Yu, Linda B Smith, Hongwei Shen, Alfredo F Pereira, and Thomas Smith. Active information selection: Visual attention through the hands. *IEEE Transactions on Autonomous Mental Development*, 1(2):141–151, 2009.

CURRICULUM VITAE

Sven Bambach
School of Informatics and Computing, Cognitive Science Program
Indiana University Bloomington
Contact: sbambach@indiana.edu, homes.soic.indiana.edu/sbambach

Education

Indiana University, Bloomington, IN

- Ph.D., Computer Science, Advisor: David Crandall — August 2016
- Ph.D., Cognitive Science (joint degree), Advisor: Chen Yu — August 2016
- M.S., Computer Science, GPA 4.0 — May 2013

Cologne University of Applied Sciences, Cologne, Germany

- B.Eng., Media and Imaging Technology — November 2010
  - Thesis: Design and Realization of an Experimental Optical Stop-Motion Capture System, Advisors: Stefan Grünvogel, Dietmar Kunz

Peer-reviewed Publications

1. **Sven Bambach**, David Crandall, Linda Smith, Chen Yu. Active viewing in toddlers facilitates visual object learning: An egocentric vision approach. *Annual Conference of the Cognitive Science Society*, 2016.
2. **Sven Bambach**, Linda Smith, David Crandall, Chen Yu. Objects in the center: How the infant's body constrains infant scenes. *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, 2016.
3. **Sven Bambach**, Stefan Lee, David Crandall, Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. *IEEE International Conference on Computer Vision (ICCV)*, 2015.

4. **Sven Bambach**, David Crandall, Chen Yu. Viewpoint integration for hand-based recognition of social interactions from a first-person view. *17th ACM International Conference on Multimodal Interaction (ICMI)*, 2015.

5. Stefan Lee, **Sven Bambach**, David Crandall, John Franchak, Chen Yu. This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Workshop on Egocentric Vision*, 2014.

6. **Sven Bambach**, John Franchak, David Crandall, and Chen Yu. Detecting hands in children's egocentric views to understand embodied attention during social interaction. *Annual Conference of the Cognitive Science Society*, 2014.

7. **Sven Bambach**, David Crandall, Chen Yu. Understanding embodied visual attention in child-parent interaction. *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, 2013.

Awards

Fellowships

- Paul Purdom Fellowship Award for Doctoral Studies in CS/Informatics 2015/2016

Research Awards

- 1st place in hand detection/classification at the *Vision for Intelligent Vehicles and Applications (VIVA) challenge (IEEE IV2015)* June 2015
- *Intel* best paper award at the EgoVision Workshop *(IEEE CVPR)* June 2014

Travel Awards and Grants

- NSF-sponsored 2016 travel award for young scientists to attend CogSci 2016
- Purdue University C Design Lab and NSF-sponsored travel award to attend CVPR 2016
- IU SOIC Ph.D. student travel grant to attend CVPR 2013 and ICDL 2013

Undergraduate Awards, Cologne University of Applied Sciences

- Best GPA among all 2010 graduates at the IMP institute November 2010