INTERACTIVE TRACKING AND ACTION RETRIEVAL TO SUPPORT HUMAN BEHAVIOR ANALYSIS

A Dissertation
Presented to
The Academic Faculty

by

Arridhana Ciptadi

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Interactive Computing

Georgia Institute of Technology
May 2016

Copyright © 2016 by Arridhana Ciptadi

INTERACTIVE TRACKING AND ACTION RETRIEVAL TO SUPPORT HUMAN BEHAVIOR ANALYSIS

Approved by:

Professor James M. Rehg,
Committee Chair
School of Interactive Computing
Georgia Institute of Technology

Professor James M. Rehg, Advisor
School of Interactive Computing
Georgia Institute of Technology

Professor Gregory D. Abowd,
co-Advisor
School of Interactive Computing
Georgia Institute of Technology

Dr. Agata Rozga
School of Interactive Computing
Georgia Institute of Technology

Professor Daniel Messinger
Department of Psychology
University of Miami

Professor Pietro Perona
Division of Engineering and Applied Science
California Institute of Technology

Date Approved: March 14, 2016

ACKNOWLEDGEMENTS

I joined the Ph.D. program not fully realizing how much effort one needs to spend in order to obtain this degree. This is certainly the most difficult thing that I have ever done in my life and it is only through the support of family and friends that I am able to complete this journey. I would like to thank my mom and dad, Purwani Eko Prihatin and Agus Sardjono, for always supporting me. Ever since I can remember, both of you always trust my judgment (maybe too much) and always believe in me. Only recently I realize how lucky I am to have been raised in such an environment. Next, I would like to thank my sister, Anindya Chandra Dewi, for always being there for me. I know no matter what happened I can always talk to you to cheer me up and make me feel better. I also would like to thank my grandpa and grandma for always showering me with love. I am incredibly fortunate to have a second set of parents like you two.

Next, I would like to thank all the people in Wall Lab. It is hard to imagine a better group of friends to share a lab with. All of you are incredibly fun to be around with. Thank you to Ahmad Humayun, Chanho Kim, Abhijit Kundu, Yin Li, Brian Goldfain and Paul Drews for being such fun lab mates and friends. I will always remember the many interesting discussions and fun conversations we had. I also would like to thank my roommate, Khaldoon Abu-Hakmeh. I didn’t think our serendipitous meeting would turn into this lasting friendship. I am really lucky to have a friend like you that I can always rely on. Next, I would like to thank Bruce Chou for always being a good friend. You are one of the nicest people that I have ever met. I also would like to thank Vanessa, Grace, Jake, Victor, Jenny, Kristin, Yi. You are all awesome. Your companionship means a lot to me. It is impossible to list all my friends who have made this journey to be so fun and memorable. I have shared many interesting and personal moments with so many of you. I am really thankful to have the support and friendship from all of you.

Last, I would like to thank all my committee members. It would be impossible for me to finish this journey without all your help and support. I would like to thank Agata Rozga for always being supportive and keeping me focused on the problem domain. I'm really fortunate to have met you at the early stage of my Ph.D. career. I would like to thank Daniel Messinger for introducing me to an interesting problem which becomes one of the central themes of this dissertation. I would like to thank Pietro Perona for always reminding me to be thorough in evaluating my methods. Last, I would like to thank both my beloved advisors, James Rehg and Gregory Abowd. Gregory, I admire you for always being organized, sharp, fair, critical and compassionate to your students. I know I can always rely on you to help me sharpen and focus my work. I still remember clearly my early interactions with you during Expeditions research meetings. Your strong focus on meaningful applications of technical research is really inspiring. Jim, you always push me to be precise and aware of how the different concepts and ideas are connected. I always learn something new after talking to you. I chose Georgia Tech in large part because I felt your enthusiasm over the phone when you were describing the behavior imaging research project that you were starting up at that time. Looking back, I definitely think I made the right decision joining you in this effort. Jim and Gregory, I cannot express how grateful I am to have both of you as my advisors. I consider both of you as good friends and mentors. Both of you certainly have made a positive and lasting impact on me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The goal of this thesis is to develop a set of tools for continuous tracking of behavioral phenomena in videos to support human behavior study. Current standard practices for extracting useful behavioral information from a video are typically difficult to replicate and require a lot of human time. For example, extensive training is typically required for a human coder to reliably code a particular behavior/interaction. Also, manual coding typically takes a lot more time than the actual length of the video (e.g., it can take up to 6 times the actual length of the video to do human-assisted single object tracking [151][1]). The time intensive nature of this process (due to the need to train expert and manual coding) puts a strong burden on the research process. In fact, it is not uncommon for an institution that heavily uses videos for behavioral research to have a massive backlog of unprocessed video data.

To address this issue, I have developed an efficient behavior retrieval and interactive tracking system. These tools allow behavioral researchers/clinicians to more easily extract relevant behavioral information, and more objectively analyze behavioral data from videos. I have demonstrated that my behavior retrieval system achieves state-of-the-art performance for retrieving stereotypical behaviors of individuals with autism in a real-world video data captured in a classroom setting. I have also demonstrated that my interactive tracking system is able to produce high-precision tracking results with less human effort compared to the state-of-the-art. I further show that by leveraging the tracking results, we can extract an objective measure based on proximity between people that is useful for analyzing certain social interactions. I validated this new measure by showing that we can

[1] This is calculated using an optimistic assumption that it takes a human one second to annotate one object in a single frame.

use it to predict qualitative expert ratings in the Strange Situation (a procedure for studying infant attachment security), a quantity that is difficult to obtain due to the difficulty in training the human expert.

CHAPTER I

INTRODUCTION AND MOTIVATION

1.1 Objective

The goal of this thesis is to demonstrate how computer vision applied to continuous tracking of behavioral phenomena in videos can assist in the study of human behavior. Video is an essential component in many studies in developmental psychology. It has been used in a wide range of applications, such as assisting the coding of child behaviors in a controlled research protocol (e.g., in The Strange Situation [9] and ADOS [95]) and functional assessment of a target behavior (e.g., stereotypies in individuals with autism [100]). Currently, the ability of psychologists and clinicians to utilize these video recordings is hampered by the lack of tools for extracting useful behavioral measures from them. I wish to alleviate this problem by creating a set of tools that allow researchers and clinicians to more easily extract useful behavioral measures from videos.

1.2 Thesis Statement

Motion-derived action representation and instance-specific tracking model enables faster extraction of behavioral measures from video data compared to manual annotation and existing computer vision techniques.

1.3 Motivation

Imagine a situation where a child with a history of stereotyped/self-injurious behavior is studying in a classroom that is equipped with cameras (see Figure 1). During the class session, the child is following the lesson quietly for about 10 minutes, but suddenly he ran away from the table. A clinician then assisted the child back into his seat but from that point on, the child seemed unsettled by exhibiting several bouts of self-injurious behavior and escaping his seat a few more times until the classroom session ended. This summary is exactly the type of information that a behavioral therapist is looking for. To obtain this information, the therapist will have to review the video recording of the class. However, a therapist is interested not only in what has happened, but also *why* the child was exhibiting certain disruptive behaviors so that the appropriate program can be created to help him.

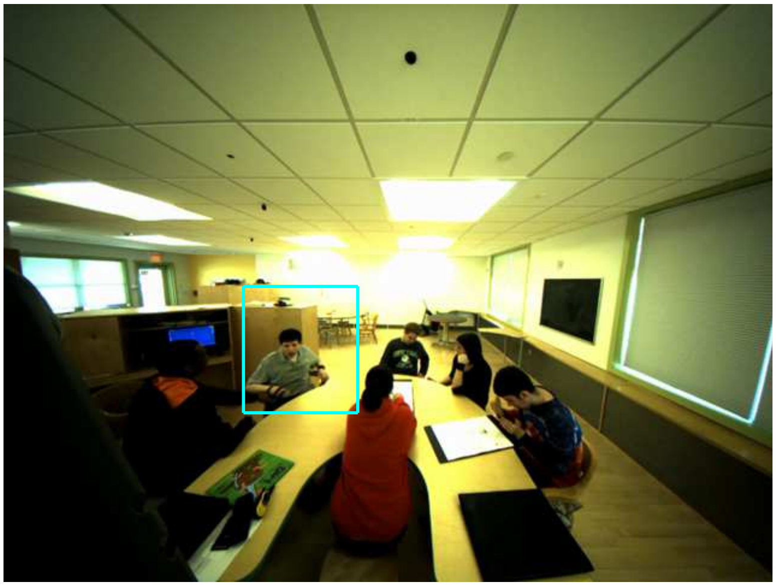


Figure 1: A child with special needs studying in a class.

Figuring out the underlying cause of a behavior is often a challenging task since there are many potential factors. For example, the child ran away might be because he is trying to get attention from somebody else in the room by actively moving to gain proximity with him, or it might be because the child is simply trying to escape from the lesson. For the self-injurious behavior, it is possible that the child did it as an attempt to get access to certain items (e.g. toys), or he did it because it fulfills a certain sensory function for him, or the child might feel that the room was unpleasant since the sunlight made it particularly bright that day. To find out exactly which ones of these factors contributed to the child's disruptive behavior, a therapist will have to review a lot more video recordings to find some patterns by looking at other instances of the target behavior. This process is incredibly time-consuming considering the amount of videos that the therapist potentially will have to review (e.g. the child might attend the class a few times per week). The process of having to watch the full video recording just to find some instances of a particular behavior is also very inefficient since the behavior itself might only be present for a few seconds in the video. A mechanism that allows a therapist to quickly browse through a large video collection to find instances of a particular behavior would greatly improve this process.

The first problem that I address is behavior retrieval in videos and its application for studying stereotyped and repetitive behaviors (stereotypies) in individuals with autism. In the domain of behavioral psychology, there is currently great interest in studying the effectiveness of behavioral therapy for children with an Autism Spectrum Disorder (ASD) [13]. These children frequently exhibit repetitive and stereotyped behaviors, known as *stereotypies*. In comparison to more traditional functional activities, stereotypies are often unique expressions of individual behavior and are therefore highly subject-specific, making it challenging to construct a general model of such behaviors. While clinicians use videos as a tool to capture these behaviors, currently the process of identifying these behaviors from a video collection is very time-consuming.

The problem of retrieving human actions is challenging due to the need to handle many sources of variations: viewpoint, size and appearance of actors, scene lighting and video quality. While many approaches have been proposed, getting accurate results remains a challenge. State-of-the-art features that are used in video action recognition/retrieval require a lot of training examples to perform well (e.g., dense trajectories [155]). This makes it challenging to apply current techniques to stereotypies since it is difficult to collect a lot of examples of such behaviors due to several factors. First, they are highly subject-specific. Although commonality certainly exists (e.g. Lam et.al. [86] found that 74.2% of individuals with autism exhibit some form of hand/finger movement stereotypies), the actual expression of the behavior itself differs from person to person. Second, the frequency of occurrence for some of the behaviors might be low. A useful behavior retrieval system should be able to work for a given new instance of target behavior. The need to collect a large number of examples, even for behaviors that have low frequency of occurrence, is certainly a problem. Last, the landscape of behaviors of interest for a particular individual is always constantly changing because of various environmental factors such as a change in diet, in medication, in sleeping pattern, etc.. Furthermore, while the act of behavior therapy itself might successfully reduce or eliminate a certain disruptive behavior, it might cause new problem behavior to be expressed if the program fails to address the underlying cause of the original disruptive behavior. Due to all these factors, an ideal behavior retrieval system should still be able to perform well even if only given a single example of the target behavior.

In this thesis, I address this problem of retrieving behavior of interest given only a single example. While this is certainly a difficult requirement to satisfy, I will later demonstrate that we can address it by exploiting the unique nature of the movements of a certain classes of stereotypies. The core of my approach is a novel feature for representing action based on the timing patterns of its movements. I demonstrate that my feature is robust to variations in viewpoint and appearance in a controlled cross-view action recognition experiment using the widely used IXMAS dataset [163]. To validate my proposed approach, I show that not only my novel feature outperform the state-of-the-art techniques for the task of retrieving stereotypies in real-world video recordings, but also I demonstrate that it is useful for the task of general action recognition, achieving state-of-the-art performance in cross-view action recognition and sports action recognition when combined with existing action representation.

The second problem that I address in this work is high precision tracking of people and objects and its application for objective analysis of interaction. More specifically, I show that by tracking people in a video we can derive an objective measure based on proximity that is useful for analyzing certain interactions. A lot of behavioral phenomena can be built up from having a continuous measurement of proximity between people or people and objects. This measure, or measures derived from it, are present in a wide range of behavioral studies such as temperament [174], attachment [9], locomotion [5], exploration [164] and personal space [137]. For example, central to a lot of the studies on personal space by Sommer is the measure of physical distance between people. Rothbart used a continuous measure of physical distance between an infant and an object as the key measure for studying infant approach behavior in the context of temperament study [125]. Similarly, a continuous measure of distance between an infant and a target goal is one of the key measures in studies of infant locomotion [5]. In The Strange Situation (a procedure to observe infant attachment to the caregiver), the proximity between the infant and the caregiver is central to how an expert rate the interaction. For example, some of the dimensions of the baby-caregiver interaction that an expert rate includes proximity-seeking (the intensity, promptness and persistence of the baby's efforts to gain contact with the caregiver), contact-maintenance (degree of activity and persistence in baby's efforts to maintain contact with the caregiver) and avoidance (intensity, persistence, duration and promptness of the baby's avoidance of proximity and interaction, even across a distance) [9].

Given the importance of this proximity measure, it is striking that standard practice of obtaining this measure still depends on human visual observation. For example, Adolph et.al. [5] instrumented the environment with physical markers (grids) so that a human observer can measure the location of the infant in a video recording during infant locomotion study. Similarly, Ainsworth and Bell in an early study of infant attachment [8] also used grids so that human observer can derive a coarse measure of distance from the video recording. Relying on human visual observation to estimate physical distance from video data is certainly not desirable since it is both inefficient (it takes time for humans to estimate distance between two objects even for a single frame) and inaccurate (e.g. humans have to rely on a certain physical marker with known size as the basis for measurement). It is clear that the ability to accurately measure distance between people or people and object will greatly impact many studies in psychology.

In this work, I demonstrate that combining high precision object tracking and RGBD camera allows for accurate extraction of temporally-dense proximity measure. To showcase the usefulness of this measure, I focus on the problem of analyzing the interaction between an infant and a caregiver in the context of infant attachment study in The Strange Situation.

The Strange Situation is a scripted interaction procedure, and is the most widely used method for studying infant attachment security to the caregiver. It is an interesting procedure to demonstrate the value of having a dense measure of proximity since currently, analysis on The Strange Situation entirely relies on expert judgment for the presence or absence of certain attachment behaviors (e.g. proximity-seeking, contact-maintenance, avoidance and resistance behaviors). However, from the theory of attachment, it is clear that the infant's attachment security will affect how he will move in the space in relations to the caregiver. Currently, a quantitative proximity measure (e.g. Fig. 2) is not being used by experts when producing attachment classification for The Strange Situation since it is difficult to obtain using human observation. I later demonstrate that not only that this quantitative measure can be used to predict the subjective ratings given by expert coder, but also having this measure allows us to easily test a new hypothesis about this procedure that is previously difficult to test.

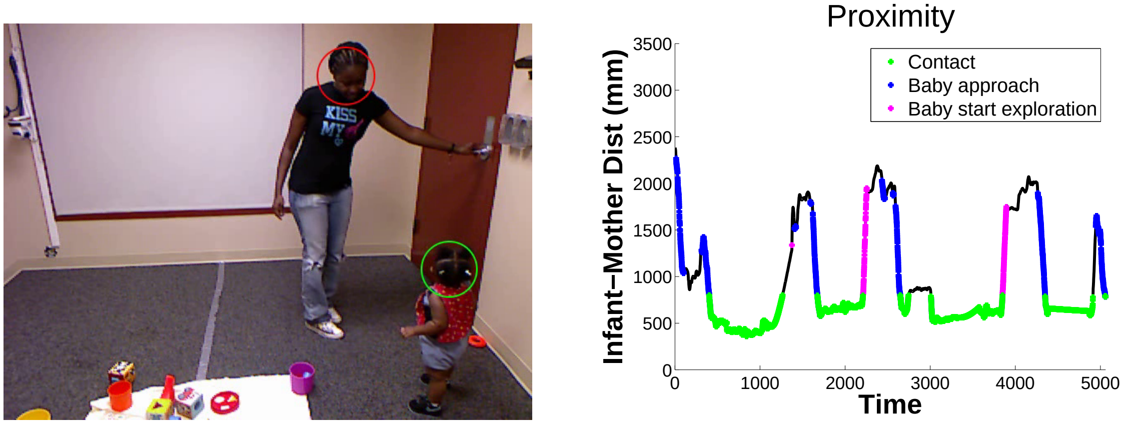


Figure 2: Infant-mother proximity derived from head tracking results. Left: tracked infant and mother. Right: infant-mother proximity during an episode in The Strange Situation procedure.

Deriving the measure of proximity from RGBD video requires accurate localization of the target objects. Tracking objects in a video with high precision is an extremely challenging problem due to several factors. First, the object might undergo a big amount of visual variations during the course of the video because of pose changes, deformations and lighting changes. Second, the object might become heavily occluded, which is another common cause for tracking failures. Last, the target object can get in and out of the video frame at any time. To this day, fully-automated high precision object tracking remains a big challenge in computer vision. In this thesis I address the problem of obtaining high precision object tracking efficiently in an *interactive* tracking framework, where high precision results is obtained by including human annotator as an oracle to correct for tracking mistakes.

Interactive object tracking techniques have been employed for applications that require accurate tracking results such as annotating video for training object detectors and preprocessing for movie special effects. However, the literature on interactive tracking is still virtually bereft of any application of these techniques to my target problem domain of studying human behavior. The main issue to be addressed in an interactive tracking framework is efficiency: how to minimize the usage of the human annotator since human time is the most expensive resource. While there have been studies on how to effectively scale-up an interactive tracking system (e.g. through crowdsourcing [151]), the issue of efficiency still has not been thoroughly addressed in the interactive tracking literature (e.g., twenty six hours of surveillance video cost tens of thousands of dollars to annotate despite using state-of-the-art annotation system [109]). In this thesis, I address the problem of minimizing human effort in interactive tracking by leveraging user annotations for incrementally learning instance specific model parameters within the tracking cost function. I will later demonstrate that using my proposed approach, accurate tracking results can be obtained more efficiently (requiring $\sim 60\%$ fewer annotations) compared to the state-of-the-art interactive tracking techniques.

1.4 Contributions

I have made the following four contributions:

1. A state-of-the-art action retrieval approach that is robust to variations in viewpoint and appearance by exploiting the dynamics of an action for temporally locating stereotyped/repetitive behaviors in video (Chapter 3). We have published our Stereotypy dataset to the research community to encourage future research on the behavior retrieval problem in realistic contexts. This work has been presented in International Meeting for Autism Research 2012 and published in European Conference on Computer Vision 2014.

2. A new computer vision interactive tracking method that obtains high-precision object track from video with less human effort compared to the state-of-the-art by leveraging human annotation to incrementally learn instance-specific model parameters of the tracking cost function (Chapter 4). This work has been published in International Conference on Computer Vision 2015.

3. An objective measure based on proximity to analyze interaction. This measure demonstrates that it is possible to depart from the traditional subjective human rating to analyze infant attachment behavior in The Strange Situation procedure (Chapter 5). This work has been presented in Society for Ambulatory Assessment Conference 2015, Association for Psychological Science Annual Convention 2015, and International Conference on Infant Studies 2016.

4. A demonstration of computer vision research for human behavior study in order to influence the practice on how clinicians and behavioral researchers perform analysis on video data.

CHAPTER II

RELATED WORK

In this chapter we review relevant related work. Section 2.1 gives an overview on action/activity representation in video and also action recognition and retrieval methods. Section 2.2 describes various object representation strategies in tracking. Following this, Section 2.3 discusses applications of tracking in behavioral study. Finally, Section 2.4 outlines work on interactive tracking.

2.1 Video Action Recognition and Retrieval

There is a vast literature on action/activity representation. A classic representation of action in videos is based on space-time templates [22, 52, 79]. While this approach captures the fine-grained detail of an action, it is challenging to achieve robustness to various sources of variations, particularly in viewpoints. Inspired by the success of sparse features in object recognition [96], local sparse space-time features combined with the bag-of-features (BOF) framework have dominated the landscape of action representation for the last decade. Some of the local features that have been explored in this domain includes interest points [88, 39], tracks of points [102, 73, 155] or frame based descriptors [146]. Given these locally defined spatial neighborhoods (e.g. points, track of points), one can extract various types of descriptors such as HOG/HOF [88, 39, 155], MBH [155], MIP [82] or shape-flow [146].

One of the major challenges in the action retrieval problem is the viewpoint effect: an action looks different depending on the camera viewpoint. While various descriptors are sufficiently specific to be discriminative and thus have been shown to be useful for action recognition, they are not robust to variations in viewpoints, and thus may not support accurate retrieval of actions across views. One way to handle this problem is by having a lot of training examples comprising of actions captured from different viewpoints. However, this requirement presents a significant practical barrier for many applications, including the one that I address in this work (see Section 3).

Recently, interesting work has been done to address the challenge of viewpoint variation in action recognition. Liu et.al. [94] tackle the viewpoint problem through transfer learning by building a mapping between codebooks from different viewpoints. However, their framework requires knowledge of the camera viewpoint associated with each action (in testing and training). In a similar spirit, Li [92] and Zhang [177] learn a series of linear transformations of the feature vector extracted from a video to make it invariant to viewpoint changes. However, a linear transformation is not guaranteed to accurately model view-invariant mapping. Also, performance of their method drops significantly in the absence of multi-view observations of actions in training examples. In addition, these methods depend on the shape-flow descriptor that requires extraction of a bounding box and silhouette of an action, which can be challenging in real-world videos. Note that these methods assume a discrete number of pre-defined camera positions, which limits applicability of the methods since the need to collect examples across viewpoints can be burdensome.

Junejo et.al. [73] propose the self similarity matrix (SSM) which exhibits invariance to viewpoint changes. They compute SSM by either point tracking or pairwise frame similarity. However, point tracking is not always accurate and computing pairwise frame similarity means the feature will not be robust to slowly changing background. Another representation robust to changes in viewpoints is the hankelet ([90]). Hankelet is a hankel matrix representation of a tracklet that is invariant to affine transformations. Results in [73] and [90] show that SSM and hankelet are susceptible to large viewpoint changes.

In contrast to the action recognition domain, relatively little work has been done on action retrieval. DeMenthon and Doermann [35] built a system for retrieving action based on a single example, but it only works for short and near duplicate actions. Jung et.al. [74] present an action retrieval system based on shape template matching of body parts. While their method obtains good results on the standard KTH dataset, estimating body parts in a more general scenario is likely to be challenging. Yu et.al. [170] present an approach for human action search by performing Hough voting using STIP features extracted from the video. In the image retrieval domain, Rubner et.al. [127] incorporated the Earth Mover's Distance (EMD) as a way to measure similarity between images by using a linear programming approach. An adaptation of this method has been recently proposed for learning common activity prototypes in a video [173].

An interesting line of work has been done on action retrieval without using even a single visual example. İkizler and Forsyth presented a language-based video action retrieval [64]. The idea is that assuming human limb tracking can be performed reliably, then searching for a particular action in a video collection can be done using only sentences as the query, without any visual examples. Wang et.al. [157] developed a system for automatically retrieving quasi-periodic events in video. This is a bottom up process to retrieve instances of repeating visual patterns in a video. The challenge in their work is defining the right unit for what can be considered as a single visual pattern. They demonstrated the utility of their method for retrieving instances of social games from a long video recording. In a similar spirit, Prabhakar et.al. [116] built a system for retrieving social games based on temporal causality analysis on the occurrences of visual words.

2.2 Object Representation in Tracking

In the tracking context, an object is defined as any physical entity that is of interest. For example, an object can be something rigid such as a car, a bicycle, a plane or something non-rigid such as a person or an animal. In the tracking literature, an object is represented by its shape and appearance. Here I review the various approaches for modeling object shape and appearances.

I first discuss the different shape modeling strategies in tracking. A classical way to capture shape is to represent an object as a collection of points with a certain spatial configuration [96, 131]. To this end, various interest point detection strategies have been explored (e.g. using eigen values of the structure tensor to detect Harris corner [59], the Shi-Tomasi method [135], or using the difference of Gaussian function [96]). The main drawbacks of this shape modeling approach is that sometimes it is difficult to reliably detect the same set of interest points from an object under various possible sources of variations (e.g. illumination, scale, viewpoint). Another approach to represent object shape is by using simple geometric primitives such as ellipse or rectangle. While this approach might be the simplest, it is by far the most popular [32, 76, 180, 68, 124, 91, 14, 15, 57, 175]. In large part, this popularity is caused by the rapid progress in object detection. Since most object detectors operate on simple geometric primitives (typically rectangles), a tracking approach that uses the same shape modeling strategy can build on the success of object detection. Another way to model object shape is by using silhouette or contour [66, 169]. While this shape representation has been shown to be capable of producing good results for tracking certain objects in a highly cluttered environment [66], modeling all possible contour variations of an object can be difficult to do. Another approach for representing shape is the articulated shape models [122, 16]. This approach is typically employed when the application demands further information beyond just knowing where the object is (e.g. when we need to know where the hand is, in addition to knowing where the person is).

I now discuss appearance modeling strategies in tracking. In general, we can classify all appearance modeling approaches into three groups: 1) templates [180, 76, 68, 124, 91, 14, 57]; 2) image statistics [32, 169, 16]; and 3) filter responses [15, 175, 83, 50].

Template-based approaches model the appearance of an object by either its raw image intensities [180, 68, 14, 57], a subspace representation based on image intensities [124, 91], or image gradients [66, 131]. An advantage of a template-based approach is that in addition to modeling the object appearance, it also encodes the spatial information. Another advantage is that a template tends to be very discriminative, which is beneficial for tracking since we often need to discriminate between the target object and similar looking objects in the background. The main drawback of template is that it is more susceptible to object appearance variations due to factors such as illumination and viewpoint.

Image statistics-based approaches describe the appearance of an object by aggregating statistics derived from image intensities. These statistics can be in the form of a simple variance of image intensities [76], a color histogram [131, 32] or histogram of oriented gradients (HOG) [97, 33]. Many recent work on object detection use a combination of image statistics features (e.g. the Aggregated Channel Features (ACF) [40] which combines gradient magnitude, HOG and color histogram). An advantage of image statistics-based appearance modeling is that it is compact, thus allows for fast comparisons (e.g. computing distance in the feature space is cheap) and also efficient to store in memory. For example, Hall's object tracker makes use of fast feature extraction (ACF) and classification pipeline (boosting) to achieve real-time speed [55]. Another advantage is that it is robust to various sources of appearance variations. On the other hand, depending on the choice of statistics, and appearance characteristics of the target object and the background, this approach might produce a representation that is not sufficiently specific to discriminate between the target and the background.

Filter responses-based approaches summarize object appearance by the filter responses of the raw image. The filters can be in the form of Haar-like features [15, 57, 175] or filters obtained from deep convolutional networks [83, 50]. An advantage of this approach is that it provides a compact representation that is discriminative and relatively robust to various sources of appearance variations. A potential drawback of this approach is that the performance of the system will largely depend on the choice of filters. However, recent work on deep convolutional neural networks [83, 50] demonstrate that this filter design problem can be alleviated by using filters obtained from deep convolutional networks trained on a large number of images.

2.3 Tracking for Behavior Analysis

The need to analyze a large number of behavioral data from video has created a big demand for computational tools that can be used to assist the process of extracting relevant behavioral information from video. Tracking in particular has been widely used in many studies of animal behaviors [34, 23, 140, 105, 75, 110, 113, 61]. To obtain high precision results, most tracking approaches in this domain are typically highly tuned to track for a particular type of animal in a very specific environmental setup. For example, Ohayon et.al. [110] developed a system for tracking mice by first manually tagging each mouse with a unique color pattern so that each target can be easily identified. They use the histogram of oriented gradients (HOG) [33] to represent appearance of a mouse, and tracking is done by using a variant of sampling approach. Perez et.al. [113] used exemplar-based approach for detecting and tracking individual objects (they experimented on mice, flies, fish and ants). A big downside of their approach is the need to collect a large number of exemplars per target (typically 3000) for their system to perform well.

A very popular framework for tracking animals is combining some variations of background subtraction (for detecting the target object) with some notion of motion modeling [105, 23, 34, 140]. For example, Branson et.al. [23] built a system for tracking flies, cockroaches or other elliptically shaped animals in a container. Their system assumes simple background (i.e. easily distinguishable from the target animal) so that detecting the target in each frame can be done by a simple background subtraction approach. Once detection has been performed, the task of data association (assigning a detection to a particular track) is done using a simple matching approach. Similarly, Straw et.al. [140] built a system for tracking flies and birds by using a background subtraction approach for object detection and extended Kalman filter (EKF) for associating detections with trajectories.

Tracking is important in many animal behavior studies since many behavioral analysis approaches use the tracking results as the basis for detecting certain events. For example, JAABA [75] used a suite of trajectory-related features such as speed, distance-to-wall, inter-animal proximity to discriminate between certain classes behaviors in flies. Similarly, Hong et.al. [61] employed a combination of image features with trajectory for recognizing mouse social behavior. It is then crucial that the tracking results contain as few errors as possible. Currently, even with the highly tuned approach described above, errors in tracking (typically ID switches) still happened occasionally. To address this issue, typically a post-processing stage is performed so that a human can interactively fix for tracking errors (e.g., CTRAX, a popular fly tracker, uses FixErrors Matlab GUI [1] to allow users to identify and fix tracking errors).

2.4 Interactive Tracking

The literature on interactive tracking is relatively sparse compared to the extensive literature on fully-automated object tracking. Early work in interactive tracking focused on creating a system that can quickly incorporate a new annotation given by the user during the interactive stage to refine the tracking result [24, 161]. The goal was to enable the user to quickly evaluate the quality of the tracking result and decide whether additional annotation is necessary. To achieve this, Buchanan and Fitzgibbon [24] combined an efficient data structure based on K-D tree and a dynamic programming approach for interactive feature tracking. The K-D tree allows for fast lookup of patches with similar appearance, while dynamic programming provides an efficient solution for inferring the trajectory of the tracking target. Wei et.al. [161] used the dynamic programming approach proposed by Buchanan and Fitzgibbon and combined it with object detection to build an interactive object tracking system. The basic idea is that given some initial annotations, an interactive tracking system should be able to anticipate likely object locations in a given frame by performing object detection (with a conservative threshold). This allows the system to more quickly respond to the user's input during the interactive stage to perform object trajectory optimization.

Another line of work in interactive tracking focuses on interpolation strategies. Wei and Chai [160] propose a weighted template model (based on color histogram) for interpolating object appearance. The idea is that the appearance of the target object in all frames can be adequately described by a linear combination of the appearance of the object in the annotated frames. The LabelMe video work by Yuen et.al. [172] presents a strategy for interpolating the location of the target object in between keyframes by using homography-preserving linear interpolation. Using linear interpolation to infer an object trajectory is an efficient alternative to the dynamic programming approach presented in [24, 161], but it assumes that annotations are performed densely such that the object moves linearly between the annotated frames. To achieve good tracking results by using linear interpolation, Vondrick et.al. [151] estimated that on average 1 out of every 5 frames would need to be annotated.

A further line of work in interactive tracking focuses on frame selection strategies to minimize the number of annotations that a user will need to perform to obtain good tracking results. Vondrick and Ramanan [152] propose an active learning framework for interactive tracking. They present an approach for deciding which frame to present based on the expected change in the tracking result if the user were to annotate that frame (similar to the popular maximum expected gradient length (EGL) algorithm for active learning [132]). In the video segmentation domain, Fathi et.al. [46] present an active learning approach based on using frame uncertainty to decide which frame to annotate. Their approach is based on the assumption that the frame with the highest uncertainty estimate is the one that will be the most informative for segmentation purposes. Vijayanarasimhan and Grauman [149] present a frame selection method for video segmentation based on expected label propagation error. In contrast to these works, the focus of my work (Chapter 4) is not on the selection of the best frame for the user to annotate. Rather, my goal is to utilize the annotation information more effectively for the task of interactive tracking.

CHAPTER III

RETRIEVING ACTIONS FROM ARBITRARY VIEWPOINTS

3.1 Introduction

Given a single example video clip containing an action of interest, the goal is to retrieve all matching instances of that action from an unstructured video collection. I refer to this problem as *action retrieval*. I am motivated by applications in behavioral and developmental psychology, where it is common practice to record video of children and adults engaged in a variety of activities. Currently, the ability of psychologists to utilize these recordings is hampered by the lack of tools to efficiently search the video collection for a specific target behavior.

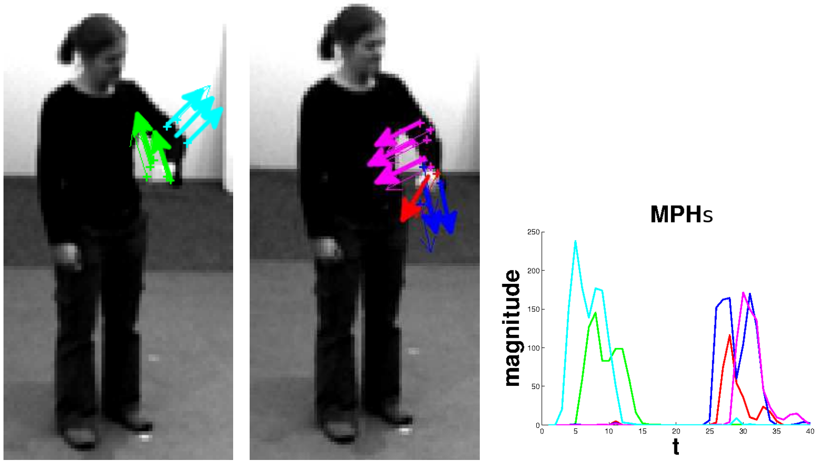
As an example, a facility that serves individual with developmental disabilities (e.g. The Center for Discovery in NY) may want to record the activities during a class session by instrumenting a room with one or multiple cameras to allow clinicians to do a more comprehensive review of what is happening during the session. From the clinician's perspective, having a visual record of the activities is immensely useful. One usage of such recordings is to assist in discovering new potential problem behaviors that might be missed by the staffs who were present in the room. For this use case, after identifying this new behavior of interest, a clinician then can review the current and previous recordings to help her answer several questions: 1) Has this individual exhibited this behavior previously? Knowing whether this behavior is a new occurrence or something that has happened before in the past is useful for understanding the cause of the behavior. 2) What is the frequency the behavior? Knowing how many times an individual exhibited a problem behavior during a particular session is a useful measure for assessing factors that might affect the behavior (e.g. if the frequency is much higher during the after-lunch sessions, the individual might be having issues with the foods). 3) What is typically happening just before the behavior occurred? Figuring out patterns around the behavior is useful for determining the potential trigger. Having a system that can be used to quickly retrieve instances of a particular problem behavior from a large video collection will be extremely useful for the clinicians working in this domain.

A big challenge in this behavior retrieval task is that the behavior or interest for a particular individual changes over time because of various environmental factors such as change in diet, medication or sleeping pattern. Furthermore, while the act of behavior therapy itself might successfully reduce or eliminate a certain disruptive behavior, it might cause new problem behavior to be expressed if the program fails to address the underlying cause of the original disruptive behavior. All these factors make it challenging for applying the classical action recognition approach to this problem since techniques in this domain often requires the collection of numerous training examples, which is impractical in this scenario. An ideal behavior retrieval system should still be able to perform well even if only given a single example of the target behavior. In this chapter, I present a technique that can be used to efficiently *search* for a given instance of a problem behavior in a large video collection.

There are two key challenges in doing action retrieval for problem behaviors: 1) We only have a single (or a few) example of the behavior. I want to support the use case of retrieving instances of new behaviors. Some of the problem behaviors occur only occasionally. Thus, the ability to search for this problem behavior from a video corpus without having to do extensive effort to collect numerous samples for training is important for practical application. 2) We need to handle many sources of variations: viewpoint, lighting, size and appearance of actors. In this chapter I introduce a novel action representation based on motion dynamics that is robust to such variations, while still being discriminative.

Currently, state-of-the-art performance in action classification is achieved by extracting dense local features (e.g., histogram of oriented gradients, motion boundary histogram) and grouping them in a bag-of-features (BOF) framework [155]. The basic BOF representation ignores information about the spatial and temporal arrangement of the local features by pooling them over the entire video volume. More recently, it has been shown that considering the spatial and temporal arrangements (dynamics) of an action (e.g., extracting a separate BOF model for each subvolume of a video [88, 155] or modelling the spatio-temporal arrangements of the interest points [171]) adds more discriminative power to the representation.

My approach is based on the observation that the dynamics of an action provide a powerful cue for discrimination. In Johansson's moving light display experiment, it was shown that humans perceive actions by abstracting a coherent structure from the spatio-temporal pattern of local movements [69]. While humans respond to both spatial and temporal information, the spatial configuration of movements that comprise an action is strongly affected by changes in viewpoint. This suggests that representing the temporal structure of an action could be valuable for reducing the effect of viewpoint. Motivated by this observation, I define human actions as a composition of temporal patterns of movements.



(a) Frame 10 (b) Frame 31 (c) MPHs

Figure 3: Movement Pattern Histogram for *checkwatch* action. (a)-(b): Arrows indicate optical flow direction and are color coded according to the flow words (flows are subsampled for presentation). (c): MPH set for *checkwatch*. (Best viewed in color)

My key hypothesis is that the temporal dynamics of an action are similar across views (assuming there is minimal occlusion). For example, the timing pattern of acceleration and deceleration of the limbs is largely preserved under viewpoint changes. In my representation, an action is decomposed into movement primitives (corresponding roughly to body parts). I encode the fine-grained temporal dynamics of each movement primitive using a representation that I call the *movement pattern histogram* (MPH). I describe an action as a collection of MPHs (see Fig. 3).

An advantage of video-level pooling methods such as BOF is that computing similarity between representations can be done reliably using $L2$ or χ^2 distance function. In part this is because these representations discard the temporal structure of an action, obviating the need for temporal alignment as a part of the matching process. In contrast, computing similarity between two sets of MPHs requires alignment and I describe a novel method to do so using a simultaneous alignment and bipartite matching formulation. Such formulation allows for matching across viewpoints and we present an efficient algorithm to solve it.

My MPH representation can be used in two ways: 1) as a stand-alone action representation for action recognition/retrieval across multiple viewpoints; and 2) to complement existing BOF representations for action recognition. I demonstrate that my approach outperforms standard representations for cross-view recognition tasks in the IXMAS dataset [163]. I also show that my representation complements existing representations for the classification task in the UCF50 dataset [120]. Finally, I show that my representation yields state-of-the-art results for the task of action retrieval in the novel Stereotypy dataset that I introduce (stereotypies are repetitive body movement patterns frequently associated with autism and are often the target of behavioral therapy). In summary, this work makes three contributions:

- I introduce the *movement pattern histogram*, a novel representation of actions as a multi-channel temporal distribution of movement primitives.
- I present a novel optimization approach to matching movement pattern histograms across videos based on maximum bipartite graph matching.

- I introduce the Stereotypy dataset, a new annotated video corpus obtained by recording children with autism in a classroom setting[1]. This dataset is publicly available.

3.2 Action Representation

In this section I describe my action representation, the *movement pattern histogram* (MPH). MPH encodes the global temporal pattern of an action without requiring explicit tracking of features over time. In Sec. 3.3 I present an iterative method for matching two sets of MPHs.

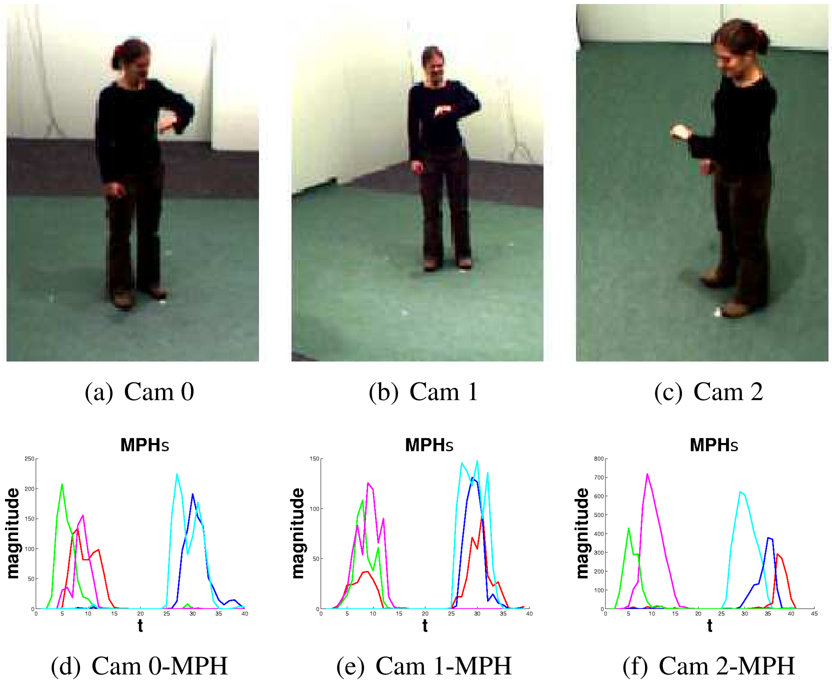
3.2.1 The Movement Pattern Histogram

To illustrate my approach, consider the action of a person checking a wrist-mounted watch seen from frontal view (Fig. 3). This action can be characterized by the upward movement of the hand and upper arm during the early part of the action (to bring the watch to a readable distance) and the downward movement of the same body parts at the end of the action. We can imagine encoding these body part movements with a cluster of flow vectors, where each cluster explains some portion of the total flow across the video. We denote these clusters as *flow words*. In the check-watch example, the upward hand movement might be mapped to a single flow word. That word would be present in the first half of the frames and absent in the other half (when the hand moves downward).

Given a set of extracted flow words, my goal is to represent an action by encoding the pattern of temporal occurrence of the flow words. In the example of Fig.3, the green and cyan words occur early in the action (when the hand and upper arm are raised) while the blue and magenta words occur later in the action. I construct an MPH for each flow word which encodes its dynamics.

I now describe the process of constructing the MPH representation. I assume that the

[1] Note that the Stereotypy dataset was collected under an IRB-approved protocol, following best-practices for research with vulnerable subject populations. Consent to publish has been obtained for all images and results.



(a) Cam 0 (b) Cam 1 (c) Cam 2

(d) Cam 0-MPH (e) Cam 1-MPH (f) Cam 2-MPH

Figure 4: (a)-(c): Three different views of the checkwatch action. (d)-(f): MPH representations of checkwatch for each view. Note the structural similarity of the MPH curves despite huge changes in viewpoint.

video is captured using a static camera (I relax this assumption in Section 3.2.2). First I compute dense optical flow over the video clip. Then, I use EM [37] to cluster together the flow vectors from all frames based only on the flow direction (we only consider flow vectors whose magnitudes are above a certain threshold). Each flow cluster defines a single flow word. In Figure 3(a)-3(b) we can see the flows color-coded according to the five flow words. I then generate an MPH for each of the flow clusters by binning the flow vectors. Each bin t in the MPH h_c corresponds to frame number t, and contains the sum of flow magnitudes for all pixel flows f that corresponds to cluster c in that frame. Let m_c denote the set of flow vectors that map to cluster c:

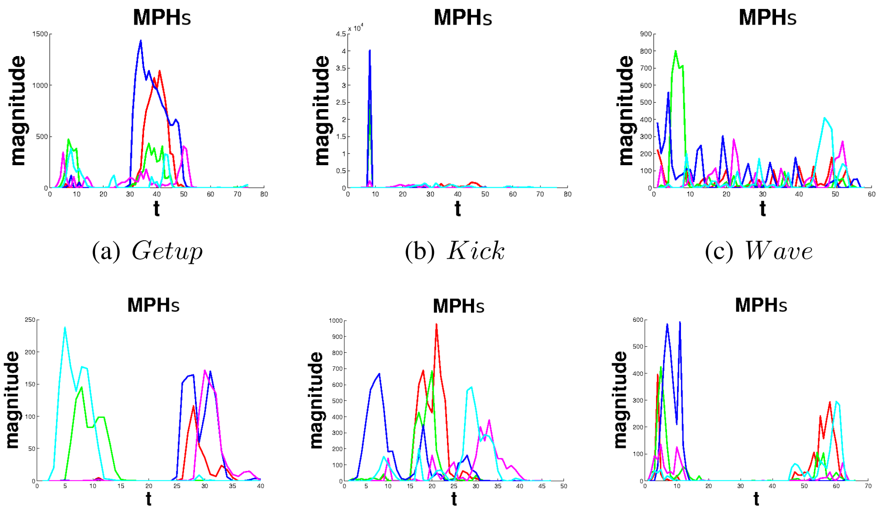
$$h_c(t) = \sum_{f(t) \in m_c} \|f(t)\| \tag{1}$$

In Fig. 3(c) we can see that the green MPH corresponding to upward hand movement is active at the beginning of the action and the blue MPH that corresponds to downward hand movement is active at the end. Note that MPH is quite different from other flow-based models such as the histograms of oriented optical flow (HOOF) [28]. HOOF models the distribution of optical flow direction in each frame, making it a viewpoint-dependent representation, while MPH models the temporal distribution of the *magnitudes* of the different flow clusters.

MPH differs in two ways from the standard histogram representations of visual words which are used in action recognition. First, each MPH corresponds to a single flow word and describes the variation in its magnitude over time. In contrast, BOF uses a single fixed histogram describing the co-occurrence of all visual words. Second, the MPH provides a very fine-grained temporal description (one bin per frame) but a very coarse spatial description (all occurrences of a word in a frame are binned together), in order to gain robustness to viewpoint variations.

Figure 4 illustrates the robustness of the MPH representation to viewpoint variation. We can see that the shapes of the MPH sets are quite similar in spite of substantial changes in viewpoint. To empirically evaluate the robustness of MPH to viewpoint variations, I perform an experiment where I use MPH for the task of cross-view action recognition: classify actions captured in a novel viewpoint given training data captured from other viewpoints (see Section 3.4.1). The results of the experiment show that MPH perform better than current state-of-the-art representations for cross-view action recognition, which demonstrates the robustness of MPH to viewpoint effects.

Figure 5 shows MPHs for different actions. The MPH representation achieves a certain invariance property under viewpoint changes because it marginalizes out information about appearance, spatial configuration, and flow direction of an action. While spatial configuration and appearance can be important for discriminating certain actions (e.g. high punch vs low punch), Figure 5 demonstrates that the temporal nature of an action can also be very discriminative. Note how MPH captures the dynamics of the different actions: wave (Fig. 5(c)) consists of hand moving left and right and this periodicity is reflected in the MPH. Even in cases where the mechanics of two actions are similar (checkwatch



(a) *Getup* (b) *Kick* (c) *Wave*

(d) *Checkwatch* (e) *Punch* (f) *Scratchhead*

Figure 5: MPHs of different actions.

and scratchhead both involve upward and downward movement of the hand), the dynamics of the actions make the MPH sets distinct (Fig. 5(d) vs 5(f)). To empirically evaluate the discriminate power of MPH, I evaluate its performance in the action recognition task (Sections 3.4.1 and 3.4.2). The results demonstrate that MPH is useful for discriminating a broad class of actions.

3.2.2 Compensating for Camera Motion

Sometimes action in the real world is captured using a moving camera. This can cause problems for our representations if we assume that all flows in the video are cause by the action. To minimize the effect of camera motion we can apply a video stabilization technique such as Grundmann et.al. [54] before computing dense optical flow. However, since we only need to remove the background motion between two consecutive frames (i.e., we don't need to produce smooth camera trajectory for the whole video), we can apply a simpler solution. We estimate the background motion by computing the homography between frames from the optical flow motion vectors (this is similar to Jain et.al. [67] but instead of assuming affine motion between frames we use homography). Using the dense optical flow computed, we select a subset of flow vectors located in textured regions (using criteria in Shi and Tomasi [135]) and perform homography estimation with RANSAC. From the estimated homography, we compute the camera-induced background motion for every pixel in that frame and then subtract the background motion from the computed flow vectors. We do this background motion estimation for every frame in the video and use the corrected flow vectors to compute MPH. Figure 6 shows the result of our motion compensation.

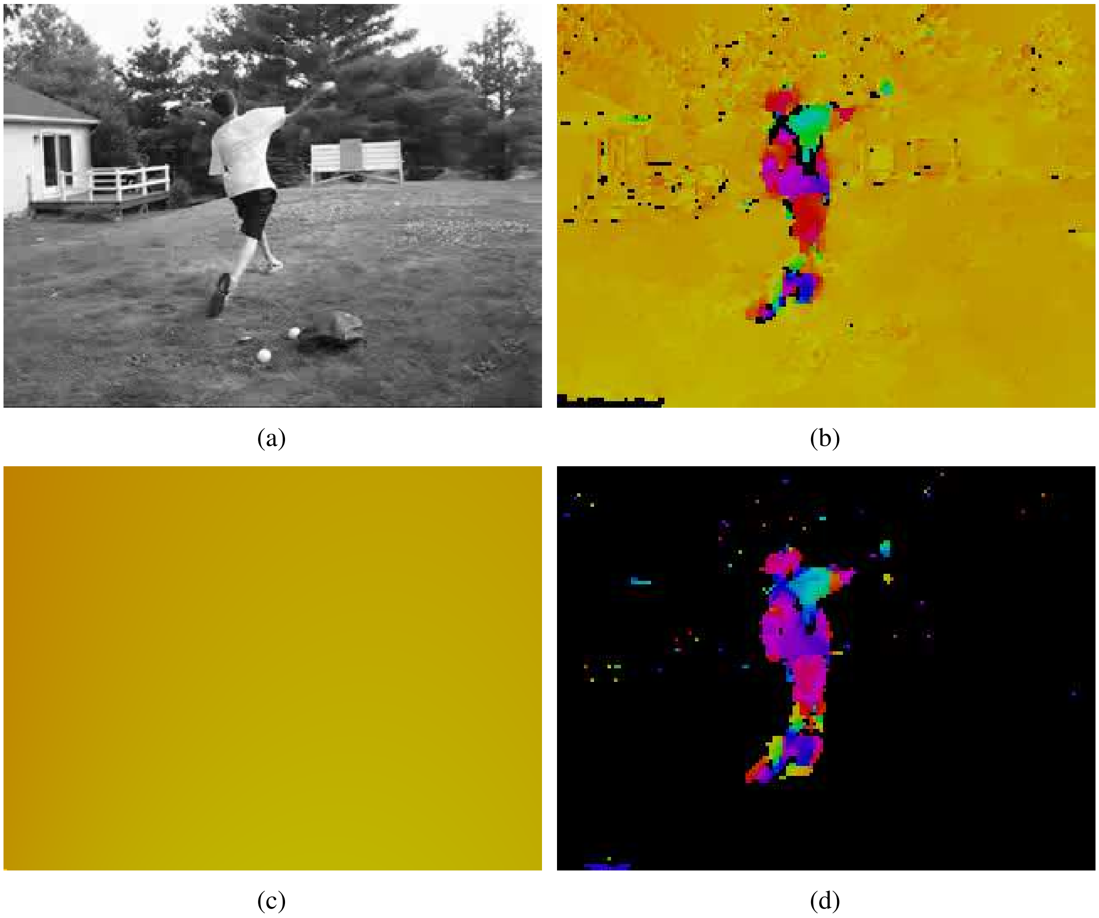


Figure 6: Motion compensation results from UCF50: b) Original flow, c) Estimated background motion, d) Motion compensated flow. Flows are color coded following the Middlebury convention.

3.3 Computing Similarity

Given my new MPH representation, how can we compute similarity between two videos – $target$ and $source$? Accurate similarity measure is important for action recognition and retrieval. My assumption is that if the two videos correspond to the same action, we can find matching in which the MPH pairs are highly correlated. Let $h_i^t \in \mathbb{R}^{l_t}$ and $h_j^s \in \mathbb{R}^{l_s}$ be the movement pattern histogram for primitives (clusters) i and j in the $target$ and $source$ videos, respectively (l_s and l_t are the number of frames of the videos). Note that since each video is clustered independently, there is no a priori relationship between MPHs from separate videos. Let $T = \{h_1^t, h_2^t, ...h_K^t\}$ and $S = \{h_1^s, h_2^s, ...h_K^s\}$, where K is the total number of flow words in the target and source video. We can construct an undirected bipartite graph $G = (V, E)$ where every single element of T is connected to every single element of S, the vertex set $V = T \cup S$, and $e_{ij} \in E$ is the edge between h_i^t and h_j^s. The weight of edge e_{ij} is the similarity measure between two signals h_i^t and h_j^s. I use the Pearson correlation coefficient (PCC) to compute e_{ij} due to its invariance to scaling: $e_{ij} = PCC(h_i^t, h_j^s) = \frac{cov(h_i^t, h_j^s)}{\sigma_{h_i^t}\sigma_{h_j^s}}$

The similarity between the target and source video is the maximum weighted bipartite matching score of graph G.

Simultaneous alignment and matching

Since an action can be performed at different speeds, the two sets of histograms S and T might not be temporally aligned. This negatively impacts my correlation measure. In order to overcome this problem, I propose a simultaneous alignment and matching method where we iteratively perform alignment and matching of S and T.

Let $\boldsymbol{H_s} = [h_1^s, h_2^s, ...h_K^s]$ and $\boldsymbol{H_t} = [h_1^t, h_2^t, ...h_K^t]$ be the matrices that we construct from S and T. Without loss of generality, let us assume that we normalize the MPH in S and T so that they all have zero mean and unit standard deviation. Also, we zero-pad each vector h_j^s and h_i^t such that $l_s = l_t = l$. Under this condition, finding the maximum weighted bipartite matching of graph G is equivalent to computing a $K \times K$ binary matrix $\boldsymbol{M}$ that minimizes $C_m = \|\boldsymbol{H_s}\boldsymbol{M} - \boldsymbol{H_t}\|_F^2$, where $\Sigma_i \boldsymbol{M}(i,j) = 1, \Sigma_j \boldsymbol{M}(i,j) = 1$.

To align $\boldsymbol{H_s}$ and $\boldsymbol{H_t}$, we can use dynamic time warping (applying DTW or its variants e.g., Zhou et.al. [179] on a time series data is a common approach for doing activity alignment) to compute binary matrices $(\boldsymbol{D_s}, \boldsymbol{D_t})$ that minimize $C_{dtw} = \|\boldsymbol{D_s H_s} - \boldsymbol{D_t H_t}\|_F^2$, where $\Sigma_j \boldsymbol{D_s}(i,j) = 1$ and $\Sigma_j \boldsymbol{D_t}(i,j) = 1$. Note that DTW optimization infers $\boldsymbol{D_s}$ and $\boldsymbol{D_t}$ using dynamic programming such that the temporal ordering of the rows in $\boldsymbol{H_s}$ and $\boldsymbol{H_t}$ is preserved. The DTW solution $(\boldsymbol{D_s}, \boldsymbol{D_t})$ are binary matrices of size $l' \times l$ where l' is the length of the alignment path between $\boldsymbol{H_s}$ and $\boldsymbol{H_t}$. Putting the previous two steps together, we get the final cost function that we want to minimize:

$$\begin{aligned} & C_{mdtw} = \|\boldsymbol{D_s H_s M} - \boldsymbol{D_t H_t}\|_F^2 \\ \text{where} \quad & \Sigma_i \boldsymbol{M}_{ij} = 1,\ \Sigma_j \boldsymbol{M}_{ij} = 1 \\ & \Sigma_j \boldsymbol{D_s}(i,j) = 1 \\ & \Sigma_j \boldsymbol{D_t}(i,j) = 1 \end{aligned} \tag{2}$$

Optimizing C_{mdtw} is a non-convex optimization problem with respect to the matching matrix $\boldsymbol{M}$ and alignment matrices $\boldsymbol{D_s}$ and $\boldsymbol{D_t}$. We can perform iterative optimization by alternating between computing $(\boldsymbol{D_s}, \boldsymbol{D_t})$ and $\boldsymbol{M}$:

1. Set $\boldsymbol{M}$ as $K \times K$ identity matrix
2. Fix $\boldsymbol{M}$ and minimize $C_{dtw} = \left\|\boldsymbol{D_s H_s^m} - \boldsymbol{D_t H_t}\right\|_F^2$, where $\boldsymbol{H_s^m} = \boldsymbol{H_s M}$, to optimize for $(\boldsymbol{D_s}, \boldsymbol{D_t})$
3. Fix $(\boldsymbol{D_s}, \boldsymbol{D_t})$ and minimize $C_m = \left\|\boldsymbol{H_s^{dtw} M} - \boldsymbol{H_t^{dtw}}\right\|_F^2$, where $\boldsymbol{H_s^{dtw}} = \boldsymbol{D_s H_s}$ and $\boldsymbol{H_t^{dtw}} = \boldsymbol{D_t H_t}$, to optimize for $\boldsymbol{M}$
4. Iterate 2-3 until convergence

Both step 2 and 3 monotonically decrease/non-increase C_{mdtw}. Since C_{mdtw} has a lower bound of 0 this optimization will converge. DTW can be solved in $O(l^2)$ and minimizing C_m using the Hungarian algorithm takes $O(K^3)$. Hence the complexity of this algorithm is $O(l^2) + O(K^3)$ and since l and K are typically small this is efficient to compute (l is typically between 60-150 depending on how long the action is. K depends on the number of peaks of the flow distribution, typically between 4-6). Empirically I observe that this optimization converges after 2-3 iterations.

To optimize for $\boldsymbol{M}$, the task is to find the set of edges $e_{ij} \in E$ that defines a perfect matching in G such that the sum of the edges in the matching is maximum. I solve this using the Hungarian algorithm to compute a set of λ for the following problem:

$$\begin{aligned} \max_{\lambda} \quad & \sum_{(i,j)\in E} \lambda_{ij} e_{ij} \\ \text{s.t.} \quad & \sum_{j\in N(i)} \lambda_{ij} = 1 \; \forall i \in source \\ & \sum_{i\in N(j)} \lambda_{ij} = 1 \; \forall j \in target \\ & \lambda_{ij} \in \{0,1\} \end{aligned} \tag{3}$$

where e_{ij} is the correlation between the i-th column of $\boldsymbol{H_s}$ and j-th column of $\boldsymbol{H_t}$, and $N(i)$ is the set of vertices that are adjacent to vertex i.

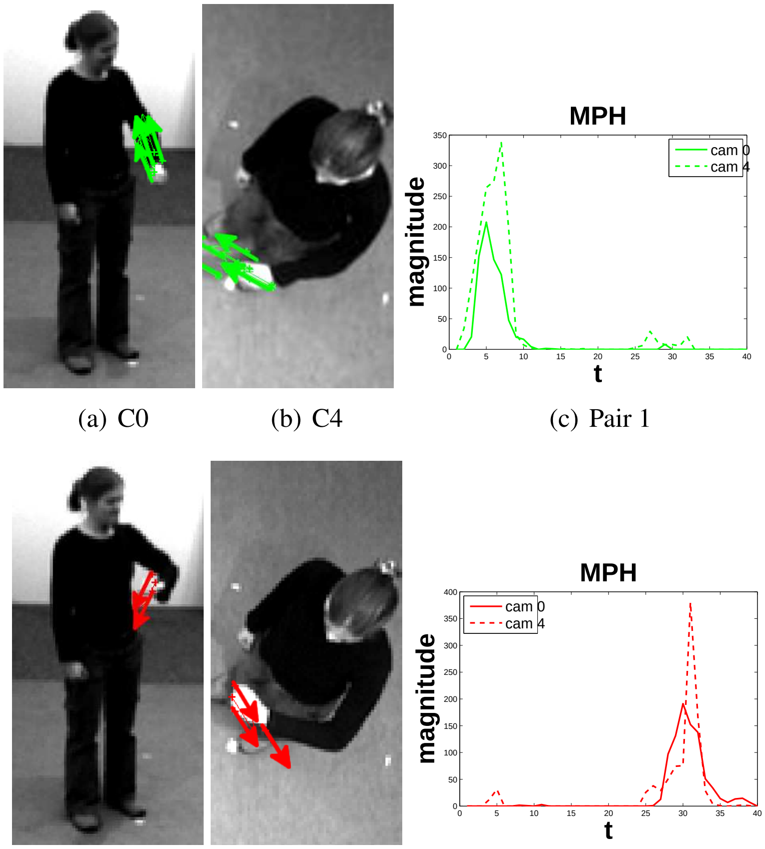
After obtaining the λ for maximum matching, I define the similarity score between two videos as the maximum weighted bipartite matching score of graph G:

$$score = \sum_{(i,j)\in E} \lambda_{ij} e_{ij} \tag{4}$$

Figure 7 illustrates an example of the matching result. Note that while the two actions are captured from widely different viewpoints, my matching algorithm is able to establish the correspondence between flow clusters by exploiting the temporal property of MPH. For instance, the matched MPH pair 1 (Fig. 7(c)) corresponds to flow words that belong to the hand while it is moving up at the beginning of the action (Fig. 7(a)-7(b)) and the matched pair 3 (Fig. 7(f)) corresponds to flow words of the hand while it is moving down at the end of the action (Fig. 7(d)-7(e)).

3.4 Experimental Results

To evaluate the performance of my method I performed experiments on the IXMAS dataset [163], UCF50 dataset [120] and a new real-world Stereotypy dataset that consists of a collection of videos ranging from 10 to 20 minutes each (with total length of 31 hours). I



(a) C0 (b) C4 (c) Pair 1

(d) C0 (e) C4 (f) Pair 3

Figure 7: Matching of cam 0 and cam 4 for checkwatch. Note how the matched MPH pair correspond to the same body part movements ((a)-(b): Hand moving up, (d)-(e): Hand moving down).

consider three tasks. First, to empirically show the robustness of my representation to variations in viewpoint, I perform cross-view action recognition experiments on the IXMAS dataset. Second, to show how my MPH complements existing BOF representation, I perform action recognition on the UCF50 dataset. Finally, I demonstrate the power of my approach on a real-world behavior search task by doing action retrieval on the Stereotypy dataset collected by our collaborators in Center for Discovery, an institution that serves individuals with developmental disabilities.

For the action retrieval task, I compare retrieval results against two BOF representations: the self-similarity matrix (SSM) [73], a representation that has been shown to be robust to viewpoint effects, and improved trajectories (IT) [156], currently the most popular action representation (e.g., most submissions in the THUMOS action recognition challenge [51] use this representation) and has been shown to achieve state-of-the-art results on many action recognition tasks. Note that BOF representation has been previously used for action retrieval [26].

To compute MPH I used GPU-based dense optical flow [165]. To select the K for MPH I examined the number of peaks in the distribution of flow directions in sample videos. I chose $K = 5$ (5 MPH per video) for all experiments. For action retrieval comparisons, I found the following k works best for the different features:

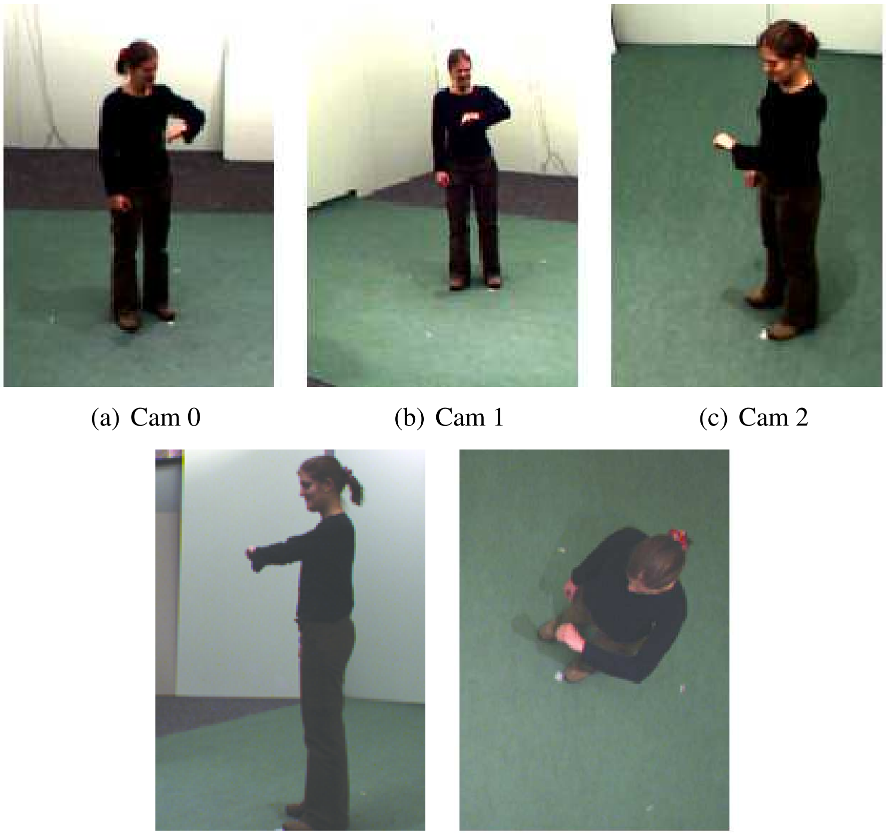
- SSM: 150
- Improved trajectories: 100

3.4.1 Robustness to Viewpoint Variations

To demonstrate the robustness of my representation to viewpoint effects, I perform the cross-view classification experiment on the IXMAS, a standard dataset to empirically test how viewpoint affects an action representation. The IXMAS dataset contains videos of 11 types of actions captured from 5 viewpoints (see Figure 8). Each action in the dataset is performed by several actors (30 examples per action) to ensure that the dataset captures the intraclass variations of actions.

For this experiment I use 1-Nearest Neighbor (1-NN) classifier and a 6-fold cross validation procedure (identical cross-validation procedure to other published results for this dataset [92, 90, 177]). I compare my results against various cross-view action recognition methods [73, 92, 90, 177].

I focus on two recognition tasks: 1) classifying videos captured from the $test$ view using training data captured from the $train$ view; and 2) classifying videos captured from the $test$ view using training data from all of the other views. It is important to note that in this particular experiment I am not assuming any view-correspondence in the training data since in many applications the need to have multi-view correspondence for training can be



(a) Cam 0 (b) Cam 1 (c) Cam 2

(d) Cam 3 (e) Cam 4

Figure 8: The 5 viewpoints in the IXMAS dataset.

burdensome.

The results for the first recognition task (classifying videos from the test view using training from the train view) can be seen in Table 3.4.1. For this task, my method improves the average recognition accuracy by 2.5% compared to the next best approach (see the highlighted cell in Table 3.4.1). Hankelet [90] is only robust to affine transformation and thus achieves low accuracy when classifying videos trained from very different viewpoints (e.g., their numbers in the c4 column are very low due to the viewpoint of cam 4 being significantly different from cam 0-3). My representation also achieves significantly better results than SSM when classifying actions from very different viewpoints (e.g. accuracy of 56.5% for c4 compared to 49.6% of SSM). Even when compared against an approach that uses multi-view correspondence for training the representation[92], my approach still produces better classification accuracy.

Table 1: Classification results by using a single view for training on IXMAS. Each row is a training view, and column a test view.

	Test View					
	c0 Ours, [92], [90], [73]	c1 Ours, [92], [90], [73]	c2 Ours, [92], [90], [73]	c3 Ours, [92], [90], [73]	c4 Ours, [92], [90], [73]	Avg. Ours, [92], [90], [73]
c0		80.3, 63.6, **83.7**, 75.2	63.6, 60.6, 59.2, **69.7**	68.5, 61.2, 57.4, **71.8**	**56.4**, 52.6, 33.6, 49.4	**67.2**, 59.5, 58.5, 66.5
c1	80.0, 61.0, **84.3**, 78.5		62.1, 62.1, 61.6, **67.9**	59.7, 65.1, 62.8, **71.5**	47.9, **54.2**, 27.0, 48.0	62.4, 60.6, 58.9, **66.5**
c2	63.6, 63.2, 62.5, **70.0**	62.1, 62.4, 65.2, **73.0**		**79.7**, 71.7, 72.0, 68.5	**75.5**, 58.2, 60.1, 55.2	**70.2**, 63.9, 64.9, 66.7
c3	67.0, 64.2, 57.1, **73.6**	65.8, 71.0, 61.5, **72.4**	**83.6**, 64.3, 71.0, 67.3		46.4, **56.6**, 31.2, 45.9	**65.7**, 64.0, 55.2, 64.8
c4	**54.5**, 50.0, 39.6, 44.5	49.4, **59.7**, 32.8, 41.5	**72.1**, 60.7, 68.1, 55.2	50.0, **61.1**, 37.4, 37.9		56.5, **57.9**, 44.5, 44.8
Avg.	66.3, 59.6, 60.9, **66.7**	64.4, 64.2, 60.8, **65.5**	**70.4**, 61.9, 65.0, 65.0	64.5, **64.8**, 57.4, 62.4	**56.5**, 55.4, 38.0, 49.6	**64.4**, 61.2, 56.4, 61.9

The results for the second recognition task (classifying videos from the test view by using training from all of the other views) can be seen in Table 2. Note that the results of Junejo et.al. [73] for this task is obtained by including videos from all the views (including the test view) for training. Even then, my approach still yields the best result, outperforming other methods by a significant margin. This demonstrates that *my method can use the additional training views more effectively due to its ability to generalize across viewpoints.*

Table 2: Cross-view recognition accuracy on IXMAS (trained on videos captured from all views except the $test$ view). Note how our representation gives a significantly more accurate result.

Method	**Test View**					
	c0	*c1*	*c2*	*c3*	*c4*	*Avg.*
Ours	**83.9**	**81.8**	**87.6**	**83.0**	**73.6**	**82.0**
[177] (test view used for transfer learning)	66.4	73.5	71.0	75.4	66.4	70.5
[92] (test view used for transfer learning)	62.0	65.5	64.5	69.5	57.9	63.9
[73] (trained on all cameras)	77.0	78.8	80.0	73.9	63.6	74.6

3.4.2 Results on UCF50 Dataset

The UCF50 dataset contains 6618 videos of 50 types of actions. For this experiment I use the leave-one-group-out (LoGo) cross validation as suggested by Reddy et.al. [120].

Many videos in UCF50 were captured using low resolution handheld cameras with various motion artifacts due to camera shake and rolling shutter. Clearly, the fine-grained motion features that my method exploits are difficult to extract in this case. However it is valuable to characterize the limitations of my approach by analyzing the UCF50. Another important characteristic of this dataset is that the scene context gives a significant amount of information about the type of action in the video. For example, many of the actions are performed using a specific set of instruments (e.g. barbell in bench press) and representing those cues can help immensely for classification. This suggests the need to combine my representation (which only models the dynamics of an action) with a complementary appearance-based representation.

I combine my approach with Fisher Vector (FV) encoding [114] (which can be seen as an extension to BOF) of the dense trajectory descriptor described in [155]. To convert my pairwise action similarity measure to a feature vector I use a method similar to ActionBank [129]. In ActionBank, the videos in the training set function as the bases of a high-dimensional action-space. For example, if we have N videos in the training set, the feature vector for video v is a vector of length N where the value of $N(i)$ is my similarity measure between video v and the i-th video in the training set. The full feature vector for each video is then simply a concatenation of the FV representation of dense trajectory and our ActionBank-like representation. For this experiment I use 1-vs-all linear SVM (with $C = 0.1$) for training and classification.

Classification results on this dataset can be seen in Table 3. The accuracy improvement obtained by adding my representation suggests that MPH encodes information that is complementary to HOG, HOF and MBH.

Comparing results of MPH + FV of dense trajectory against only FV of dense trajectory, the most significant improvement in accuracy comes from the class PizzaTossing (an improvement of 10.5% from 65.8% to 76.3%). A large part of this improvement comes from a better discrimination between PizzaTossing and Nunchucks classes. Many of the videos of these these two classes share a significant similarity in appearance: a person performing an action in a small room captured from close to frontal view. Thus, MPH (which

Table 3: Classification results on UCF50.

Method	**Accuracy (LoGo)**
Ours (MPH) + FV of DT [155]	**90.5**
Dense trajectories [155] w/ FV encoding	88.9
MBH + scene context[120]	76.9
GIST3D + STIP [136]	73.7
MIP [82]	72.7

models the dynamics of the action) increases discrimination between these two classes. Another notable improvement comes from the class RockClimbing (an improvement of 6.9% from 85.4% to 92.3%). About half of the improvement for this class comes from a better discrimination against RopeClimbing. While the actual movement of climbing a rope vs climbing a wall with a rope is different, the context of these two classes are very similar since wall and rope tend to be the prominent features in the video. Thus, MPH provides a powerful cue to help discriminate between these two classes. On the other hand, MPH can also increase confusion between classes. I observe the biggest drop in accuracy in the class HorseRace (a decrease of 3.1% from 98.4% to 95.3%) partly due to increased confusion with Biking. This is likely due to the fact that from a distance, the movement dynamics of HorseRace and Biking look similar: people moving on a trajectory with their body moving slightly up-and-down with a particular frequency. Human action is a complex concept defined by the interplay of a number of elements: movements, human pose, instruments used, and surrounding background context. A better approach to modellng any of these elements is a step towards a better action representation.

3.4.3 Results on Stereotypy Dataset

Now I will use my approach to address the problem of *action retrieval*: Given a single example video clip containing an action of interest, the task is to retrieve all matching instances of that action from an unstructured video collection. The strength of my bottom-up matching approach is that it can compute a similarity measure between activities without learning. It can therefore be used in situations where the space of possible activities is very

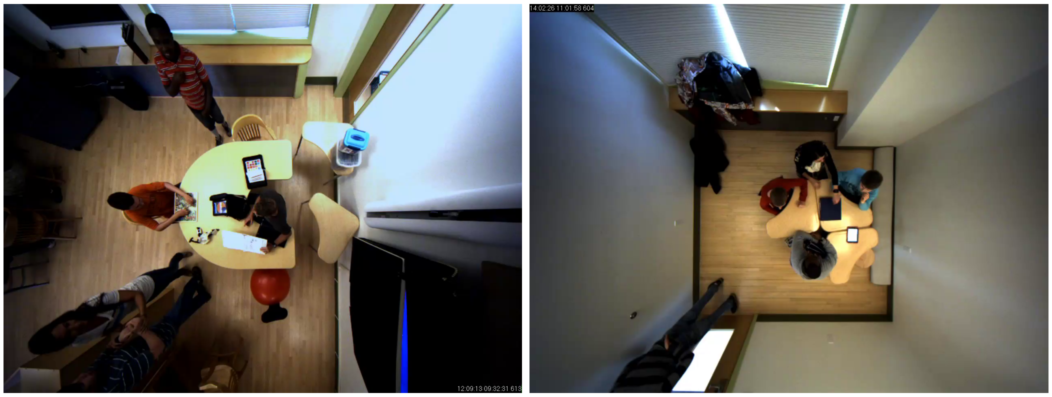


Figure 9: Two representative images from the Stereotypy dataset.

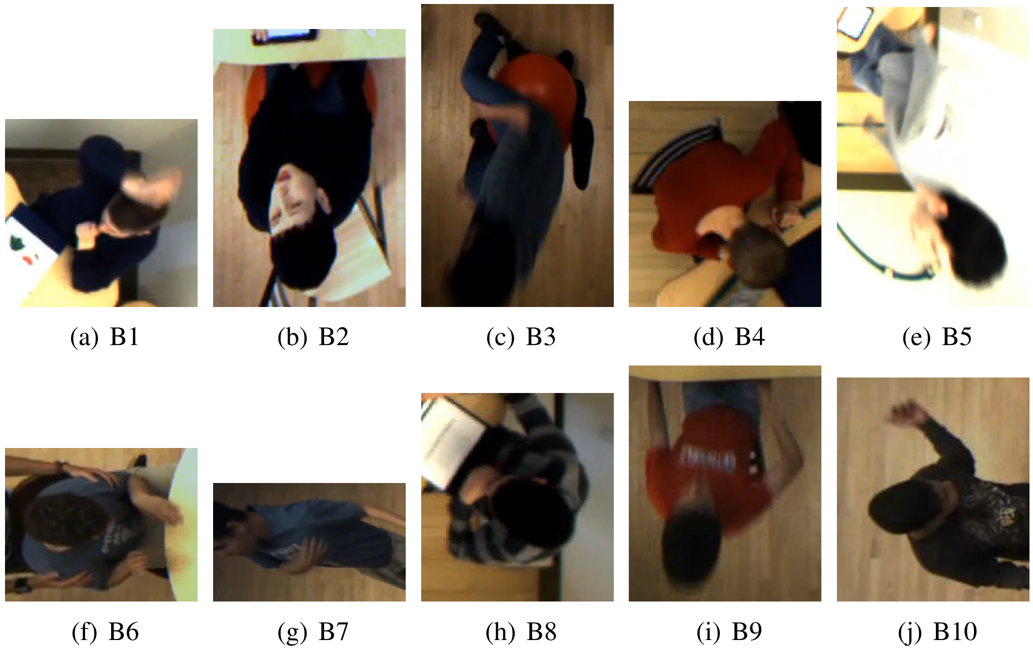
large and difficult to define *a priori* and when it is difficult to find an extensive amount of training examples across different views.

In the domain of behavioral psychology, there is currently great interest in studying the effectiveness of behavioral therapy for children with autism. These children frequently exhibit repetitive motor movements, known as *stereotypies*. In comparison to more traditional functional activities, stereotypies are often unique expressions of individual behavior, making it challenging to construct a general model of such behaviors. At the same time, it would be very useful to be able to retrieve all instances of a particular stereotypy exhibited by a child across multiple recording sessions given only a single example. I conducted an experiment to evaluate the effectiveness of my algorithm in this context.

I collaborated with experimental and educational psychologists in the Center for Discover, an institution that serves individuals with developmental disabilities, on analyzing videos obtatined of children with autism who engage in stereotypies in a classroom setting. The dataset consists of 31 hours of videos from two overhead cameras located in two separate rooms. Representative frames are shown in Figure 9. Note that even though the activities in both rooms are recorded from an overhead camera, there is still a significant variation in the orientation of the people with respect to the camera since the people are free to move within the room area (see Figure 10). Our collaborators have identified 10 different target behaviors from these videos (Table 4), out of which 9 are subject–specific stereotypies and 1 is a general behavior (out of seat). Note that the target behaviors encompass many of the common classes of stereotypical behaviors [86]. For the annotation, each bout of the target behavior is assumed to last for 2 seconds. A psychologist with autism expertise and familiarity of the children provided the groundtruth labels for the target behaviors.

Table 4: Behaviors of interest in the Stereotypy dataset. All behaviors except B5 are subject–specific.

Behavior name	**Count**
Hand to the head (B1, Fig. 10(a))	24
Bounce on ball (B2, Fig. 10(b))	27
Jump on ball (B3, Fig. 10(c))	54
Head on arm (B4, Fig. 10(d))	49
Out of seat (B5, Fig. 10(e))	199
Paddling on objects (B6, Fig. 10(f))	23
Paddling on self (B7, Fig. 10(g))	23
Bite hand (B8, Fig. 10(h))	60
Body rocking (B9, Fig. 10(i))	23
Play with hand (B10, Fig. 10(j))	30



(a) B1 (b) B2 (c) B3 (d) B4 (e) B5

(f) B6 (g) B7 (h) B8 (i) B9 (j) B10

Figure 10: Visual examples of the target behaviors in the Stereotypy dataset.

Table 5: $\%$ of database one must go through to get 50% recall (lower is better).

Behavior	**Ours** 1 sample	**Ours** 2 samples	**IT** 1 sample	**IT** 2 samples	**SSM** 1 sample	**SSM** 2 samples
B1	9.7%	9.0%	16.5%	15.7%	12.8%	12.1%
B2	7.3%	6.5%	16.0%	13.8%	13.5%	12.1%
B3	8.3%	8.0%	14.9%	14.1%	16.2%	16.3%
B4	7.2%	6.6%	18.2%	17.0%	14.5%	13.9%
B5	4.9%	4.5%	9.2%	7.6%	10.9%	10.4%
B6	8.8%	8.6%	13.2%	12.3%	13.9%	13.1%
B7	6.3%	5.0%	13.4%	11.4%	14.8%	12.8%
B8	9.0%	8.7%	18.3%	17.2%	13.0%	12.6%
B9	2.9%	1.8%	13.6%	11.5%	13.8%	13.8%
B10	7.6%	7.1%	17.0%	15.6%	10.6%	10.0%
Avg.	7.2%	6.6%	15.0%	13.6%	13.4%	12.7%

For the experiment, I used sliding window to split the videos spatially into 5×8 overlapping windows. For each of these windows, I temporally segment them into a series of 60-frame clips (2 seconds) with stride length of 15. This process segments each video into many small subvolumes which I would refer to as clips from this point forward.

Action retrieval task: We identified all clips containing stereotypies, and used each of those clips as the $target$ input for retrieval. Given a $target$ clip, I computed the similarity of the clip against the rest of the clips in the dataset. The similarity score for my algorithm is described in Section 3.3. For BOF, I found that the $L2$ distance between the normalized feature vectors yielded the best results. I then ranked the videos according to the similarity score and measured performance by using the Cumulative Match Characteristic (CMC) curve, a common metric for retrieval. The CMC curve describes how far along one has to go down the retrieval result (x-axis) to achieve a certain recall rate for the target (the y-axis). I counted a clip as a hit if it temporally overlaps with at least 50% of the groundtruth.

In addition to performing retrieval using only a single example, I also perform the retrieval experiment if the system were to be given two examples of the same target behavior. This is done to show the effect of having additional information to the different representations. For this experiment, the similarity score is simply the max of the similarity to the two $target$ clips.

One of the main use cases of a behavior retrieval system is for a clinician to search for additional examples of a new target behavior given minimal (one or two) input samples. Individuals with ASD sometimes exhibit a new problem behavior which can be caused by a number of reasons (e.g., environmental factors, change in sleeping pattern). When a clinician observed a new problem behavior in a video recording, he will be interested to see additional examples of such behavior. Note that the goal here is not 100% recall, but rather for the clinician to see enough examples of the new behavior so that he can understand the behavior better (e.g., to form a hypothesis for the underlying cause, to assess the severity of the behavior, etc.). To simulate this use case, I show how much of the videos in the dataset one must go through in order to get 50% recall of the target behavior (Table 5).

We can clearly see how my approach outperforms the competing methods in this scenario. My method performs well on behaviors that involve gross motor movements such as body rocking (B9), out of seat (B5) and paddling on self (B7). The limitation of my method is that it is only encoding motion, thus my approach sees a drop in accuracy when used for retrieving behaviors that are largely characterized by the body/limb configuration such as hand to the head (B1) and bite hand (B8). The improved trajectories (IT) feature performs the poorest in this setting since it is heavily dependent on the viewpoint and appearance information. SSM performs better than IT since it is robust to viewpoint changes. However, the SSM representation is highly affected by the presence of clutter in the background (e.g., other people/objects) since it relies on computing pairwise similarity between the frames. So, even though it works well on a constrained setting when the video only contains the target subject [73], it performs poorly in this setup. Adding an additional sample as an input improves the performance of all methods (see results with 2 samples in Table 5). However, my approach, even when using only a single input sample, still performs much better than the competing methods with 2 input samples. This demonstrates the effectiveness of my representation for this task.

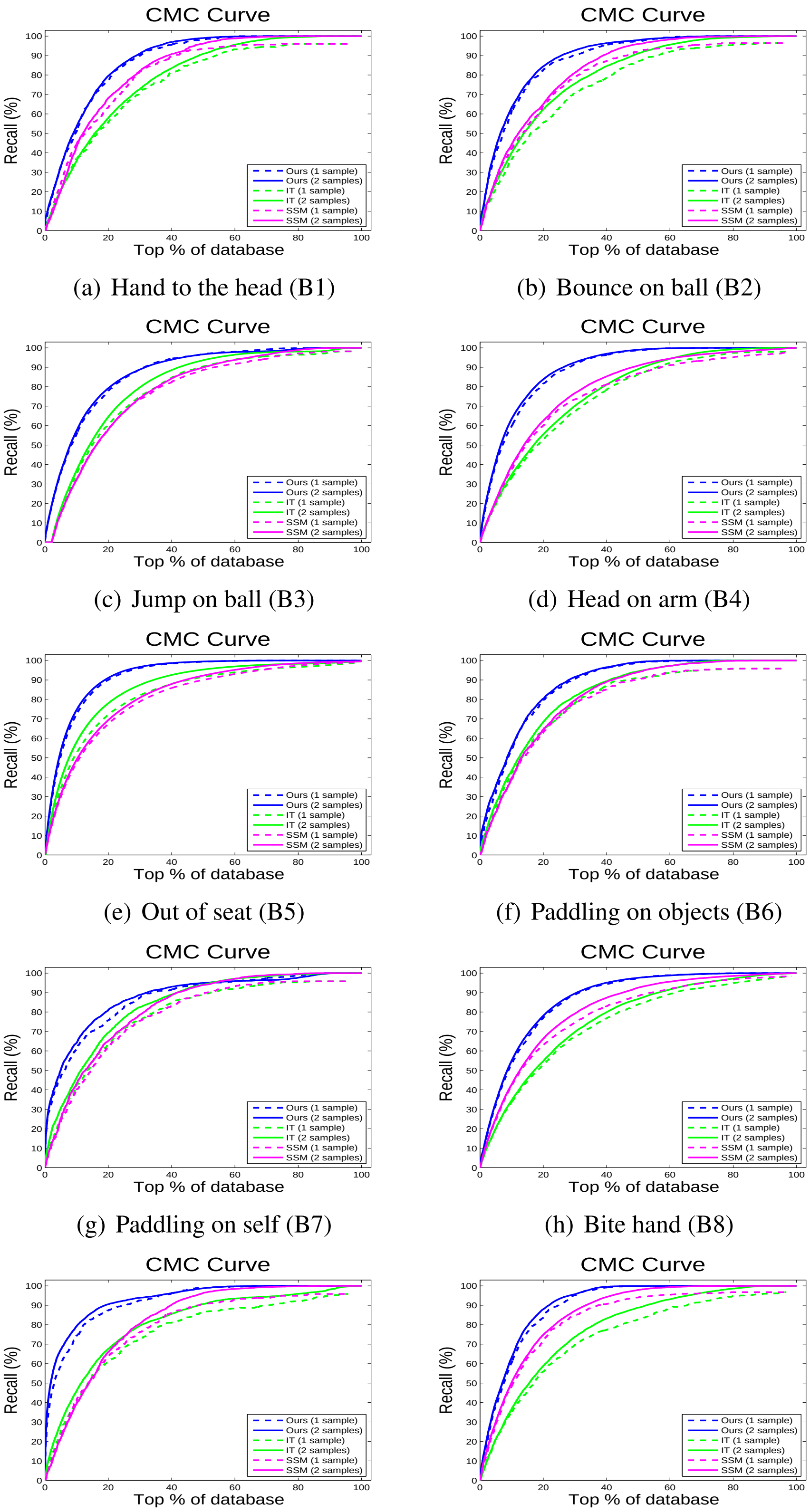
The CMC curves for retrieving stereotypies can be seen in Figure 11. Note that my method performed significantly better than the competing BOF representations across all behaviors. There are several factors that contributed to this: the child (target) might be wearing different clothing in the different video, a therapist sometimes came to interact with the child during the course of the class, and also the child often moved, changing his relative angle to the camera (for example, Figure 10 illustrates how the different seating positions affect a person's orientation with respect to the camera). All these variations will affect any representations that heavily depend on appearance information. My approach, which encodes the temporal pattern of the movements of a behavior, is more robust to these variations (as demonstrated by the higher recall rate across all behaviors in Figure 11).

3.5 Conclusions

Behavior retrieval from arbitrary viewpoints is a novel and practically-important problem that has not been addressed before. I tackle the retrieval problem in two stages: 1) formulation of a feature (MPH) that is easy to extract and robust to viewpoint variations; and 2) simultaneous matching and alignment formulation that explicitly handles variations in the dynamics of an action and allows matching of features extracted from different viewpoints.

The experimental results from the Streotypy Dataset demonstrate that current action representation approaches are not suitable for the behavior retrieval task which requires robustness to many sources of appearance variations. I have demonstrated that by leveraging the unique pattern of the dynamics of a behavior, my approach achieves the best performance on retrieving instances of problem behaviors exhibited by individuals with autism. This can potentially impact how clinicians perform behavior analysis on a large video collection by providing them with a mechanism to more quickly search for relevant behaviors.

In addition to obtaining state-of-the-art performance in behavior retrieval, I have also showed that my representation performs well on the action recognition problem. My approach obtains state-of-the-art results on the cross-view action recognition task, and when complemented with the existing BOF representation, performs well on the classical action recognition task.



(a) Hand to the head (B1)

(b) Bounce on ball (B2)

(c) Jump on ball (B3)

(d) Head on arm (B4)

(e) Out of seat (B5)

(f) Paddling on objects (B6)

(g) Paddling on self (B7)

(h) Bite hand (B8)

(i) Body rocking (B9)

(j) Play with hand (B10)

Figure 11: CMC curves for all behaviors.

CHAPTER IV

MINIMIZING HUMAN EFFORT IN INTERACTIVE TRACKING BY INCREMENTAL LEARNING OF MODEL PARAMETERS

4.1 Introduction

The past decade has seen an explosive growth of video data. The ability to easily annotate/track objects in videos has the potential for tremendous impact across multiple application domains. For example, in computer vision annotated video data can be used as an extremely valuable source of information for the training and evaluation of object detectors (video provides continuous view of how an object's appearance might change due to viewpoint effects). In sports, video-based analytics is becoming increasingly popular (e.g. the Italian company Deltatre employed 96 people to pour over multiple video footage for live player tracking during the 2014 World Cup). In behavioral science, video has been used to assist the coding of children's behavior (e.g. for studying infant attachment [9], typical development [121] and autism [95]).

The tracking problem is challenging because of the often dramatic appearance variations in the object being tracked (e.g., due to lighting and viewpoint change) and occlusions. As a result, fully-automated high precision object tracking remains an open problem. Note that getting accurate object tracks is important in many applications. For example, biologists who use video to monitor the movement of animals care about accurately tracking these animals at all times. Errors in tracking are unacceptable since they can contaminate the research findings. To obtain practically useful accurate tracking, several *interactive approaches* have been pursued (e.g., LabelMe Video [172] and the crowdsourcing method of Vondrick et.al. [151]). Unfortunately, most existing interactive tracking approaches are not optimized for human effort. However, minimizing human annotation effort is extremely important in practice since video can be prohibitively expensive to label (e.g., twenty six hours of surveillance video cost tens of thousands of dollars to annotate despite using a state-of-the-art annotation system [109]).

In this chapter I propose an interactive tracking system that is designed to minimize the amount of annotation required to obtain high precision tracking results. I achieve this by leveraging user annotations for incrementally learning *instance specific* model parameters of the tracking cost function. This is in contrast to the common practice of hand-tuning the model parameters on a training set and applying the same fixed parameters on any new testing data. This approach is both time consuming (due to hand-tuning) and gives suboptimal accuracy on individual tracking instances. I cast the problem of learning the optimal model parameters as the problem of learning a structured prediction model in a maximum margin framework. My key insight is that the incremental nature of an interactive tracking process is particularly well-suited for *efficient* maximum margin learning of model parameters.

Related to this work is the work by Taskar et.al. [143] in learning structured prediction model using large margin approach. However, their work is not focused in the tracking problem (they apply their approach to protein structure prediction). Szummer et.al. [141] apply Taskar's [143] and Tsochantaridis' [148] large margin approach to the problem of learning CRF using graph cuts in the context of image segmentation. In contrast to these works, my approach exploits the sequential nature of an interactive tracking process for online *incremental* learning of a structured model, departing from the classical training-testing paradigm.

I show that my approach significantly outperforms the current best practice of using hand-tuned model parameters on two datasets: the VIRAT challenge dataset and the Infant-Mother Interaction dataset that I introduce. The main contribution of this work is an annotation-driven maximum margin framework for efficiently learning instance-specific model parameters.

4.2 Object Tracking

In this section I describe the framework that I use for estimating object track in a video. I first give a description of the object representation technique in Section 4.2.1. I then present the formulation for estimating object trajectory given a set of observations (Section 4.2.2). Finally, I describe an efficient approach to optimize for object trajectory in Section 4.2.3.

4.2.1 Object Representation

The choice of object representation significantly impacts tracking performance. For this work, I use the Aggregated Channel Features (ACF) [40] which has been shown to achieve good performance in the task of object detection. ACF is constructed by combining three features: normalized gradient magnitude (Grad), histogram of oriented gradients (HOG) [33] and color channel in LUV colorspace:

$$x = [Grad\ HOG\ LUV]^T \tag{5}$$

To model the global appearance of the object, I use a discriminative approach. For each annotated frame, I use the annotated bounding box and some perturbed version of it as the positive instances and extract a large number of negative bounding boxes that do not overlap (or have very minimal overlap) with the annotation. To learn the object model, I use the positive and negative instances to train a linear SVM. For every frame, I detect K object candidates using the learned model (I use a very conservative value of $K = 500$ to avoid false negative).

4.2.2 Tracking Model

The task is to track an object in an image sequence of length T frames. An object track is a set of T object locations $Y = \{y_t\}_{t=1...T}$. With each y_t is associated x_t, our object representation based on HOG and color histogram. The set of all x_t is denoted as X.

A track is initialized by bounding box annotations l_i made by the user in a set of *keyframes*. Note that the user could select only a single keyframe. The annotations are represented by their locations $L = \{l_i\}_{i \in N}$, with $1 \leq i \leq T$ and $|N| \leq T$. Under this model, a tracking algorithm can be described as a method that takes L as an input and returns Y, the trajectory of the object for the entire image sequence.

Given the description above, we now define the cost function that serves as a measure of the track quality:

$$E(Y; w) = \sum_t e(y_t; w) \tag{6}$$

$$e(y_t; w) = [w_1 w_2 w_3] \begin{bmatrix} d(x_t) \\ s_{app}(x_t, x_{t-1}) \\ s_{mot}(y_t, y_{t-1}) \end{bmatrix} \tag{7}$$

where $d(\cdot)$ is the cost of deviating from the global appearance model of the object (we use the SVM score), $s_{app}(\cdot)$ is the appearance smoothness cost, and $s_{mot}(\cdot)$ is the cost of deviating from the location predicted by optical flow. The contribution of $d(\cdot)$, $s_{app}(\cdot)$, and $s_{mot}(\cdot)$ to the overall cost is described by the parameters of the cost function: $w = [w_1, w_2, w_3]$. Note that the value of these parameters significantly impacts the tracking performance for a given video (see Section 4.3).

In this formulation, the tracking problem is reduced to finding the trajectory Y that minimizes the cost function $E(Y; w)$. In addition, we also have to ensure that the hard constraints of $y_i = l_i$ for all $i \in N$ are satisfied. In order to be robust to occlusion, we augment Y with an occlusion flag to reduce the penalty when an object undergoes occlusion.

4.2.3 Tracking Optimization

The task is to find the best track Y that minimizes the cost function described in Equation 7 subject to the constraints $y_i = l_i$ for all $i \in N$. If we assume there are K candidate locations for the object in each frame, a naive approach to finding the best track would take $\mathcal{O}\binom{K}{T}$ time. Fortunately, this problem exhibits optimal substructure that lends itself to an efficient dynamic programming (DP) solution (interested reader can refer to previous work on DP [17, 24] for more details).

Let K_t be the set of object candidates at frame t. Let y_t^k be the k-th candidate location of the object at frame t. Let $C_t(y_t^k)$ be the cumulative cost of the track up until y_t^k, if y_t^k is picked as a part of the object track. We can compute $C_t(y_t^k)$ for all $t \in T, k \in K_t$ in $\mathcal{O}(TK^2)$ by using forward recursion:

$$\begin{aligned}
C_0(y_0^k) &= w_1 d(x_0^k) \\
C_t(y_t^k) &= w_1 d(x_t^k) + \min_{j \in K_{t-1}} P_{t-1}^j(y_t^k) \\
P_{t-1}^j(y_t^k) &= C_{t-1}(y_{t-1}^j) + w_2 s_{app}(x_t^k, x_{t-1}^j) \\
&\quad + w_3 s_{mot}(y_t^k, y_{t-1}^j)
\end{aligned} \tag{8}$$

To obtain the best track, we can store the pointer to the match in the previous frame (Eq. 9) and backtrack from the location with the lowest cost in the last frame in T.

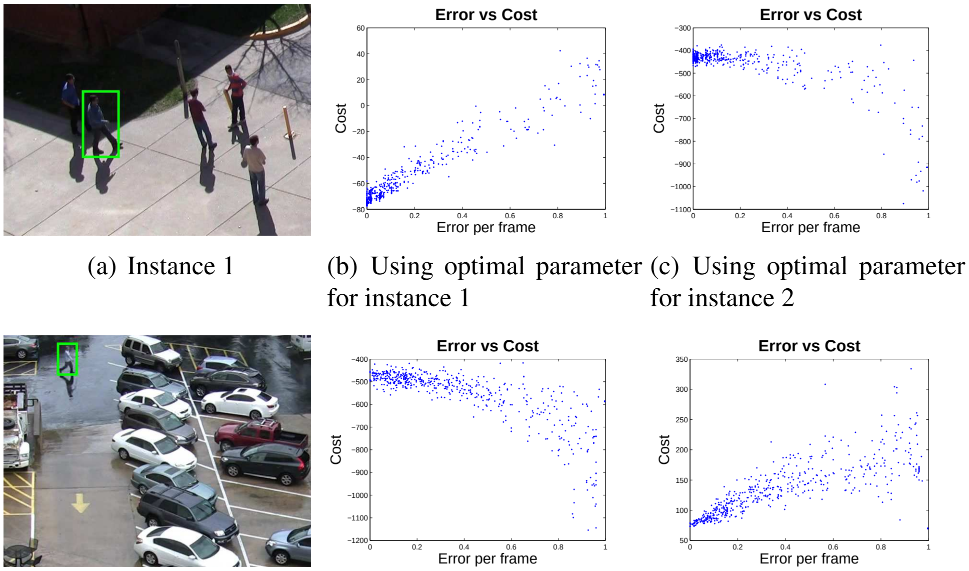
$$M_t^k(y_t^k) = \arg\min_{j \in K_{t-1}} P_{t-1}^j(y_t^k) \tag{9}$$

To ensure that the track satisfies the hard constraints $y_i = l_i$ for all $i \in N$, we simply set $d(x_t^k) = -\infty$ for all of the manually annotated locations l_i. Similar to Buchanan and Fitzgiboon [24], to account for occlusion we augment the set of object candidates in each frame with an occlusion state ($[y_t]_{occ} = 1$ means the object is occluded), effectively modifying the cost function into the following:

$$E(Y; w) = \sum_t \begin{cases} e(y_t; w) & [y_t]_{occ} = 0 \\ \lambda_o & [y_t]_{occ} = 1, [y_{t-1}]_{occ} = 0 \\ \lambda_r & [y_t]_{occ} = 1, [y_{t-1}]_{occ} = 1 \\ \lambda_v & [y_t]_{occ} = 0, [y_{t-1}]_{occ} = 1 \end{cases} \tag{10}$$

I set $\lambda_v = \lambda_o$, and $\lambda_r = 0.4\lambda_o$, so there is only one parameter to choose a value for.

This optimization method is very efficient. It takes less than 2 seconds to compute the globally optimal solution for $T = 1000$ and $K = 500$. That means that for every new annotation that a user has made, he/she can immediately observe how it affects the tracking result. This is a very desirable property for an interactive system. Note that this formulation has been been used in a number of interactive tracking work [24, 161, 152]. Thus, our approach to improve the cost function (Sec. 4.3) applies more broadly.



(a) Instance 1

(b) Using optimal parameter for instance 1

(c) Using optimal parameter for instance 2

(d) Instance 2

(e) Using optimal parameter for instance 1

(f) Using optimal parameter for instance 2

Figure 12: Error vs cost for two different sets of parameter values. I sample a number of trajectories that are close to the groundtruth, and plot the error for each of these trajectories under two different parameter settings. Note that the optimal parameter value for one instance can result in a bad model for the other instance in the sense that low cost is assigned to the trajectories that in fact have high error. In these scatter plots the ideal distribution is a line with a slope of 1, reflecting a cost function which matches the error relative to groundtruth.

4.3 Instance Specific Tracking Model

An important question that needs to be addressed is how do we *weight* the contributions of the different parts of the cost function. In other words, how do we select the appropriate values for $w = [w_1, w_2, w_3]$ in Equation 7? Currently, a popular solution for this parameter selection task is hand-tuning: the parameters that minimize the average training error are identified and used for all new testing videos. There are three problems with this approach: 1) There is no single value that is optimal for all of the possible testing videos. This is a major problem from the perspective of highly accurate tracking in which every video is important, as by minimizing the average error we accept the possibility of large error on specific video clips; 2) It can be very time consuming to exhaustively search for the best parameter value; and 3) Adding new components to the cost function requires substantial additional work. For example, if we want to incorporate an additional way to model global appearance into the cost function, we have to redo the parameter search step.

To illustrate the problem of using a single set of weights for all videos, consider two instances of a basic tracking task illustrated in Figure 12: tracking a person in the parking lot with other people around (instance 1) and without (instance 2). I sample a number of object trajectories that are close to the groundtruth trajectory and I compute the cost (according to (7)) for each of these trajectories with two different weight values, that correspond to the optimal weights for instances 1 and 2 (these values are computed by using our approach presented in Section 4.3.1). In Figure 12(e) we can see that an optimal set of weights for instance 1 results in a very bad model for instance 2 (and vice versa) where the trajectories that have more error actually have less cost. Note that even though in both instances we are tracking people, the *context* is different. In video 1 there are other objects with similar appearance in the scene (other people), in video 2 there are no objects present with similar appearance. Ideal weight parameters should reflect this difference in the nature of the tracking problem. Indeed, my approach is able to learn that very little weight should be assigned to the global appearance in instance 1 (since there are other people in the scene with very similar appearance) and instead the motion should be emphasized. In the subsequent sections I present my approach to incrementally learning the optimal value of the weight parameters *for each object trajectory in an interactive setting*.

4.3.1 Learning The Optimal Weights

In tracking optimization, the goal is to find a trajectory that has the lowest cost. The underlying assumption is that the groundtruth object trajectory has the lowest cost compared to all other possible trajectories. Therefore, by optimizing the cost function, we can obtain the groundtruth trajectory. Let Y^{gt} be the groundtruth trajectory. We can express this property mathematically as follow:

$$E(Y; w) > E(Y^{gt}; w) \qquad \forall Y \neq Y^{gt} \tag{11}$$

I have discussed in the previous section how the choice of w plays a critical role in determining the validity of the above assumption. If this assumption is violated, then optimizing the cost function is a fool's errand because it does not reflect the quality of the trajectory. In interactive tracking, this translates into the user having to provide substantial manual annotations to correct for tracking mistakes, which are inevitable since the costs are wrong. This is extremely wasteful given that a better choice of w could greatly alleviate this problem.

My goal is to find the optimal value for the weight parameter w for each tracking instance such that the groundtruth configuration has the lowest cost. The inequalities in (11) can have infinitely many solutions (e.g. a simple scaling of w will not change the inequality since the cost function is linear in w). A common trick to resolve this type of issue is to frame the problem as a maximum margin learning problem where the task is to find w that will maximize the margin between the groundtruth trajectory and all other trajectories:

$$\begin{gathered} \min \tfrac{1}{2}||w||^2 \\ E(Y; w) - E(Y^{gt}; w) \geq 1 \qquad \forall Y \neq Y^{gt} \end{gathered} \tag{12}$$

Due to the modeling limitation of the cost function and noise in the data, the above program may not have any solution. To address this issue we add the slack variables ξ_n. Thus we allow the individual margins to be smaller, but this is discouraged by adding the slack variables into the objective.

$$\begin{aligned} &\min \tfrac{1}{2}||w||^2 + \tfrac{C}{N}\sum_n \xi_n \\ E(Y;w) - E(Y^{gt};w) &\geq 1 - \xi_n \qquad \forall Y \neq Y^{gt} \end{aligned} \tag{13}$$

The program described above assigns unit margin to all of the trajectories that are not groundtruth (0-1 loss). While this should work well in an ideal scenario, if there is noise in the data the algorithm might produce suboptimal results since the optimization enforces the same margin on all of the trajectories (i.e. the same weight is assigned to all of the trajectories). The algorithm will be more likely to produce the desired result if we can instead use a better loss measure. This is the essence of the maximum margin learning approach to structured prediction [143], which we would adopt.

In tracking, we can measure loss by using the Hamming distance between a trajectory and the groundtruth trajectory $\Delta(Y, Y^{gt})$. In this sense, we can view the problem of learning the optimal weight parameter in tracking as an instance of maximum margin structured prediction learning. By using the Hamming distance as our loss measure, the constraints in (13) now become:

$$E(Y;w) - E(Y^{gt};w) \geq \Delta(Y, Y^{gt}) - \xi_n \quad \forall Y \neq Y^{gt} \tag{14}$$

The above constraint means that we desire larger margin for the trajectories that are further from the groundtruth. Or in other words, this loss-scaled margin means that the trajectories that are further from the groundtruth should have higher cost than the trajectories that are closer (smaller margin). This is certainly a very desirable property for a cost function. Unfortunately, it is not feasible to solve the above program due to two factors: 1) we do not know the groundtruth trajectory Y^{gt}; and 2) there are an exponential number of constraints (K^T assuming there are K object candidates in every frame in a T-frame long video sequence).

During the interactive tracking process, a user incrementally adds one annnotation at a time. As a result of this, we obtain a series of trajectory estimates $Y^1, Y^2, \ldots Y^N$ (assuming the user has made N annotations) where Y^{i+1} is likely to be closer to the groundtruth than Y^i. My insight is that we can exploit this process to incrementally learn w. So instead of using the groundtruth trajectory (which we do not have) as the positive instance for max margin learning, *we can use the current best estimate of the trajectory as the positive instance* and perform the optimization over a much smaller set of constraints that correspond to the other previously estimated trajectories that we have obtained during the interactive tracking process. So for every new annotation a user has made, we can estimate the parameter value that will make the most recent trajectory estimate have the lowest cost. This process is aligned with our original formulation where we desire parameters that will make the cost function assign lower cost to the trajectory that is closer to the groundtruth (i.e. the latest trajectory estimate Y^N) compared to the trajectories that are further from the groundtruth (i.e. other previously obtained trajectories $Y^1, Y^2, \ldots Y^{N-1}$). We can implement this as the following optimization:

$$\begin{gathered}\min \tfrac{1}{2}||w||^2 + \tfrac{C}{N}\sum_{i=1}^{N}\xi_i \\ E(Y^i; w) - E(Y^N; w) \geq \Delta(Y^i, Y^N) - \xi_i \quad {\scriptstyle i=1\ldots N-1} \\ w_j \geq 0 \qquad \forall w_j \in w\end{gathered} \tag{15}$$

By solving the above program after every annotation, we are guaranteed to have w that assigns the lowest cost to the latest trajectory estimate (within some slack tolerance). Note that if the user annotated the whole video sequence ($N = T$), the above program reduces to the original formulation in Equation 14, but with a much smaller set of constraints.

To account for the fact that we now optimize over a significantly smaller set of constraints compared to the original formulation in (14), we add an additional set of constraints to enforce every single element of w to be nonnegative. This is a subtle but important addition since this set of constraints serve as a way to represent the trajectories that are far from the groundtruth in the optimization. Many of the high loss trajectories will have high values of $d(\cdot)$, $s_{app}(\cdot)$ or $s_{mot}(\cdot)$. Consider for example a trajectory that jumps from one corner of the image to a different corner in successive frames. This trajectory will have a very high $s_{mot}(\cdot)$ (similar examples can be drawn for the other two components of the cost function, $d(\cdot)$ and $s_{app}(\cdot)$). Since our constraint set consists of only trajectories that are close to the groundtruth, it will most likely not contain examples of those high-loss trajectories. Because of this, there is a possibility that we obtain a negative w which can result in the high-loss trajectories (which are not represented in the constraint set) to obtain the lowest cost. Adding the nonnegativity constraint for w alleviates this problem.

To illustrate the result of our incremental learning of w, let's revisit our earlier example of tracking a person in the presence of other people (Fig. 12(a)). Due to the existence of similar looking objects in the scene (other people), we know that intuitively the global appearance component should carry less weight in the overall cost function. Our incremental weight learning approach is able to quickly learn this context information (see Table 6). Also note how given the same set of annotations, the w that we learn incrementally results in a better cost function for the problem (which is reflected by the lower error rate).

Table 6: Incremental learning of w. This table illustrates the effect of my incremental learning of the cost function parameters. I annotate a 300-frame long sequence at 4 uniformly-spaced locations, and I perform trajectory estimation given those annotations with 4 different w values (the starting w and w that is learned incrementally after annotations 2, 3 and 4). Note that my approach is able to learn to place less and less weight on the global appearance cost (w_1) since there are many similar-looking objects in the scene (Fig. 12(a)).

N **annotations**	w_1	w_2	w_3	**Error/frame**
1	0.33	0.33	0.33	0.5100
2	0.18	0.46	0.36	0.3800
3	0.08	0.40	0.52	0.0733
4	0.03	0.37	0.60	0.0367

4.3.2 Improving The Objective

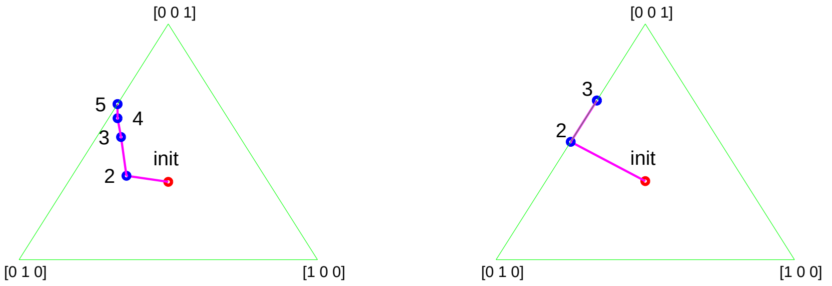
A potential problem with the loss-scaled constraint in Equation 15 is that the algorithm may give a suboptimal solution since it focuses on the constraints with high loss. Since we scale the margin by the loss, a w that gives Y^N the lowest energy (which is our goal) may not be selected in the optimization if there are any high-loss constraints that do not have a large enough margin. This means that the earlier trajectory estimates (which are the constraints that have high loss) can potentially overwhelm the ultimate objective which is finding a w that gives the most recent trajectory estimate Y^N the lowest cost. In order to compensate for this, we can add directly to the objective the difference between the cost of the two latest trajectory estimates, given a w parameter ($E(Y^N; w) - E(Y^{N-1}; w)$). This can be interpreted as putting more emphasis for the algorithm to search for the solution that maximizes the separation between the two data points that are closest to the decision boundary. This acts as a counter-weight to the high loss constraints. The final objective then becomes the following:

$$\min \frac{1}{2}||w||^2 + \frac{C_1}{N}\sum_{i=1}^{N}\xi_i + C_2(E(Y^N; w) - E(Y^{N-1}; w)) \tag{16}$$

This formulation is similar to Szummer et.al. [141] and Tsochantaridis et.al. [148] but is adapted to my sequential formulation.

To illustrate the effect of the new objective on the parameter learning process, let us consider once more the interactive tracking task in Figure 12(a) (tracking a person in the presence of other people). We start with $w = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$ and perform interactive tracking on the sequence by doing annotation one frame at a time. We use the same annotation schedule (same set of frames with the same annotation ordering) and compare the convergence behavior of the two objectives. Starting from the initial annotation, after each subsequent annotation we compute the optimal w according to the two objectives. We normalize the w to sum to 1 and plot its value on a simplex (note that normalizing w does not change the inequality constraints in (15)). Figure 13 illustrates the convergence behavior of the two objectives.

Notice that even though both objectives essentially converged to the same value (both learned to place no weight on the global appearance due to the presence of similar looking objects in the scene), the improved objective found the optimal parameter value much more quickly than the original objective, converging after only 3 annotations instead of 5. My



(a) Original objective

(b) Improved objective

Figure 13: Convergence behavior of the original objective (15) and the improved objective (16) on the tracking instance in Fig. 12(a) (tracking an object in the presence of similar looking objects). The green simplex is the solution space. Red dot is the starting value of w and the blue dots are the value of w after each annotation. Note how the improved objective converged quicker to the optimal solution.

hypothesis is that the additional term in the objective allows the algorithm to quickly converge to the optimal solution by admitting a solution that does not provide enough margin to the high loss constraint (in this case, constraint induced by the first trajectory estimate Y^1). I look at the value of the slack variable ξ_1 after annotation 3 to confirm this hypothesis, and indeed the value of this variable in the new objective is higher than that in the original objective. This confirms my idea that the additional term in the new objective can serve as a balancing term to the high loss constraints, allowing the algorithm to focus more on the solution that maximizes the separation between data points that are closest to the decision boundary.

4.4 Experiments

To demonstrate the advantage of my instance specific max-margin tracking parameter learning approach, I perform experiments on two datasets: 1) the VIRAT challenge dataset [109]; and 2) the Infant-Mother Interaction dataset that I introduce. The VIRAT dataset consists of over 300 surveillance videos captured in various locations. My task in this dataset is to track *moving* people and cars (in VIRAT there are a lot of objects that are

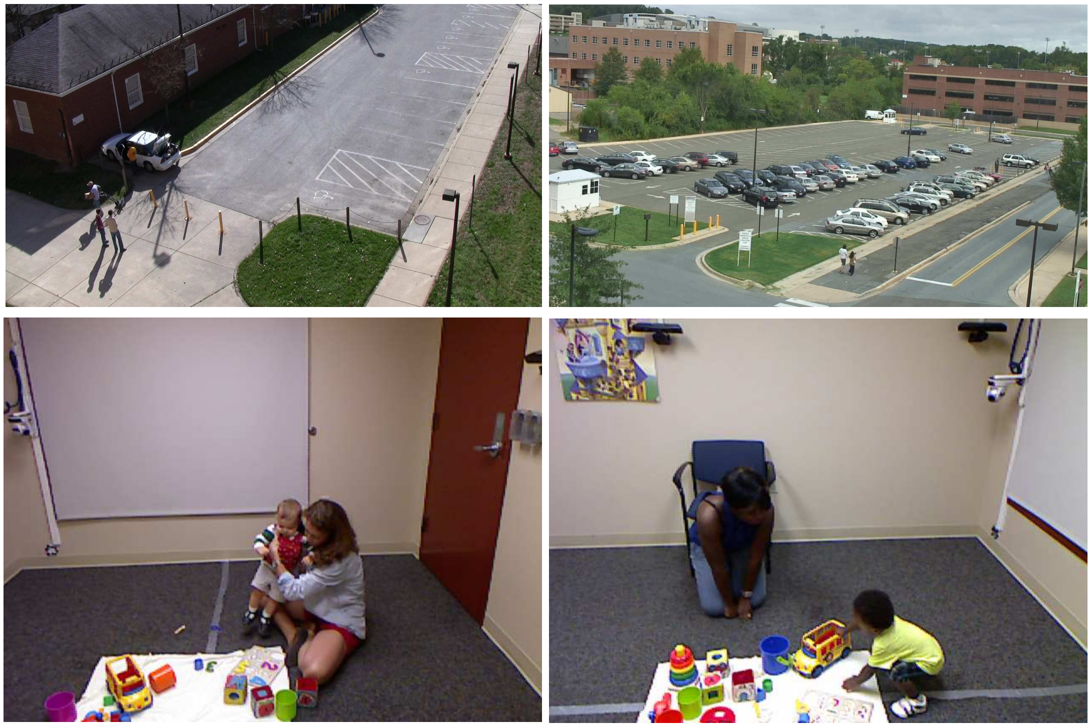


Figure 14: Dataset used for experiments: VIRAT dataset (top row) and Infant-Mother Interaction dataset (bottom row).

completely stationary, which is trivial to track). The Infant-Mother Interaction dataset consists of 15 videos of a dyadic interaction between an infant and a mother in a room during a behavioral study called The Strange Situation [9]. This dataset serves as an important practical application for interactive tracking since being able to obtain high precision track of the people in the scene has a tremendous amount of utility for quantifying the behavior of the individuals in the study (see Chapter 5). The task in this dataset is to track the head of the people in the scene. A representative set of frames from the two datasets can be seen in Figure 14. Note that groundtruth bounding box annotations of object tracks are provided in both datasets.

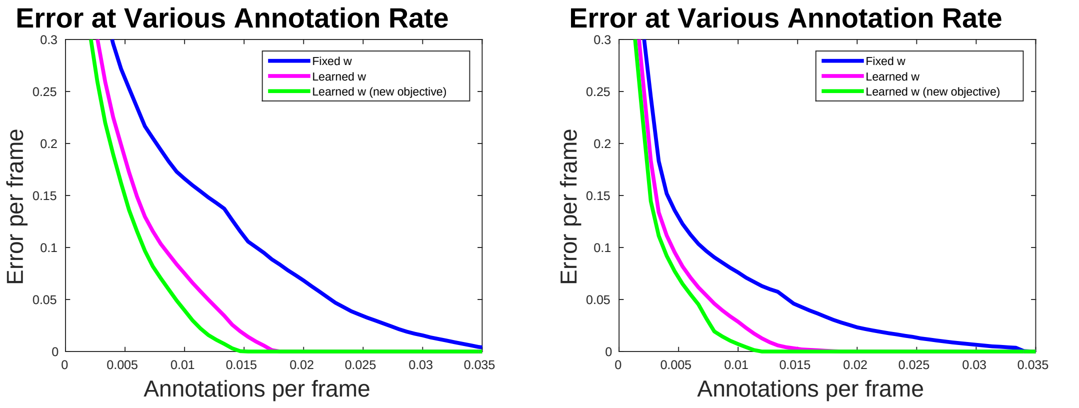
I compare my incremental weight learning approach against the traditional fixed-weight approach (hand-tuned to each of the datasets). I measure tracking error based on how well the tracker is able to estimate the groundtruth annotations. For every frame, an object track is considered to be correct if its intersection over union (IoU) with the groundtruth is at least 0.3 (a similar metric is used in [152]). I quantify tracking error by the error-per-frame metric (i.e. an error of 0.01 means that for every 100 frames there is 1 frame where the IoU is less than 0.3). To quantify human effort, I use the annotations-per-frame metric (an annotations-per-frame of 0.1 means that a user annotated 10 frames out of 100). For an interactive tracking system, the goal is to obtain high precision tracking results with *as few annotations* as possible. To capture this, for each experiment I report the number of annotations-per-frame that is required from the user to achieve a certain error-per-frame target.

The interactive tracking experiments are conducted assuming that an oracle exists for the system to request annotations from. In other words, user annotations are simulated by the oracle (which uses the provided groundtruth). I perform the experiments using two different frame selection strategies. The first one is the sequential strategy where the interactive tracking system requests for a bounding box annotation from the oracle whenever the object track starts to drift from the groundtruth (IoU < 0.3). This is to simulate the use case when a user first annotated the target object in the first frame, and continue watching the video (possibly in super-realtime speed) to correct for tracking mistakes as they happen. The second strategy is the uniform frame selection. This is a frame selection strategy commonly employed in a setting where multiple users are available to provide annotations (e.g., crowdsourcing [151]). For this use case, a system simply requests for annotations for the target object on a set of keyframes that are uniformly distributed (temporally) on the video.

The results for the VIRAT dataset using the sequential frame selection strategy can be seen in Figure 15(a). My approach is able to outperform the fixed weight approach by a large margin. For example, on average, by learning the weight parameter during the annotation process using our method, my method is able to achieve 0.01 error tracking results using only 0.0133 annotations-per-frame, compared to the 0.0327 annotations-per-frame that is required by the fixed weight approach. This is an improvement of $\sim 60\%$

which means that by using our approach, I can annotate this dataset to the same desired accuracy with only 40% of the effort. This can potentially translate to a saving in the order of tens of thousands of dollars for a dataset this size. Also note that the improved objective that I propose gives a considerable improvement over the standard maximum margin objective.

Similar to the VIRAT dataset, my approach is able to significantly improve the annotation efficiency in the Infant-Mother Interaction dataset (see Figure 15(b)). For the target error rate of 0.01, my approach is able to achieve the same tracking accuracy with only 36.6% of the human effort (going from 0.0273 annotations-per-frame to 0.01). Note that the Infant-Mother Interaction dataset represents the ideal dataset for the hand-tuned fixed weight approach since on the surface there seems to be minimal variations in the scene (there is only one type of target object and all of the videos are captured in the same room). However, even in this setup my approach is still able to provide a large improvement. This means that even on videos captured from similar scene with the same type of target object, there is always a significant variability in the individual tracking instances. Note that as is the case in VIRAT, the proposed new objective gives the best results.

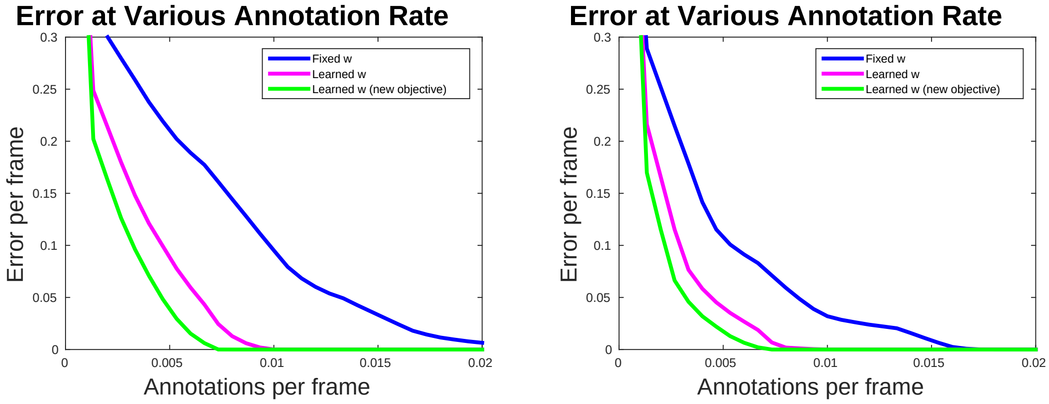


(a) Results on VIRAT dataset

(b) Results on Infant-Mother Interaction dataset

Figure 15: Results on VIRAT and the Infant-Mother Interaction dataset using sequential annotation procedure. y-axis is the error rate, x-axis is the annotations rate.

Similar to the sequential frame selection strategy, my approach gives a significant improvement using the uniform frame selection strategy (Figure 16). In the VIRAT dataset, my approach uses 64.3% fewer annotations compared to the baseline to achieve 0.01 error-per-frame (going from 0.0187 annotations-per-frame to 0.0067). Similar improvement is also observed in the Infant-Mother Interaction dataset (see Figure 16(b)).



(a) Results on VIRAT dataset

(b) Results on Infant-Mother Interaction dataset

Figure 16: Results on VIRAT and the Infant-Mother Interaction dataset using uniform annotation procedure. y-axis is the error rate, x-axis is the annotations rate.

4.5 Conclusion

I have highlighted the importance of having instance-specific model parameters in the tracking by detection framework. I have presented a novel approach to address this critical problem of determining the parameter value of the cost function. I leverage the sequential nature of interactive tracking to formulate an efficient approach for learning instance-specific model parameters through a maximum margin framework. I have demonstrated that by using my approach we can save the required number of annotations by $\sim 60\%$ to achieve high precision tracking results, a significant improvement in efficiency compared to the existing approach.

CHAPTER V

OBJECTIVE ANALYSIS OF INTERACTION BY DENSE MEASURE OF PROXIMITY

5.1 Introduction

The measure of proximity (physical distance between two objects) is central to many studies in psychology. For example, in studying how people develop a sense of personal space, having a measure of physical distance between people is crucial [137]. Similarly, studies in infant locomotion use a continous measure of infant location with respect to the target goal as one of the primary features [5]. Physical distance between an infant and an object is also used as a measure when studying approach behavior [125], which is one of the dimensions of temperament (aspects of an individual's personality that are often regarded as innate rather than learned). In the attachment domain, proximity-seeking and exploratory behaviors are some of the important characteristics when looking at infant-mother interaction [9].

Currently, it is very difficult to obtain accurate, dense measure of proximity. In an early personal space study, Sommer [137] used the arrangement of chairs around a dining table as a way to measure proximity. Adolph et.al. [5] instrumented the space with physical marker (grids) so that a human observer can measure the location of infant in a video recording during infant locomotion study. Rothbart [125] measured infant approach behavior in a highly restrictive tabletop setting for studying temperament. In an early study of infant attachment, Ainsworth and Bell [8] also used grid-lines as a way to measure infant exploration behavior. The need to use actual physical markers (e.g. grids, arrangement of objects) to get a proximity measure is certainly not desirable. Not only it limits the experiment setup (e.g. the grids have to be always visible to the human observer to make it useful), but also it only gives us a very coarse measure of proximity, both spatially (the maximum accuracy is limited by the size of the grids) and temporally (having a human observer measures proximity for every frame in the video is impractical, thus it is typically done in a very coarse time-step, e.g. 15 seconds in the study done by Ainsworth and Bell [8]). All of those examples clearly suggest that a better way to measure proximity will have a big impact on a wide array of studies.

While it is extremely challenging for humans to estimate accurate proximity measure between objects just by looking at a video, recent development in affordable depth camera (e.g. Microsoft Kinect) has made it possible for us to obtain this measure. With depth sensor, the task of continously measuring the 3D spatial location of an object can be reduced to the video object tracking problem (i.e. we only have to localize where the object is on the image plane in order to determine its 3D location in the world). In this chapter, I address this issue of obtaining accurate, temporally dense measure of proximity by combining state-of-the-art interactive tracking technique with modern sensing technology (depth camera).

The focus of this work is on demonstrating how we can use dense measure of proximity to objectively analyze infant-mother interaction during The Strange Situation, a protocol for studying infant attachment security. In The Strange Situation, an infant is put through a series of interaction episodes involving a caregiver (typically the mother) and a stranger where the goal is to activate the attachment system. The idea is that different behaviors will be exhibited by an infant depending on where he/she is on the attachment spectrum. An expert then review the interaction through a video for these attachment-related behaviors and and classify the infant attachment into one of three categories: secure, insecure-avoidant and insecure-resistant.

The Strange Situation is an excellent test scenario for showcasing the power of having dense measure of proximity since it is designed based on the theory that infant attachment security will affect the balance between attachment behaviors (wanting to be close to the caregiver) and exploration behaviors (wanting to explore the environment) when the attachment system is stressed. It is clear that a measure of infant-caregiver proximity directly relates to the attachment and exploration behavior. However, currently an objective measure of proximity is not being taken into account by experts when producing the final attachment classification even though early study of the Strange Situation incorporated a coarse measure of proximity [8]. One reason for this might be due to the difficulty in obtaining accurate, continous measure of proximity from human observation. Instead, current practice of assigning attachment classification for The Strange Situation involves a complex guide based on qualitative behavioral observation. The main problem with this approach is that training a human to be able to code for these behaviors reliably is both time consuming and expensive (e.g. The University of Minnesota Strange Situation training takes at least a full week [2]).

In this chapter I will demonstrate that dense measure of infant-mother proximity can be used to predict expert's infant attachment classification in the Strange Situation. This finding can be seen as the first step towards an objective, quantitative way to analyze social interactions. In addition, another benefit of having a low-level quantitative measure is that the data can be easily teased apart in some new way, allowing for a convenient way to test for new hypothesis. Towards that end, in this chapter I will also demonstrate how we can use the proximity measure to test for new hypothesis on the Strange Situation.

5.2 Dataset

I collaborated with The Early Play and Development Laboratory at the University of Miami to collect several recordings of The Strange Situation. The procedure is executed in a 9×9 square-foot room. Within the room, there are two chairs for the adults (the stranger and the caregiver) to sit on, and also a set of toys for the infant to play with. The room is instrumented with 4 Kinect cameras (see Figure 17 to capture RGBD videos of the interaction. The videos were recorded with spatial resolution of 640×480 pixels and temporal resolution of 25 frames-per-second. The dataset comprised of 34 fully recorded, and 1 partially recorded session of the Strange Situation procedure from 35 different infants (each infant only did the procedure once).

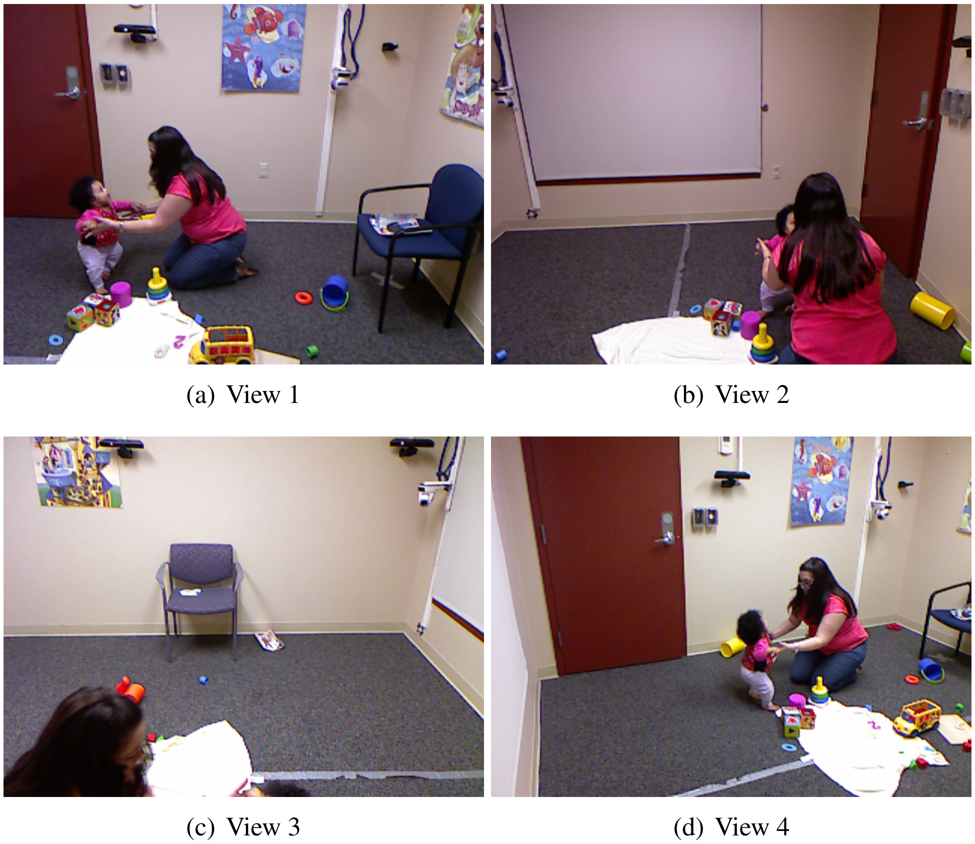

(a) View 1
(b) View 2
(c) View 3
(d) View 4

Figure 17: Four Kinects capturing The Strange Situation.

5.2.1 The Strange Situation Procedure

Before moving forward, I will now briefly explain the procedure of The Strange Situation [8]. The Strange Situation is designed to elicit a wide range of behaviors pertinent to attachment and exploration by bringing an infant through a situation of novelty and alarm in a laboratory setting. There are 4 individuals involved in this procedure: the mother (M), the baby (B), the stranger (S) and an observer (O). The interactions in The Strange Situation comprise of eight standard episodes (same order is followed for all subjects). The procedure is designed in such a way that it is novel enough to elicit exploratory behavior, but at the same time not so strange that it causes a high degree of fear to the infant. Following is the

8 episodes in The Strange Situation (visual illustration can be seen in Figure 18):

1. Episode 1 (M, B, O). Mother accompanied by an observer carried the baby into the room, then the observer left.

2. Episode 2 (M, B). Mother put the baby down in the specified place, then sat quietly in her chair, participating only if baby sought her attention. Duration 3 minutes.

3. Episode 3 (S, M, B). A stranger entered, sat quietly for 1 minute, conversed with the mother for 1 minute, and then gradually approached baby, showing him a toy. At the end of the third minute, the mother left the room unobtrusively.

4. Episode 4 (S, B). If baby was engaged in play, the stranger was nonparticipant. If the baby was inactive, the stranger tried to interest him in the toys. If the baby was distressed, the stranger tried to distract him or to comfort him. If the baby could not be comforted, the episode was curtailed — otherwise it lasted 3 minutes.

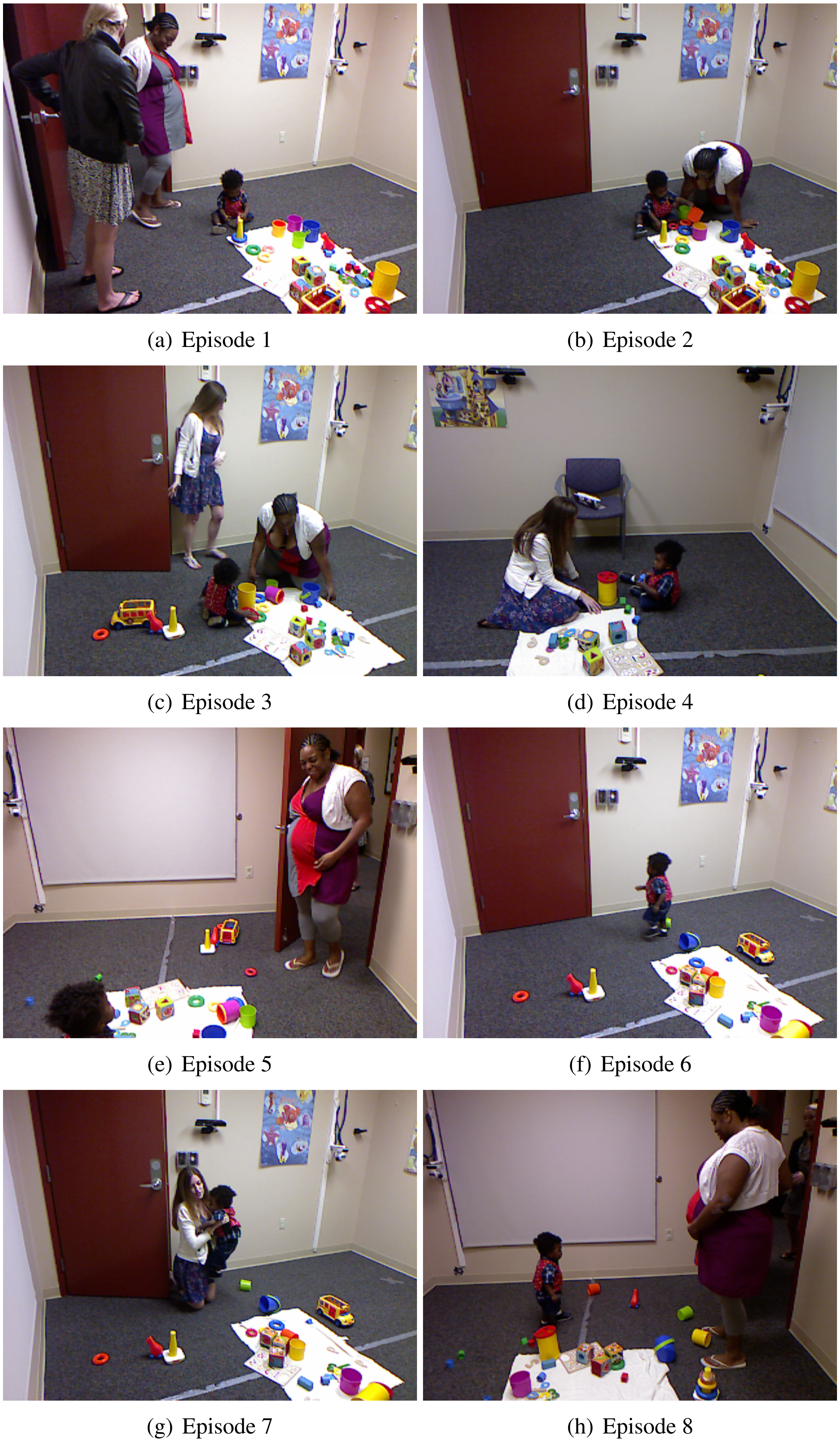
5. Episode 5 (M, B). Mother entered, paused in the doorway to give the baby an opportunity to mobilize a spontaneous response to her. The stranger then left unobtrusively. What M did next was not specified — except that she was told that after baby was again settled in play with the toys, the mother was to leave again.

6. Episode 6 (B alone). The baby was left alone for 3 minutes, unless he was so distressed that the episode had to be curtailed.

7. Episode 7 (S, B). The stranger entered and behaved as in episode 4 for 3 minutes.

8. Episode 8 (M, B). The mother returned, the stranger left, and after the reunion had been observed, the situation was terminated.

These episodes were arranged so that the less disturbing ones (from the infant's perspective) came first. As a whole, the procedure is intended to be no more disturbing than what an infant might encounter daily. The goal of The Strange Situation is to activate the



(a) Episode 1

(b) Episode 2

(c) Episode 3

(d) Episode 4

(e) Episode 5

(f) Episode 6

(g) Episode 7

(h) Episode 8

Figure 18: The eight episodes in The Strange Situation.

attachment system by putting the baby through a period of separation from the mother (episode 4, 6 and 7). The basic idea is that when the baby feels stressed, his attachment behavior is likely to be activated. As the mother comes back into the room (episode 5 and 8), the baby then will elicit certain behaviors to gain back a sense of security which then allows the exploratory system to be activated by the novel environment. In the attachment literature, this phenomenon is termed ”using the mother as a secure base from which to explore.” As such, when classifying The Strange Situation, an expert typically focus more on the two reunion episodes (episode 5 and 8). Similarly, in this work I focus on analyzing the infant-mother interactions during the two reunion episodes through the lens of a dense proximity measure.

5.2.2 Expert Ratings in The Strange Situation

After the procedure has been run, an expert then performs a detailed coding of the infant behaviors for each of the two reunion episodes by referring to the video recordings. There are four classes of behaviors scored by the expert on The Strange Situation: proximity-seeking, contact-maintenance, interaction-avoiding, and contact-resisting. The coding is done on a 7-point scale based on the assumption that the infant might exhibit the each behaviors with different intensities. Following is a brief description of the four classes of behaviors:

1. *Proximity-seeking.* The intensity, promptness and persistence of the babys efforts to gain contact to caregiver.

2. *Contact-maintenance.* Degree of activity and persistence in baby’s efforts to maintain contact with mother once contact has been gained.

3. *Interaction-avoiding (avoidant).* Intensity, persistence, duration and promptness of the baby’s avoidance of proximity and interaction, even across a distance.

4. *Interaction-resisting (resistant).* Intensity and frequency or duration of resistant behavior during contact.

The behaviors above are considered as the ingredients of the infant attachment in the Strange Situation. There are a lot of factors that an expert has to consider when assigning a rating for these behaviors during each of the reunion episodes. Also, the difference between the ratings are sometimes subtle and hard to define precisely. For example, differentiating a 7 and a 6 in the rating scale often involves determining whether a particular action by the infant is deemed as *very* active as opposed to just active. This makes assigning a rating reliably to these behaviors difficult. In fact, a coder will have to go through at least a week-long training to become an expert [2]. In this work, I hypothesize that by looking at infant-mother proximity, a purely objective quantity derived directly from video data, we can accurately replicate the expert rating for these attachment behaviors. I will describe how this can be done in Section 5.4.

After performing a detailed coding of the attachment behaviors, an expert then infers the attachment security classification of the infant. The infant attachment can be classified into three groups: A (insecure-avoidant), B (secure) and C (insecure-resistant). Note that as described in Ainsworth's Patterns of Attachment [9] and Fraley and Spieker [49], there is no straight mapping between the rating of the attachment behaviors and the attachment classification. Rather, these behaviors only serve as a guide for the expert to make the final judgment on the classification. The basic idea is that coding these behaviors informs the expert for making the right classification. Following are some excerpts from the criteria for attachment classification as outlined in Ainsworth's Patterns of Attachment:

- Group A. ”Conspicuous avoidance of proximity to or interaction with the mother in the reunion episodes. If there is approach, the baby tends to mingle his welcome with avoidance responses–turning away, moving past. Little or no tendency to seek proximity to or interaction or contact with the mother.”

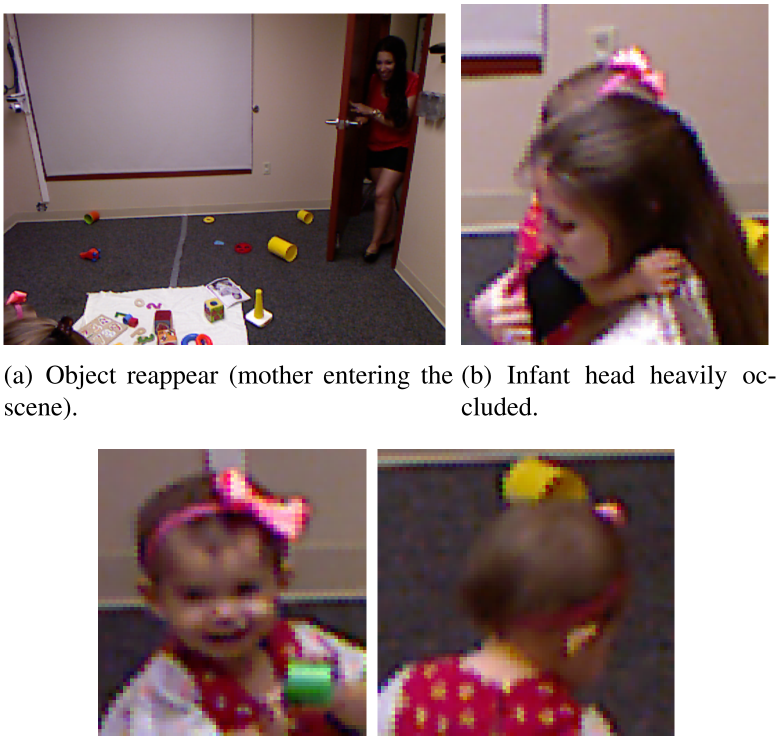
- Group B. ”The baby wants either proximity and contact with his mother or interaction with her, and he actively seeks it. The baby responds to his mother’s return in the reunion episodes with a tendency to approach. Little or no tendency to avoid his mother in the reunion episodes.”
- Group C. ”The baby displays conspicuous contact- and interaction-resisting behavior. He also shows moderate-to-strong seeking of proximity and contact and seeking to maintain contact once gained.”

Assigning an attachment classification involves analyzing a lot of fine details on the behaviors during the reunion episodes. Therefore, producing a reliable classification requires a lot of training and experience. It is not uncommon for an expert to ask for a second opinion when assigning a classification. Looking at the description of the different attachment groups, we certainly get a sense that infant-mother proximity plays a big role in discriminating the groups. Similar to previously, I hypothesize that we can use infant-mother proximity during the reunion episodes, a fully objective measure obtained directly from video data, to produce attachment classification in the Strange Situation (Section 5.4).

5.3 Extracting Dense Measure of Infant-Mother Proximity from Kinect Recordings

I use the distance between the baby’s and the mother’s head as the measure of infant-mother proximity. There are two reasons for choosing head as the target: 1) It is is easier to track compared to other body parts. The shape of the head is always approximately ellipsoid independent of the viewpoint of the camera. This facilitates easy detection and tracking; 2) Knowing the 3D location of the head allows an easy way to identify certain events such as discriminating between whether an infant is crawling, walking or being carried by the mother (this can be done easily by looking at both the height of the head with respect to the floor and the proximity between infant’s and mother’s head.

The first step that I did to obtain a temporally-dense measure of infant-mother proximity from the multi-view Kinect recordings is to track the head of the infant and the mother in the videos. The 2D head tracking result will later be used as a way to determine the 3D position of the heads, so the tracking needs to be done with high precision. Performing high precision tracking fully-automatically is not possible with the current state of computer vision technology due to the challenging nature of the tracking problem in this setup: 1) The head can disappear/reappear at any point in the recordings (Figure 19(a)); 2) The head can be heavily occluded in the video (Figure 19(b)); and 3) There can be a wide variation in the appearance of the head at various time points in the video (Figure 19(c)). To obtain high precision tracking of the heads, I use the interactive tracking framework that I have developed in Chapter 4. An example of the tracking result can be seen in Figure 5.3.



(a) Object reappear (mother entering the scene).

(b) Infant head heavily occluded.

(c) Wide variation in the appearance of the head.

Figure 19: Challenges in tracking the head of the infant and the mother in The Strange Situation.

In order to extract the 3D head location from the 2D tracking result, we first need to map the point cloud recorded from the four Kinect cameras to a single global coordinate system.

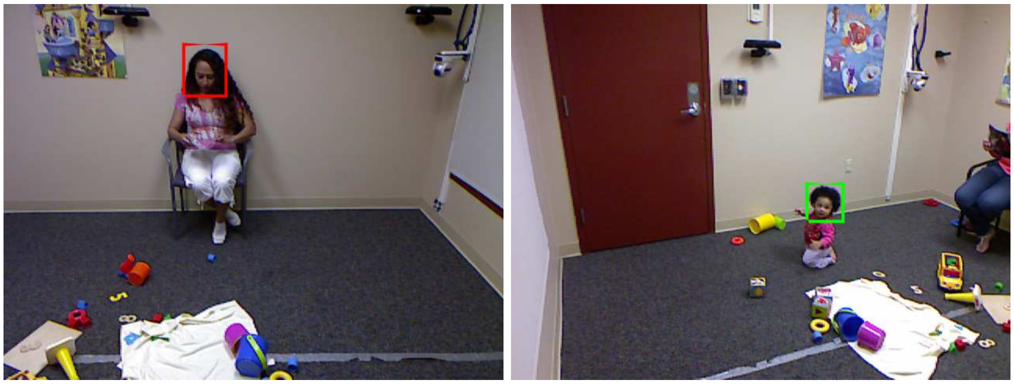


Figure 20: Interactive tracking result.

To do this, I do a one-time calibration to compute the 3D rigid transformation (6 degrees of freedom: 3D rotation + translation) between the coordinate system of the different Kinects. The first step of this calibration process is to identify a number of common physical points in the image captured from the 4 Kinect cameras (a minimum of 2 points are required to compute the 3D rigid transformation). To be robust to the noise in the Kinect point cloud data, the transformation between the different Kinects is computed by using at least 24 point correspondences that are identified manually. Using the point correspondences, the 3D rigid transformation is computed by using the Singular Value Decomposition (SVD) method [11] which has been shown to be more stable than the other methods such as the Orthonormal Matrices (OM) and the Dual Quaternions (DQ) [41]. An example of a fused point cloud data from the multiple Kinects can be seen in Figure 21.

Once we have mapped the data from the multiple Kinects into a common frame of reference, we can automatically track the infant's and the mother's head location in 3D by making use of the video head tracking result that have been obtained previously. The steps to track the head in 3D is as follow:

1. Get the 3D template of the head. In this work, I model the heads as spheres with circumference of 45cm for the baby and 55cm for the mother (those two numbers represent the $50th$ percentile of an infant and a female head circumference in the population [48, 25]).

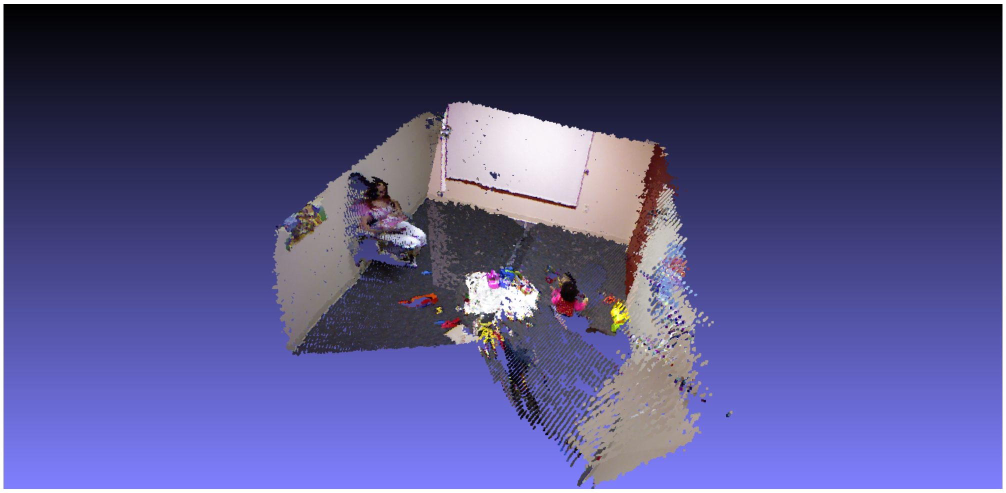


Figure 21: Fused point cloud data from multiple Kinects.

2. Compute the initial guess for the location of the head in 3D by fitting the head template to the set of point clouds that fall within the bounding box obtained from the 2D tracking result. To do the template fitting, I use the Iterative Closest Point (ICP) with robust outlier rejection [176] to make the fitting process robust to background noise and partial occlusion of the head.

3. Compute final estimate of the head location in 3D by using Kalman filter [77], treating the location of the head obtained in the previous step as the observation of the system. Note that the estimate for the 3D head location computed form the previous step can be noisy due to factors such as heavy occlusions. A known technique in tracking to address this problem is to use Kalman filter as a way to account for noise in the observation and smooth the tracking result. In this work, I use the standard linear Kalman filter with constant velocity assumption which has been shown to work well for tracking people [19]. An example of the 3D tracking result can be seen in Figure 22.

Finally, to compute the dense (25 frames-per-second) proximity measure between the infant and the mother, I simply compute the distance between the 3D location of the heads. An example of the proximity plot during a single reunion episode can be seen in Figure

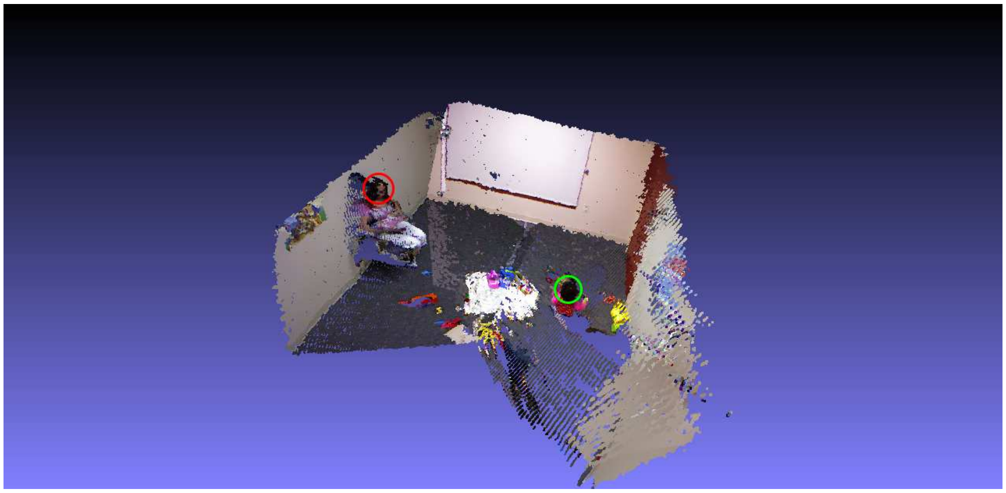


Figure 22: 3D tracking result (red circle: mother's head, green circle: baby's head).

23. In summary, I use the following four steps to obtain a dense measure of infant-mother proximity from the multi-view Kinect recordings:

1. Track the infant and the mother's head in image plane (Figure 5.3).
2. Map the point cloud obtained from the different Kinects to a single global coordinate system (Figure 21).
3. Track the heads in 3D by making use of the result of the previous two steps (Figure 22).
4. Compute the distance between the infant's and the mother's head in 3D (Figure 23).

5.4 The Strange Situation Through The Lens of Infant-Mother Proximity

The goal of The Strange Situation is to activate the baby attachment system to a high level through a period of separation with the mother to elicit a wide range of attachment and exploratory related behaviors. The theory that underlies this procedure is that the state of being attached, together with the presence of the attachment object (i.e. the caregiver), may support and facilitate exploratory behavior. Therefore, by observing the infant's behaviors

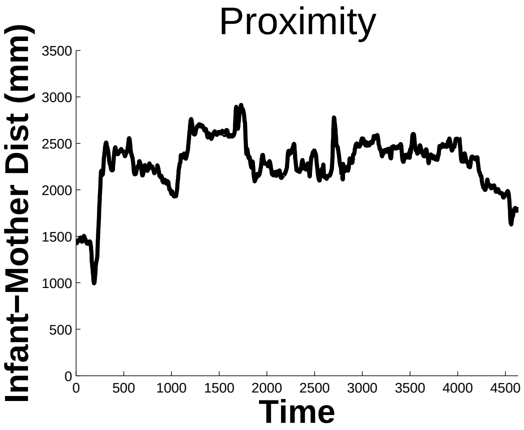


Figure 23: Infant-mother proximity during a reunion episode.

during the reunion episodes (a period where the infant attachment system is likely to be activated), an expert then can look for patterns in the attachment behaviors and classify the baby attachment into three groups: A (insecure-avoidant), B (secure) and C (insecure-resistant).

In this work I posit a new way to analyze attachment behaviors by using infant-mother proximity, an objective measure derived directly from video data. To validate this new measure, I will show that we can use it to predict both the expert's ratings for the 4 attachment behaviors (proximity-seeking, contact-maintenance, resistant and avoidant) and the expert attachment classification. Though simple, infant-mother proximity is a very rich measure. It can tell us periods when the baby is in close contact with the mother, moments when the baby is actively approaching the mother, the baby's locomotion when the mother is approaching him, and many other behavioral events that are relevant for analyzing infant-mother attachment. To use this measure for predicting the rating of attachment behaviors and attachment classification, I first need to reduce it into a set of features. This process is akin to designing features for audio classification (deriving features from a 1d signal). The features should be designed in such a way that they capture many of the attachment-relevant behaviors exhibited by the infants during the Strange Situation. Based on the description of both the attachment classification and attachment behaviors in Ainsworth's Patterns of

Attachment, I formulate the following feature set:

1. Contact time: The length of time (in seconds) that the baby is in close contact to the mother. I define contact as whenever the baby's head is within an arm's length (80cm) of the mother's head. This feature will speak to whether contact is achieved during the reunion, and for how long. We expect that baby in the A group will have a low or zero contact with the mother, baby in the B group will have medium to high amount of contact, and baby in the C group will have a high amount of contact due to the unsuccessful attempt by the mother to sooth the baby.

2. Proximity change: Change in infant-mother proximity (in meters) that is caused by the baby's locomotion during the first 5 seconds of the reunion episode. A securely attached baby is more likely to respond to the mother's return by actively approaching the mother. The intensity and persistence of the approach will be captured by this feature.

3. Contact initiation: The number of times the baby makes an effort to achieve contact. An instance of contact can be initiated by the baby, the mother or both. To capture the baby's effort, I look at the 5 second time window before contact is achieved and look whether the baby is actively moving towards the mother within that time window. I give a 0 or 1 score to the baby for each instance of contact.

4. Exploration initiation: The number of times the baby actively moves away from the mother after contact. The theory of the balance between attachment and exploratory system suggests when attachment system is activated to a high degree, the main goal for the infant would be to gain close contact to the mother. However, as the infant feels more secure, the attachment system will go to a lower activation state which allows the novel feature of the environment to activate the exploratory system. This can manifest in the baby being comfortable enough to gradually move away from the mother to explore the environment. This feature will capture this behavior.

5. Approach initiation: The number of times the baby is actively moving closer towards the mother. Approach behavior is one of the important elements that expert pays attention to when assigning ratings for attachment behaviors and attachment classification. For example, the baby's approach behavior is one of the key factors in in determining the rating for the proximity-seeking and avoidant behavior.

6. Positive response to mother approach: The number of times the baby is actively moving closer towards the mother whenever the mother is approaching the baby. For every single instance of mother approach, I increment the value of this feature if the baby is responding by locomoting towards the mother. This feature along with two other features (neutral/negative response to mother approach) are intended to capture the baby's response to the mother's attempt in gaining proximity, which is an important factor in determining the baby's attachment classification.

7. Neutral response to mother approach: The number of times the baby stays relatively at the same place whenever the mother is approaching the baby. For every single instance of mother approach, I increment the value of this feature if the baby is responding by ignoring (by not moving closer or further) the mother.

8. Negative response to mother approach: The number of times the baby is actively moving away from the mother whenever the mother is approaching the baby. For every single instance of mother approach, I increment the value of this feature if the baby is responding by avoiding (moving away) from the mother.

9. Baby being carried time: The length of time (in seconds) that the baby is being carried by the mother. I determine whether the baby is being carried or not by looking at the height of the baby's head with respect to the floor. If the height is above a certain threshold (90 cm), that means the mother is carrying the baby. This is significant in the Strange Situation since being carried usually indicates that the baby is crying. A high value for this features indicates that the baby is taking a long time

to use the mother to regulate his feelings and gain back a sense of security after being away during the period of separation.

10. Mean distance between baby and mother: Average distance between the baby and the mother during the whole reunion episode. This feature captures the overall proximity profile between the baby and the mother (i.e. whether the baby overall stays relatively close or far from the mother during the reunion).

In the following sections I will use the above feature set for the task of predicting the rating of attachment behaviors and attachment classification.

5.4.1 Predicting Rating of Attachment Behaviors Using Infant-Mother Proximity

As the first step in showing the usefulness of my proximity measure, I will demonstrate that this measure can be used to predict expert's rating for the 4 attachment behaviors in The Strange Situation. Looking at the detailed description on the rating system in Ainsworth's Patterns of Attachment, one certainly gets a sense that infant-mother proximity plays a big role in determining these ratings. For example, following is the description for a rating of 7 in proximity-seeking: ”The baby purposefully approaches the adult, creeping, crawling, or walking. He goes the whole way and actually achieves the contact.” Contrast this with the description for a rating of 2 in proximity-seeking: ”The baby begins to approach but stops, having gone only a short way.” Indeed, various distance-related events such as *approach*, *contact*, and *move away* are used as a way to describe the baby's behavior not only for the proximity-seeking dimension, but also for the other three categories (contact-maintenance, resistant, and avoidant). It is only natural that we should be able to use this proximity feature to predict these ratings.

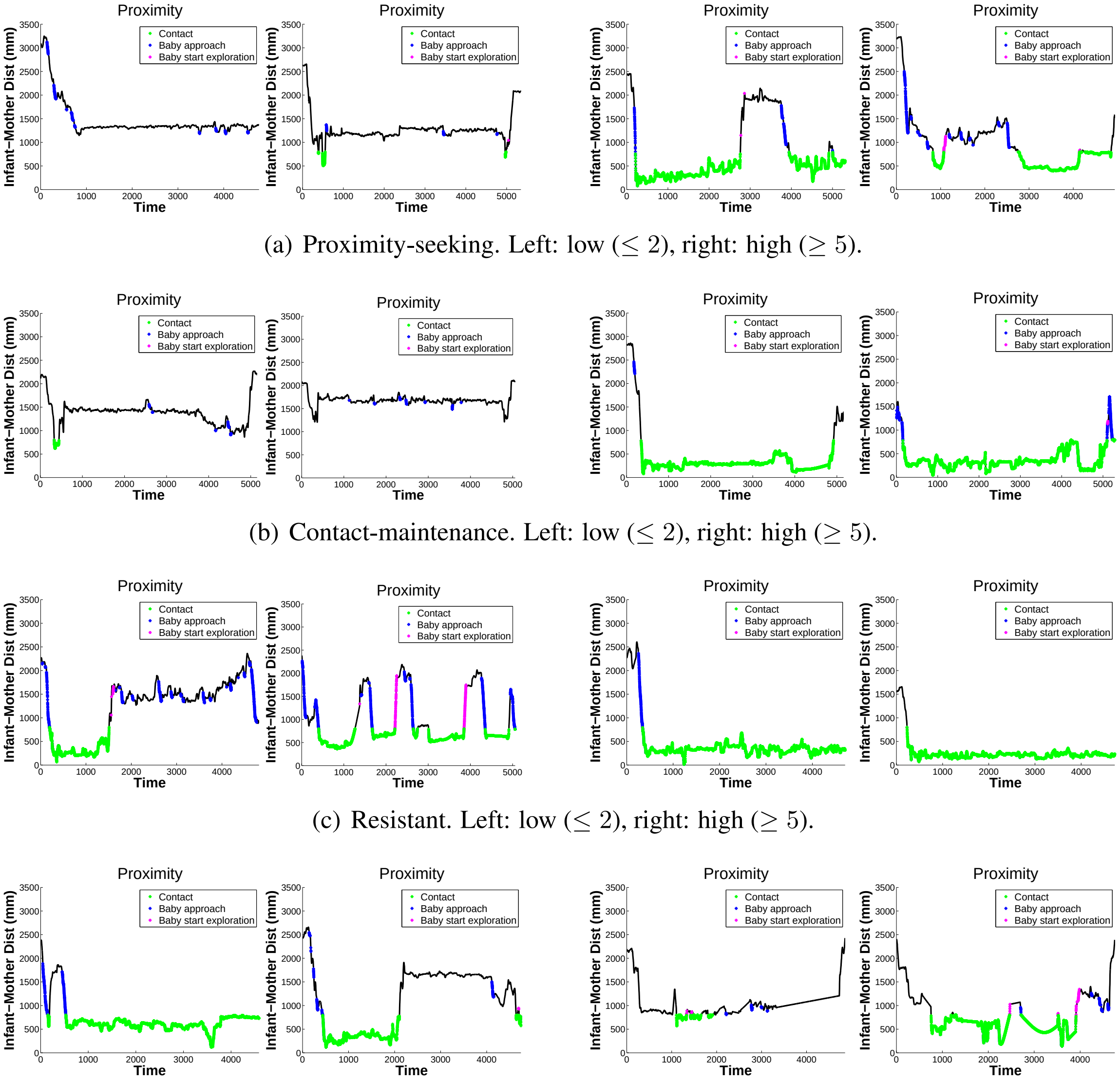
To further illustrate this point, I contrast the proximity profiles for each of the attachment behaviors in Figure 24. In Figure 24(a) we can easily see that the babies that are assigned high score in proximity-seeking show a period of active approach preceding contact with the mother. In contrast, the babies with low proximity-seeking score show minimal contact, and even when contact happened it is not preceded by an approach behavior. In Figure 24(c), the babies with high resistant score spend almost the entire length of the reunion episode in close contact with the mother, regardless of whether this contact is preceded by baby approach behavior or not. On the opposite end of the spectrum, we can see that babies with low resistant score are willing to be away from the mother for a significant period of time. Similar contrast can also be observed in the contact-maintenance and resistant behaviors. Note how infant-mother proximity measure allows us to easily produce visualization that can give us a quick way to get an overview of what is happening during a reunion episode.

In addition to being useful for visualization purpose, we can also use the proximity measure for predicting the expert rating of the attachment behaviors. I formulate this task as a linear regression problem where the task is to learn a linear mapping between the features derived from proximity (as described in Section 5.2.2) to the expert rating for each of the attachment behaviors. Let y_i^j be the expert rating for behavior j during reunion instance i, and $x_i = [x_{i1} \dots x_{i10}]^T$ be the 10 dimensional feature vector derived from the proximity measure of reunion instance i. The task is to learn a linear mapping β_j for each of the attachment behavior j such that:

$$y_i^j = \beta_j^T \begin{pmatrix} 1 \\ x_i \end{pmatrix} + \epsilon_i^j \tag{17}$$

where ϵ_i^j is all other factors which influence the expert rating y_i^j other than the our proximity-derived features x_i.

The problem of computing the linear mapping β_j between the proximity features x_i and the behavior rating y_i^j can be solved by using linear regression. For this work, I use the iterative reweighted least-squares method [60] that has been shown to be robust to outliers.



(a) Proximity-seeking. Left: low (≤ 2), right: high (≥ 5).

(b) Contact-maintenance. Left: low (≤ 2), right: high (≥ 5).

(c) Resistant. Left: low (≤ 2), right: high (≥ 5).

(d) Avoidant. Left: low (≤ 2), right: high (≥ 5).

Figure 24: Contrasting infant-mother proximity profiles in two opposite ends of the rating scales for the different attachment behaviors. Highlighted in green are periods of close contact with the mother, blue are moments when the baby is approaching the mother, and magenta are instances when the baby is moving away from the mother. Note the contrast in the proximity profile with respect to the ratings.

In this experiment I use the 34 fully-recorded sessions of The Strange Situation procedure. Since each session comprises of 2 reunion episodes, in total there are 68 datapoints for each of the attachment behaviors. The distribution of the ratings for the different behaviors can be seen in Figure 25. Note the change in the distribution of the ratings in reunion 2 (R2) compared to reunion 1 (R1). A likely cause for this is that the attachment system is more likely to be activated at a higher level during R2, so the infants tend to show a more pronounced attachment behaviors. I will discuss this further about this phenomenon in Section 5.4.2 when analyzing attachment classification.

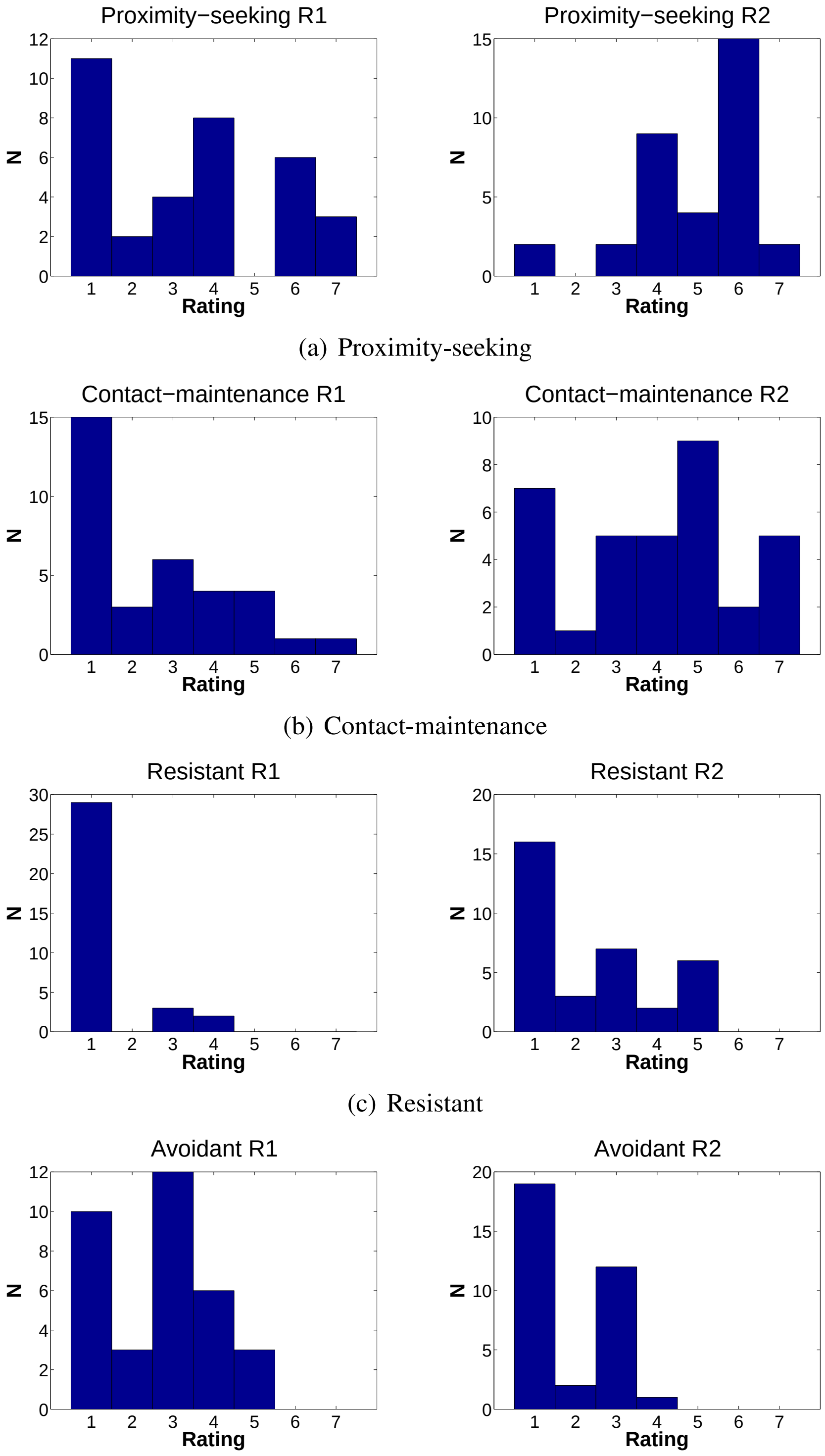
To see whether we can indeed use the features extracted from the proximity measure to predict expert attachment behavior ratings, I compute the linear mapping β_j for each of the the attachment behavior j using all 68 datapoints. Using the β_j we can easily compute our rating prediction:

$$\hat{y}_i^j = \beta_j^T \begin{pmatrix} 1 \\ x_i \end{pmatrix} \tag{18}$$

Note that the difference between the true expert rating y_i^j (Eq. 17) and the predicted rating $\hat{y}_i^j$ above is simply the noise term e_i^j. If the expert rating solely depends on proximity, the value of this variable should be close to 0. If infant-mother proximity is a major factor in determining the rating, the value for this noise term should be small.

To evaluate how well the linear regression model matches the expert rating, I compute the Pearson correlation coefficient (r-value), mean absolute difference (MD) and standard deviation (SD) between the expert rating y_i^j and the rating obtained by using the linear mapping $\hat{y}_i^j$ (Table 7). The results indicate that my predicted rating matches closely with expert rating for all attachment behaviors. This is shown by the statistically significant correlation value accompanied with low MD and SD. This results confirm my hypothesis that infant-mother proximity is a big factor that influences the expert rating.

Next, I perform an ablation experiment to investigate the significance of each individual features in predicting these expert ratings. I trained a linear regression model without



(a) Proximity-seeking

(b) Contact-maintenance

(c) Resistant

(d) Avoidant

Figure 25: The distribution of expert ratings for the different attachment behaviors. Note the noticeable difference in the distribution between reunion 1 (R1) and reunion 2 (R2) indicating a change in the trend of the behaviors.

Table 7: Pearson correlation coefficient (r-value), mean absolute difference (MD) and standard deviation (SD) between the expert rating y_i^j and the rating obtained by using the linear mapping $\hat{y}_i^j$. ** indicates highly statistically significant results ($p < 0.01$). The linear regression model is learned using all 68 datapoints.

Variable	**r-value**	**MD**	**SD**
Proximity-seeking - Reunion 1	0.780^{**}	1.154	1.344
Proximity-seeking - Reunion 2	0.610^{**}	0.990	1.281
Contact-maintenance - Reunion 1	0.790^{**}	0.896	1.092
Contact-maintenance - Reunion 2	0.826^{**}	0.996	1.136
Resistant - Reunion 1	0.626^{**}	0.464	0.702
Resistant - Reunion 2	0.595^{**}	0.929	1.252
Avoidant - Reunion 1	0.699^{**}	0.727	0.946
Avoidant - Reunion 2	0.706^{**}	0.552	0.738

using the target feature and calculate the p-value of the difference in the F-statistic of the ratings obtained with and without using the target feature. The results can be seen in Table 8. From the table, we can see that different features contribute differently in predicting the ratings. For example, the feature *proximity change* (change in infant-mother proximity that is caused by the baby's locomotion during the first 5 seconds of the reunion episode) play a significant role in determining the proximity-seeking, resistant, and avoidant ratings, but not contact-maintenance ($p > 0.1$). Intuitively, this aligns with how an expert is supposed to rate the contact-maintenance behavior, which is by observing the infant's efforts to maintain contact with mother once contact has been gained, and not by observing the change in proximity during the early moments of the reunion. Having said that, interpreting the results in Table 8 should be done with care. Since many of the features are correlated, it does not necessarily mean that a feature with $p > 0.1$ is not important for predicting a particular rating.

The previous set of experiments show that we can learn a model that maps these quantitative features to the expert ratings. Although the finding is encouraging, it is not a true prediction task since the linear mapping β_j that is used to produce the prediction $\hat{y}_i^j$ is learned using the expert rating y_i^j. In the prediction problem, the task is to produce the

Table 8: p-value of the difference in the F-statistic of the ratings obtained with and without using a particular feature.

Feature	Prox.-seeking	Contact-maint.	Resistant	Avoidant
Contact time	0.766	0.100	0.216	0.061
Proximity change	0.003	0.559	0.058	0.001
Contact initiation	0.051	0.850	0.889	0.039
Exploration initiation	0.559	0.636	0.544	0.248
Approach initiation	0.837	0.361	0.355	0.228
Positive response to mother approach	0.765	0.725	0.072	0.816
Neutral response to mother approach	0.207	0.618	0.091	0.889
Negative response to mother approach	0.851	0.061	0.137	0.615
Baby being carried time	0.100	0.803	0.271	0.698
Mean infant-mother distance	0.014	0.045	0.727	0.823

prediction without having access to the groundtruth of the testing data during the training phase. In this next experiment, I use the leave-one-subject-out experiment setup where for each i, I learn β_j^i using training set that comprises of reunions from all subjects other than the target subject, effectively reducing the number of datapoints from 68 ($34subject \times 2$) to 66 ($33subject \times 2$). The results of this experiment can be seen in Table 9. Although the accuracy is not as high as before, the numbers still indicate that the prediction matches closely with the expert ratings (significant correlation between $\hat{y}_i^j$ and y_i^j accompanied with low MD and SD). This combined with the previous results demonstrate that it is indeed possible to predict this qualitative expert ratings by using solely objective measure. Although there are definitely a number of other behavioral cues used by expert in determining these ratings (e.g. reach-in hand gesture), infant-mother proximity captures a large part of the differences in these ratings. Augmenting proximity with other behavioral feature will certainly help in producing a more accurate prediction system.

5.4.2 Predicting Attachment Classification Using Infant-Mother Proximity

After demonstrating that we can use the proximity measure to predict expert rating for attachment behaviors, in this Section I will show that we can also use this measure to discriminate the 3 attachment groups: insecure-avoidant (A), secure (B) and insecure-resitant

Table 9: Pearson correlation coefficient (r-value), mean absolute difference (MD) and standard deviation (SD) between the expert rating y_i^j and the rating obtained by using the linear mapping $\hat{y}_i^j$ in the leave-one-subject-out experiment setup. ** indicates highly statistically significant results ($p < 0.01$). The linear regression model for each $\hat{y}_i^j$ is learned using 66 datapoints (all reunions except for the target subject).

Variable	**r-value**	**MD**	**SD**
Proximity-seeking - Reunion 1	0.614**	1.446	1.719
Proximity-seeking - Reunion 2	0.444**	1.285	1.577
Contact-maintenance - Reunion 1	0.689**	1.127	1.384
Contact-maintenance - Reunion 2	0.693**	1.327	1.504
Resistant - Reunion 1	0.442**	0.758	1.210
Resistant - Reunion 2	0.326*	1.301	1.735
Avoidant - Reunion 1	0.454**	0.939	1.225
Avoidant - Reunion 2	0.453**	0.792	1.058

(C). For this experiment, I use the same 10 features as I did in the previous rating prediction experiment. The dataset that I use for this experiment comprises of 34 fully-recorded sessions and 1 partially-recorded session (only contains reunion 2) of The Strange Situation. Out of the 34 fully-recorded sessions, 1 infant is classified in the A group, 27 are in the B group and 6 are in the C group. The one infant in the partially-recorded session is classified to be in the A group. Note that this dataset is highly unbalanced reflecting the real-world distribution of these three classes [9], where the B group significantly outnumbers the other two groups.

First, I want to see whether the three attachment groups can be perfectly separated in the proximity feature space. It has been noted in Ainsworth's Patterns of Attachment [9] and Fraley and Spieker [49] that the three attachment groups are not perfectly linearly separable by their attachment behavior ratings, which is why there exists no direct mapping between expert behavior rating and attachment classification. To see whether the three groups are linearly separable in the proximity feature space, I train a support vector machine (SVM) model with linear kernel using all 34 datapoints from the fully-recorded sessions. Each session is represented by a 20 dimensional proximity-derived feature (10 from each reunion). Given a training data, SVM finds a hyperplane that best separates the different classes in the feature space. Therefore, 100% training accuracy in SVM means that the three classes are linearly separable in the feature space. The results of this experiment can be seen in Table 10. The table shows that the three classes are linearly separable (100% training accuracy) in my proximity-derived feature space. This is a very encouraging result since it means that the proximity-derived features that I propose fully encodes the differences in behaviors of the infants with respect to the attachment classification in the 34 Strange Situation sessions that I examine.

To investigate the informativeness of each individual feature, I perform another set of experiments to look at the training accuracy if the model is limited to only use a single feature (Table 10). Notice how none of the individual features can separate the three classes perfectly. Some features such as *approach initiation*, *positive response to mother approach*, and *mean infant-mother distance* is sufficient for differentiating the avoidant group from the rest (82.35% training accuracy). However, most individual features are unable to separate the 27 infants in the secure (B) group and the 7 in the insecure (A and C) groups, resulting in all infants being lumped together in a single group (79.41% training accuracy). Only by combining the features we can achieve perfect separation between the three groups.

Next, I perform an experiment to see which reunion episodes is most informative in separating the three attachment groups. It has been hypothesized that the infant attachment behavior is more pronounced during the second reunion since the infant attachment system is more likely to be activated to a higher level then. We have seen some evidence of this by comparing the distribution of the expert behavior rating in Figure 25. To further test this hypothesis, I perform an experiment to see whether we can separate the three attachment groups using features only from the individual reunion (i.e. features from R1 or R2 only). The results of this experiment can be seen in Table 11. It is very interesting to see that features extracted for reunion 1 poorly separate the three attachment groups, only achieving 85.3% training accuracy. This lends support to the hypothesis that during reunion 1,

Table 10: Training accuracy using various proximity-derived features to predict attachment classification. Accuracy of 100% means that the classes are linearly separable in the feature space.

Feature	**Accuracy**
Contact time	79.41%
Proximity change	79.41%
Contact initiation	79.41%
Exploration initiation	79.41%
Approach initiation	82.35%
Positive response to mother approach	82.35%
Neutral response to mother approach	79.41%
Negative response to mother approach	79.41%
Baby being carried time	79.41%
Mean infant-mother distance	82.35%
All combined	100%

Table 11: Training accuracy using features extracted from the individual reunions. Note how the features from reunion 1 are not very discriminative for the task of attachment classification.

Classification	**Features**		
	Reunion 1 only	Reunion 2 only	Both reunions
Insecure-Avoidant (A)	$1/1$	$1/1$	$1/1$
Secure (B)	$26/27$	$27/27$	$27/27$
Insecure-Resistant (C)	$2/6$	$6/6$	$6/6$
All	$29/34(85.3\%)$	$34/34(100\%)$	$34/34(100\%)$

the attachment system for some of the infants is not yet activated to a high level. This causes their attachment behaviors to be more muted during this reunion episode [9]. My proximity-derived features, which are designed to encode these attachment behaviors, capture this effect by showing less clear separation between the three groups compared to the results from reunion 2.

After showing that the three attachment groups are linearly separable in my proximity-derived feature space, I will now demonstrate the performance of my features for the task

Table 12: Attachment classification prediction accuracy using leave-one-subject-out training setup.

Classification	Accuracy
Insecure-Avoidant (A)	$1/2$
Secure (B)	$26/27$
Insecure-Resistant (C)	$3/6$
All	$30/35(85.7\%)$

of attachment classification prediction. For this task, I decided to only use features from reunion 2 since I have previously shown that this is sufficient for separating the three groups. Furthermore, by using only data from R2, I can now include the one session that is missing the R1 data. Similar to the previous setup for rating prediction, in this experiment I use the leave-one-subject-out experiment setup. The results from this experiment can be seen in Table 12. Notice that although my features are still able to produce very good accuracy for classifying the B group, on the surface they seem to perform poorly for the other groups. However, note that the size of the groups are very imbalanced where the A and C group only has 2 and 6 datapoints respectively. This means that in the leave-one-subject-out experiment setup, these groups will suffer the most from the reduction of datapoint during training. Having said that, the results of this experiment is still very encouraging since we are still able to correctly classify 30 out of the 35 datapoints.

5.5 Conclusions

The measure of physical distance is central in many studies in psychology. In this chapter I have demonstrated how we can move away from the traditional qualitative analysis of interaction by using a dense measure of proximity to deconstruct these complex, qualitative expert ratings. More specifically, I have shown how we can apply this measure to the problem of analyzing interaction between an infant and a mother in the context of an attachment study in The Strange Situation procedure. I validated this new measure by showing that we can use it to predict qualitative expert ratings, a quantity that is difficult to obtain due to the difficulty in training the human expert. I have also showed that we can use this measure to answer for new hypotheses that one might have about social interactions. For example, by comparing how well we can separate the three attachment groups in the proximity feature-space using features from reunion 1 and 2, I have added a new evidence to the hypothesis that infants tend to exhibit a more pronounced attachment behaviors in reunion 2 in The Strange Situation. Compared to expert rating, my proximity measure is easier to obtain since it does not require expert training and can be derived automatically from video data (given tracking output). This can potentially impact how researchers study and analyze social interactions.

CHAPTER VI

CONCLUSIONS AND FUTURE WORK

Videos have long been used by researchers and clinicians to assist in human behavior understanding. However, extracting useful behavioral measures from video data is often a difficult endeavor since it typically scales linearly with the length of the video and has to be done with painstaking attention to detail by a human expert. Despite this, more and more behavioral studies heavily incorporate videos in their analysis pipeline. For example, studies on infant locomotion [5, 6], preferential looking in individuals with autism [81, 71], and infant attachment [18, 43] base their findings on certain behavioral measures that are extracted from the recorded videos. The rapid growth of video data in this context combined with the time-intensive nature of the behavioral measure extraction process certainly create a data processing problem. It is not uncommon for an institution that heavily uses videos for behavioral research to have a backlog of *years* of unprocessed video data. This underscores the importance of better tools that can be used to assist in the extraction of useful behavioral measures from video data.

6.1 Summary

In this thesis, I address this problem by introducing two distinct approaches. The first one is a new method that can be used to assist the process of retrieving instances of a class of behaviors in a large video collection. A faster way to search for a certain target behavior in a video collection provides a tremendous value in many situations. One example that I focus on in this thesis is the task of retrieving instances of problem behaviors in individuals with autism. With the constantly changing landscape of problem behaviors, the ability to quickly retrieve additional examples of a problem behavior given one, or very few examples is valuable to aid the clinicians in understanding the behavior (e.g., in the context of behavioral therapy). I have demonstrated that my approach achieves state-of-the-art results in this real-world task of behavior retrieval from videos recorded in a classroom setting.

The second approach that I introduce is a novel way to adapt the cost function in interactive object tracking. A lot of behavioral phenomena can be derived from a continuous measure of object location (i.e. tracking). I have demonstrated that we can obtain high precision tracking results more efficiently (fewer human annotations) by leveraging user annotations to adapt the tracking cost function. Furthermore, I have also showed that we can use this tracking results to quantitatively analyze interaction through the measure of proximity. I have demonstrated how this quantitative proximity measure can be used to analyze the interaction between an infant and a mother in the context of The Strange Situation procedure, in contrast to the traditional human-based qualitative ratings.

6.2 Future Work

The aim of the approaches introduced in this thesis is to facilitate faster extraction of useful behavioral measures from video data through behavior retrieval and object tracking. There are a number of future work that can be done to further realize this goal.

6.2.1 Visualizing the Behavior Retrieval Results

The main goal of a behavior retrieval system is to provide a user the ability to quickly find other instances of a target behavior. From the algorithmic side, this means building a system that obtains high precision results at high recall, which is the problem that I address in Chapter 3. However, no retrieval method can obtain 100% precision at 100% recall. This means that a user will still have to sort through the search results to filter out the false positives. There is very little work that has been done on this problem. A common way to present video search results is by visualizing it as a list accompanied by a thumbnail image taken from a particular frame in the video (e.g., YouTube search results). However, this approach will not work for the behavior search problem since many behaviors are characterized by *movements* instead of a single canonical pose (e.g., it will be very hard to differentiate body rocking from sitting from just a single image). Playing the search results one at a time seems inefficient but on the other hand, playing many of them in a single page might overload the user's cognitive system. An effective way to present the behavior search results to the user will greatly improve the efficiency of a behavior retrieval system.

6.2.2 Interactive Multi-Object Tracking

Many videos contain more than a single object. Currently, most published work on interactive tracking [24, 161, 152, 151] focus on single object tracking. Tracking multiple objects in a video using any of these systems entails a repeated process of tracking a single object. There is very little study on how to efficiently track multiple objects in an interactive setting. From the user's perspective, Vondrick et.al. [151] find that simply asking the user to annotate all of the objects in a current frame overloads the user's cognitive system, resulting in a slower user performance compared to asking the user to just focus on annotating one object at a time. However, it seems unlikely that this is the most optimal way to ask for user annotation in interactive multi-object tracking. A potentially promising approach would be to selectively pick the set of objects for the user to annotate in a given frame, which is a middle ground to asking to annotate just a single object or all of the objects.

From the algorithmic perspective, a formulation that allows for an efficient way to optimize multiple object tracks simultaneously can improve the accuracy of the tracking results (and as a direct consequence, the system will require fewer human annotations which is the desired property of an interactive system). The dynamic programming single object tracking formulation first presented by Buchanan and Fitzgibbon [24] has been adopted by many interactive tracking approaches [161, 152, 151] because of its efficiency while still guaranteeing optimality of the results. An approach with the same property but for multiple objects setup will have a big impact on this problem.

REFERENCES

[1] "Fixerrors matlab gui." http://ctrax.sourceforge.net/fixerrors.html. Accessed: 2016-01-25.

[2] "Strange situation training." http://attachment-training.com/at/home/training/. Accessed: 2016-01-04.

[3] Adam, A., Rivlin, E., and Shimshoni, I., "Robust fragments-based tracking using the integral histogram," in *CVPR*, vol. 1, pp. 798–805, IEEE, 2006.

[4] Adelson, E. H. and Bergen, J. R., "Spatiotemporal energy models for the perception of motion," *America*, vol. 2, no. 2, pp. 284–299, 1985.

[5] Adolph, K. E., Bertenthal, B. I., Boker, S. M., Goldfield, E. C., and Gibson, E. J., "Learning in the development of infant locomotion," *Monographs of the Society for Research in Child Development*, pp. i–162, 1997.

[6] Adolph, K. E. and Robinson, S. R., "Motor development," *Handbook of child psychology and developmental science*, 2015.

[7] Aggarwal, J. and Ryoo, M. S., "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, p. 16, 2011.

[8] Ainsworth, M. D. S. and Bell, S. M., "Attachment, exploration, and separation: Illustrated by the behavior of one-year-olds in a strange situation," *Child development*, pp. 49–67, 1970.

[9] Ainsworth, M. D. S., Blehar, M. C., Waters, E., and Wall, S., "Patterns of attachment: A psychological study of the strange situation.," 1978.

[10] Ali, K., Hasler, D., and Fleuret, F., "Flowboostappearance learning from sparsely annotated video," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1433–1440, IEEE, 2011.

[11] Arun, K. S., Huang, T. S., and Blostein, S. D., "Least-squares fitting of two 3-d point sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 5, pp. 698–700, 1987.

[12] Aslandogan, Y., "Techniques and Systems for Image and Video Retrieval," *Knowledge and Data Engineering,*, 1999.

[13] Association, A. P., *Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR, Fourth Edition, Text Revision.* American Psychiatric Pub., 2000.

[14] Avidan, S., "Support vector tracking," *PAMI*, vol. 26, no. 8, pp. 1064–1072, 2004.

[15] BABENKO, B., YANG, M.-H., and BELONGIE, S., "Robust object tracking with online multiple instance learning," *PAMI*, vol. 33, no. 8, pp. 1619–1632, 2011.

[16] BALAN, A. O. and BLACK, M. J., "An adaptive appearance model approach for model-based articulated object tracking," in *Computer Vision and Pattern Recognition*, vol. 1, pp. 758–765, IEEE, 2006.

[17] BELLMAN, R. E. and DREYFUS, S. E., "Applied dynamic programming," 1962.

[18] BELSKY, J., HOUTS, R. M., and FEARON, R. P., "Infant attachment security and the timing of puberty testing an evolutionary hypothesis," *Psychological Science*, vol. 21, no. 9, pp. 1195–1201, 2010.

[19] BENFOLD, B. and REID, I., "Stable multi-target tracking in real-time surveillance video," in *Computer Vision and Pattern Recognition (CVPR)*, pp. 3457–3464, IEEE, 2011.

[20] BESL, P. J. and MCKAY, N. D., "Method for registration of 3-d shapes," in *Robotics-DL tentative*, pp. 586–606, International Society for Optics and Photonics, 1992.

[21] BLEIWEISS, A. and WERMAN, M., "Fusing time-of-flight depth and color for real-time segmentation and tracking," in *Dynamic 3D Imaging*, pp. 58–69, Springer, 2009.

[22] BOBICK, A. and DAVIS, J., "The recognition of human movement using temporal templates," *PAMI*, 2001.

[23] BRANSON, K., ROBIE, A. A., BENDER, J., PERONA, P., and DICKINSON, M. H., "High-throughput ethomics in large groups of drosophila," *Nature methods*, vol. 6, no. 6, pp. 451–457, 2009.

[24] BUCHANAN, A. and FITZGIBBON, A., "Interactive feature tracking using kd trees and dynamic programming," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1, pp. 626–633, IEEE, 2006.

[25] BUSHBY, K., COLE, T., MATTHEWS, J., and GOODSHIP, J., "Centiles for adult head circumference.," *Archives of disease in childhood*, vol. 67, no. 10, pp. 1286–1287, 1992.

[26] CAO, L., JI, R., GAO, Y., LIU, W., and TIAN, Q., "Mining spatiotemporal video patterns towards robust action retrieval," *Neurocomputing*, 2012.

[27] CHANG, C.-C. and LIN, C.-J., "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[28] CHAUDHRY, R., RAVICHANDRAN, A., HAGER, G., and VIDAL, R., "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," *CVPR*, 2009.

[29] Choi, W., Pantofaru, C., and Savarese, S., "Detecting and tracking people using an rgb-d camera via multiple detector fusion," in *ICCV Workshops*, pp. 1076–1083, IEEE, 2011.

[30] Collins, R. T., Liu, Y., and Leordeanu, M., "Online selection of discriminative tracking features," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 10, pp. 1631–1643, 2005.

[31] Comaniciu, D. and Meer, P., "Mean shift: A robust approach toward feature space analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 603–619, 2002.

[32] Comaniciu, D., Ramesh, V., and Meer, P., "Kernel-based object tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 5, pp. 564–577, 2003.

[33] Dalal, N. and Triggs, B., "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 886–893, IEEE, 2005.

[34] Dankert, H., Wang, L., Hoopfer, E. D., Anderson, D. J., and Perona, P., "Automated monitoring and analysis of social behavior in drosophila," *Nature methods*, vol. 6, no. 4, pp. 297–303, 2009.

[35] DeMenthon, D. and Doermann, D., "Video retrieval using spatio-temporal descriptors," *ACM Multimedia*, 2003.

[36] Dementhon, D. and Doermann, D., "Video Retrieval of Near Duplicates using k Nearest Neighbor Retrieval of Spatio Temporal Descriptors," *Collections*, 2004.

[37] Dempster, A. P., Laird, N. M., and Rubin, D. B., "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[38] Dexter, E. and Laptev, I., "Multi-view synchronization of human actions and dynamic scenes," *BMVC*, 2009.

[39] Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S., "Behavior Recognition via Sparse Spatio-Temporal Features," *ICCV-VS PETS*, 2005.

[40] Dollár, P., Appel, R., Belongie, S., and Perona, P., "Fast feature pyramids for object detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 8, pp. 1532–1545, 2014.

[41] Eggert, D. W., Lorusso, A., and Fisher, R. B., "Estimating 3-d rigid body transformations: a comparison of four major algorithms," *Machine Vision and Applications*, vol. 9, no. 5-6, pp. 272–290, 1997.

[42] Elhayek, A., Stoll, C., Kim, K., Seidel, H.-P., and Theobalt, C., "Feature-based multi-video synchronization with subframe accuracy," in *Pattern Recognition*, pp. 266–275, Springer, 2012.

[43] Esposito, G., del Carmen Rostagno, M., Venuti, P., Haltigan, J. D., and Messinger, D. S., "Brief report: Atypical expression of distress during the separation phase of the strange situation procedure in infant siblings at high risk for asd," *Journal of autism and developmental disorders*, vol. 44, no. 4, pp. 975–980, 2014.

[44] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A., "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.

[45] Farhadi, A., "Learning to Recognize Activities from the Wrong View Point," *ECCV*, 2008.

[46] Fathi, A., Balcan, M.-F., Ren, X., and Rehg, J. M., "Combining self training and active learning for video segmentation.," in *BMVC*, pp. 1–11, 2011.

[47] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D., "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, 2010.

[48] for Disease Control, C. and Prevention, N. C. f. H. S., *Data Table of Infant Head Circumference-for-age Charts.*

[49] Fraley, R. C. and Spieker, S. J., "Are infant attachment patterns continuously or categorically distributed? a taxometric analysis of strange situation behavior.," *Developmental psychology*, vol. 39, no. 3, p. 387, 2003.

[50] Girshick, R., Donahue, J., Darrell, T., and Malik, J., "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587, IEEE, 2014.

[51] Gorban, A., Idrees, H., Jiang, Y.-G., Roshan Zamir, A., Laptev, I., Shah, M., and Sukthankar, R., "THUMOS challenge: Action recognition with a large number of classes." `http://www.thumos.info/`, 2015.

[52] Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R., "Actions as Space-Time Shapes.," *PAMI*, 2007.

[53] Grabner, H., Leistner, C., and Bischof, H., "Semi-supervised on-line boosting for robust tracking," in *ECCV*, pp. 234–247, Springer, 2008.

[54] Grundmann, M., Kwatra, V., and Essa, I., "Auto-directed video stabilization with robust l1 optimal camera paths," in *CVPR*, 2011.

[55] HALL, D. and PERONA, P., "Online, real-time tracking using a category-to-individual detector," in *ECCV*, pp. 361–376, Springer, 2014.

[56] HANJALIC, A., LAGENDIJK, R. L., MEMBER, S., and BIEMOND, J., "Automated High-Level Movie Segmentation for Advanced Video-Retrieval Systems," *IEEE Transactions on Circuits and Systems*, vol. 9, no. 4, pp. 580–588, 1999.

[57] HARE, S., SAFFARI, A., and TORR, P. H., "Struck: Structured output tracking with kernels," in *ICCV*, pp. 263–270, IEEE, 2011.

[58] HARMAN, H. H., "Modern factor analysis.," 1960.

[59] HARRIS, C. and STEPHENS, M., "A combined corner and edge detector.," in *Alvey vision conference*, vol. 15, p. 50, Manchester, UK, 1988.

[60] HOLLAND, P. W. and WELSCH, R. E., "Robust regression using iteratively reweighted least-squares," *Communications in Statistics-theory and Methods*, vol. 6, no. 9, pp. 813–827, 1977.

[61] HONG, W., KENNEDY, A., BURGOS-ARTIZZU, X. P., ZELIKOWSKY, M., NAVONNE, S. G., PERONA, P., and ANDERSON, D. J., "Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning," *Proceedings of the National Academy of Sciences*, vol. 112, no. 38, pp. E5351–E5360, 2015.

[62] HORN, B. K., "Closed-form solution of absolute orientation using unit quaternions," *JOSA A*, vol. 4, no. 4, pp. 629–642, 1987.

[63] HSU, E., PULLI, K., and POPOVIĆ, J., "Style translation for human motion," *ACM TOG*, vol. 24, no. 3, pp. 1082–1089, 2005.

[64] İKIZLER, N. and FORSYTH, D. A., "Searching for complex human activities with no visual examples," *International Journal of Computer Vision*, vol. 80, no. 3, pp. 337–357, 2008.

[65] INRIA, "Inria children dataset (*http://4drepository.inrialpes.fr/public/viewgroup/2*)."

[66] ISARD, M. and BLAKE, A., "Condensationconditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[67] JAIN, M., JÉGOU, H., BOUTHEMY, P., and OTHERS, "Better exploiting motion for better action recognition," *CVPR*, 2013.

[68] JIA, X., LU, H., and YANG, M.-H., "Visual tracking via adaptive structural local sparse appearance model," in *Computer vision and pattern recognition (CVPR)*, pp. 1822–1829, IEEE, 2012.

[69] JOHANSSON, G., "Visual Perception of Biological Motion and a Model for Its Analysis," *Attention, Perception, & Psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[70] JOHNSON, A. E. and BING KANG, S., “Registration and integration of textured 3d data,” *Image and vision computing*, vol. 17, no. 2, pp. 135–147, 1999.

[71] JONES, W. and KLIN, A., “Attention to eyes is present but in decline in 2-6-month-old infants later diagnosed with autism,” *Nature*, vol. 504, no. 7480, pp. 427–431, 2013.

[72] JUNEJO, I., DEXTER, E., LAPTEV, I., and PÉREZ, P., “Cross-View Action Recognition from Temporal Self-Similarities,” in *ECCV*, 2008.

[73] JUNEJO, I., DEXTER, E., LAPTEV, I., and PÉREZ, P., “View-Independent Action Recognition from Temporal Self-Similarities,” *PAMI*, 2010.

[74] JUNG, S.-H., GUO, Y., SAWHNEY, H., and KUMAR, R., “Action video retrieval based on atomic action vocabulary,” in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pp. 245–252, ACM, 2008.

[75] KABRA, M., ROBIE, A. A., RIVERA-ALBA, M., BRANSON, S., and BRANSON, K., “Jaaba: interactive machine learning for automatic annotation of animal behavior,” *nature methods*, 2012.

[76] KALAL, Z., MIKOLAJCZYK, K., and MATAS, J., “Tracking-learning-detection,” *PAMI*, vol. 34, no. 7, pp. 1409–1422, 2012.

[77] KALMAN, R. E., “A new approach to linear filtering and prediction problems,” *Journal of Fluids Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[78] KE, Y., SUKTHANKAR, R., and HEBERT, M., “Spatio-temporal Shape and Flow Correlation for Action Recognition,” *CVPR Workshop on Visual Surveilance*, 2007.

[79] KE, Y., SUKTHANKAR, R., and HEBERT, M., “Volumetric features for video event detection,” *IJCV*, 2010.

[80] KIM, K., CHALIDABHONGSE, T. H., HARWOOD, D., and DAVIS, L., “Real-time foreground–background segmentation using codebook model,” *Real-time imaging*, vol. 11, no. 3, pp. 172–185, 2005.

[81] KLIN, A., JONES, W., SCHULTZ, R., VOLKMAR, F., and COHEN, D., “Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism,” *Archives of general psychiatry*, vol. 59, no. 9, pp. 809–816, 2002.

[82] KLIPER-GROSS, O., GUROVICH, Y., HASSNER, T., and WOLF, L., “Motion interchange patterns for action recognition in unconstrained videos,” *ECCV*, 2012.

[83] KRIZHEVSKY, A., SUTSKEVER, I., and HINTON, G. E., “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems (NIPS)*, pp. 1097–1105, 2012.

[84] KUHN, H. W., "The Hungarian Method for the Assignment Problem," *Naval Research Logistics Quarterly*, 1955.

[85] KWON, J. and LEE, K. M., "Visual tracking decomposition," in *CVPR*, pp. 1269–1276, IEEE, 2010.

[86] LAM, K. S. and AMAN, M. G., "The repetitive behavior scale-revised: Independent validation in individuals with autism spectrum disorders," *Journal of autism and developmental disorders*, vol. 37, no. 5, pp. 855–866, 2007.

[87] LAPTEV, I., INRIA, I., BEAULIEU, C., and CEDEX, R., "On Space-Time Interest Points," *IJCV*, vol. 64, pp. 107–123, 2005.

[88] LAPTEV, I., MARSZALEK, M., SCHMID, C., and ROZENFELD, B., "Learning Realistic Human Actions from Movies," in *CVPR*, 2008.

[89] LE, Q. V., ZOU, W. Y., YEUNG, S. Y., and NG, A. Y., "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3361–3368, 2011.

[90] LI, B., CAMPS, O. I., and SZNAIER, M., "Cross-view activity recognition using hankelets," *CVPR*, 2012.

[91] LI, H., SHEN, C., and SHI, Q., "Real-time visual tracking using compressive sensing," in *CVPR*, pp. 1305–1312, IEEE, 2011.

[92] LI, R. and ZICKLER, T., "Discriminative Virtual Views for Cross-View Action Recognition," *CVPR*, 2012.

[93] LING, H. and OKADA, K., "An efficient Earth Mover's Distance algorithm for robust histogram comparison.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, pp. 840–53, May 2007.

[94] LIU, J., SHAH, M., KUIPERS, B., and SAVARESE, S., "Cross-View Action Recognition via View Knowledge Transfer," in *CVPR*, 2011.

[95] LORD, C., RUTTER, M., DILAVORE, P., RISI, S., and GOTHAM, K., *ADOS: Autism Diagnostic Observation Schedule*. Western Psychological Services, 2008.

[96] LOWE, D. G., "Object recognition from local scale-invariant features," in *International Conference on Computer Vision (ICCV)*, vol. 2, pp. 1150–1157, Ieee, 1999.

[97] LOWE, D. G., "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[98] LUCAS, B. D., KANADE, T., and OTHERS, "An iterative image registration technique with an application to stereo vision.," in *IJCAI*, vol. 81, pp. 674–679, 1981.

[99] M. R. Naphade, T. Kristjansson, B. Frey, T. S. H., "Probabilistic multimedia objects (multijects): a novel approach to video indexing and retrieval in multimedia systems," *International Conference on Image Processin*, 1998.

[100] MacDonald, R., Green, G., Mansfield, R., Geckeler, A., Gardenier, N., Anderson, J., Holcomb, W., and Sanchez, J., "Stereotypy in young children with autism and typically developing children," *Research in Developmental Disabilities*, vol. 28, no. 3, pp. 266–277, 2007.

[101] Maji, S., Berg, A. C., and Malik, J., "Classication using Intersection Kernel Support Vector Machines is Efcient," *CVPR*, 2008.

[102] Messing, R., Pal, C., and Kautz, H., "Activity Recognition Using the Velocity Histories of Tracked Keypoints," in *ICCV*, 2009.

[103] Meyer, M., Desbrun, M., Schröder, P., and Barr, A. H., "Discrete differential-geometry operators for triangulated 2-manifolds," in *Visualization and mathematics III*, pp. 35–57, Springer, 2003.

[104] Mian, A. S., Bennamoun, M., and Owens, R., "Three-dimensional model-based object recognition and segmentation in cluttered scenes," *PAMI*, vol. 28, no. 10, pp. 1584–1601, 2006.

[105] Mirat, O., Sternberg, J. R., Severi, K. E., and Wyart, C., "Zebrazoom: an automated program for high-throughput behavioral analysis and categorization," *Frontiers in neural circuits*, vol. 7, 2013.

[106] Newcombe, R. A., Davison, A. J., Izadi, S., Kohli, P., Hilliges, O., Shotton, J., Molyneaux, D., Hodges, S., Kim, D., and Fitzgibbon, A., "Kinectfusion: Real-time dense surface mapping and tracking," in *ISMAR*, pp. 127–136, IEEE, 2011.

[107] Nguyen, T.-N., Michaelis, B., Al-Hamadi, A., Tornow, M., and Meinecke, M., "Stereo-camera-based urban environment perception using occupancy grid and object tracking," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 13, no. 1, pp. 154–165, 2012.

[108] Niebles, J., Wang, H., and Li, F.-F., "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," *IJCV*, vol. 79, no. 3, pp. 299–318, 2008.

[109] Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.-C., Lee, J. T., Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., and others, "A large-scale benchmark dataset for event recognition in surveillance video," in *Computer Vision and Pattern Recognition (CVPR)*, pp. 3153–3160, IEEE, 2011.

[110] Ohayon, S., Avni, O., Taylor, A. L., Perona, P., and Egnor, S. R., "Automated multi-day tracking of marked mice for the analysis of social behaviour," *Journal of neuroscience methods*, vol. 219, no. 1, pp. 10–19, 2013.

[111] OIKONOMIDIS, I., KYRIAZIS, N., and ARGYROS, A. A., “Efficient model-based 3d tracking of hand articulations using kinect.,” in *BMVC*, pp. 1–11, 2011.

[112] PARK, D., RAMANAN, D., and FOWLKES, C., “Multiresolution models for object detection,” in *Computer Vision–ECCV 2010*, pp. 241–254, Springer, 2010.

[113] PÉREZ-ESCUDERO, A., VICENTE-PAGE, J., HINZ, R. C., ARGANDA, S., and DE POLAVIEJA, G. G., “idtracker: tracking individuals in a group by automatic identification of unmarked animals,” *Nature methods*, vol. 11, no. 7, pp. 743–748, 2014.

[114] PERRONNIN, F., SÁNCHEZ, J., and MENSINK, T., “Improving the fisher kernel for large-scale image classification,” *ECCV*, pp. 143–156, 2010.

[115] POPPE, R., “A survey on vision-based human action recognition,” *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[116] PRABHAKAR, K., OH, S., WANG, P., ABOWD, G. D., and REHG, J. M., “Temporal causality for the analysis of visual events,” in *Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2010.

[117] PRISACARIU, V. A. and REID, I. D., “Pwp3d: Real-time segmentation and tracking of 3d objects,” *IJCV*, vol. 98, no. 3, pp. 335–354, 2012.

[118] PRISACARIU, V. A., SEGAL, A. V., and REID, I., “Simultaneous monocular 2d segmentation, 3d pose recovery and 3d reconstruction,” in *ACCV*, pp. 593–606, Springer, 2013.

[119] RAO, C., GRITAI, A., and SHAH, M., “View-invariant Alignment and Matching of Video Sequences 2 . View-invariant Alignment of Video,” *Electrical Engineering*.

[120] REDDY, K. K. and SHAH, M., “Recognizing 50 human action categories of web videos,” *Machine Vision and Applications*, 2012.

[121] REHG, J. M. and OTHERS, “Decoding children’s social behavior,” in *Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2013.

[122] REHG, J. M. and KANADE, T., “Model-based tracking of self-occluding articulated objects,” in *International Conference on Computer Vision*, pp. 612–617, IEEE, 1995.

[123] REN, C. Y., PRISACARIU, V. A., MURRAY, D. W., and REID, I. D., “Star3d: Simultaneous tracking and reconstruction of 3d objects using rgb-d data,” *ICCV*, 2013.

[124] ROSS, D. A., LIM, J., LIN, R.-S., and YANG, M.-H., “Incremental learning for robust visual tracking,” *IJCV*, vol. 77, no. 1-3, pp. 125–141, 2008.

[125] ROTHBART, M. K., “Temperament and the development of inhibited approach,” *Child Development*, pp. 1241–1250, 1988.

[126] ROTHER, C., KOLMOGOROV, V., and BLAKE, A., “Grabcut: Interactive foreground extraction using iterated graph cuts,” in *ACM Transactions on Graphics (TOG)*, vol. 23, pp. 309–314, ACM, 2004.

[127] RUBNER, Y., TOMASI, C., and GUIBAS, L. J., “The Earth Movers Distance as a Metric for Image Retrieval,” *IJCV*, vol. 40, no. 2, pp. 99–121, 2000.

[128] RUSSELL, B. C., TORRALBA, A., MURPHY, K. P., and FREEMAN, W. T., “Labelme: a database and web-based tool for image annotation,” *International journal of computer vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

[129] SADANAND, S. and CORSO, J. J., “Action bank: A high-level representation of activity in video,” in *CVPR*, 2012.

[130] SATAKE, J. and MIURA, J., “Robust stereo-based person detection and tracking for a person following robot,” in *ICRA Workshop on People Detection and Tracking*, 2009.

[131] SERBY, D., MEIER, E., and VAN GOOL, L., “Probabilistic object tracking using multiple features,” in *International Conference on Pattern Recognition (ICPR)*, vol. 2, pp. 184–187, IEEE, 2004.

[132] SETTLES, B., “Active learning literature survey,” *Computer Sciences Technical Report 1648, University of Wisconsin Madison*, 2009.

[133] SHEN, Y., FOROOSH, H., and FLORIDA, C., “View-Invariant Action Recognition Using Fundamental Ratios,” *Database*, pp. 1–6, 2008.

[134] SHI, F., PETRIU, E., and LAGANIERE, R., “Sampling strategies for real-time action recognition,” in *CVPR*, pp. 2595–2602, IEEE, 2013.

[135] SHI, J. and TOMASI, C., “Good features to track,” *CVPR*, 1994.

[136] SOLMAZ, B., ASSARI, S. M., and SHAH, M., “Classifying web videos using a global video descriptor,” *Machine Vision and Applications*, 2012.

[137] SOMMER, R., “Personal space. the behavioral basis of design.,” 1969.

[138] SONG, S. and XIAO, J., “Tracking revisited using rgbd camera: Unified benchmark and baselines,” in *ICCV*, ICCV, 2013.

[139] SORKINE, O. and ALEXA, M., “As-rigid-as-possible surface modeling,” in *Symposium on Geometry processing*, vol. 4, 2007.

[140] STRAW, A. D., BRANSON, K., NEUMANN, T. R., and DICKINSON, M. H., “Multicamera real-time three-dimensional tracking of multiple flying animals,” *Journal of The Royal Society Interface*, p. rsif20100230, 2010.

[141] SZUMMER, M., KOHLI, P., and HOIEM, D., “Learning crfs using graph cuts,” in *ECCV*, pp. 582–595, Springer, 2008.

[142] TANG, F., HARVILLE, M., TAO, H., and ROBINSON, I. N., "Fusion of local appearance with stereo depth for object tracking," in *CVPR*, IEEE, 2008.

[143] TASKAR, B., CHATALBASHEV, V., KOLLER, D., and GUESTRIN, C., "Learning structured prediction models: A large margin approach," in *Proceedings of the 22nd international conference on Machine learning*, pp. 896–903, ACM, 2005.

[144] THI, T. H., CHENG, L., ZHANG, J., and WANG, L., "IMPLICIT MOTION-SHAPE MODEL : A GENERIC APPROACH FOR ACTION MATCHING National ICT of Australia (NICTA) & University of New South Wales , 2032 , NSW , Australia Toyota Technological Institute at Chicago , 60637 , Illinois , USA," pp. 1477–1480, 2010.

[145] TORRESANI, L. and HERTZMANN, A., "Automatic non-rigid 3d modeling from video," in *ECCV*, pp. 299–312, Springer, 2004.

[146] TRAN, D. and SOROKIN, A., "Human Activity Recognition with Metric Learning," *ECCV*, 2008.

[147] TRESADERN, P. A. and REID, I. D., "Video synchronization from human motion using rank constraints," *CVIU*, 2009.

[148] TSOCHANTARIDIS, I., JOACHIMS, T., HOFMANN, T., and ALTUN, Y., "Large margin methods for structured and interdependent output variables," in *Journal of Machine Learning Research*, pp. 1453–1484, 2005.

[149] VIJAYANARASIMHAN, S. and GRAUMAN, K., "Active frame selection for label propagation in videos," in *European Conference on Computer Vision (ECCV)*, pp. 496–509, Springer, 2012.

[150] VIOLA, P. and JONES, M., "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. I–511, IEEE, 2001.

[151] VONDRICK, C., PATTERSON, D., and RAMANAN, D., "Efficiently scaling up crowdsourced video annotation," *International Journal of Computer Vision*, vol. 101, no. 1, pp. 184–204, 2013.

[152] VONDRICK, C. and RAMANAN, D., "Video annotation and tracking with active learning," in *Advances in Neural Information Processing Systems*, pp. 28–36, 2011.

[153] WANG, H., "Evaluation of local spatio-temporal features for action recognition," *Methods*.

[154] WANG, H., KL, A., SCHMID, C., and LIU, C.-L., "Action Recognition by Dense Trajectories," *CVPR*, 2011.

[155] WANG, H., KLÄSER, A., SCHMID, C., and LIU, C.-L., "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[156] WANG, H. and SCHMID, C., "Action recognition with improved trajectories," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3551–3558, 2013.

[157] WANG, P., ABOWD, G. D., and REHG, J. M., "Quasi-periodic event analysis for social game retrieval," in *International Conference on Computer Vision*, IEEE, 2009.

[158] WANG, Y., GUPTA, M., ZHANG, S., WANG, S., GU, X., SAMARAS, D., and HUANG, P., "High resolution tracking of non-rigid motion of densely sampled 3d data using harmonic maps," *IJCV*, vol. 76, no. 3, pp. 283–300, 2008.

[159] WEI, X., ZHANG, P., and CHAI, J., "Accurate realtime full-body motion capture using a single depth camera," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 6, p. 188, 2012.

[160] WEI, X. K. and CHAI, J., "Interactive tracking of 2d generic objects with spacetime optimization," in *Computer Vision–ECCV 2008*, pp. 657–670, Springer, 2008.

[161] WEI, Y., SUN, J., TANG, X., and SHUM, H.-Y., "Interactive offline tracking for color objects," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8, IEEE, 2007.

[162] WEINLAND, D., BOYER, E., and RONFARD, R., "Action Recognition from Arbitrary Views using 3D Exemplars," *IJCV*, 2007.

[163] WEINLAND, D., RONFARD, R., and BOYER, E., "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 249–257, 2006.

[164] WEISLER, A. and MCCALL, R. R., "Exploration and play: Resume and redirection.," *American Psychologist*, vol. 31, no. 7, p. 492, 1976.

[165] WERLBERGER, M., TROBIN, W., POCK, T., WEDEL, A., CREMERS, D., and BISCHOF, H., "Anisotropic Huber-L1 optical flow," *BMVC*, 2009.

[166] WERMAN, M., "Fast and Robust Earth Mover s Distances," *Most*.

[167] WHITAKER, R. T., "A level-set approach to 3d reconstruction from range data," *IJCV*, vol. 29, no. 3, pp. 203–231, 1998.

[168] YILMAZ, A., JAVED, O., and SHAH, M., "Object tracking: A survey," *Acm computing surveys (CSUR)*, vol. 38, no. 4, p. 13, 2006.

[169] YILMAZ, A., LI, X., and SHAH, M., "Contour-based object tracking with occlusion handling in video acquired using mobile cameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1531–1536, 2004.

[170] YU, G., YUAN, J., and LIU, Z., "Real-time human action search using random forest based hough voting," in *Proceedings of the 19th ACM international conference on Multimedia*, pp. 1149–1152, ACM, 2011.

[171] Yuan, C., Li, X., Hu, W., Ling, H., and Maybank, S., "3d r transform on spatio-temporal interest points for action recognition," *CVPR*, 2013.

[172] Yuen, J., Russell, B., Liu, C., and Torralba, A., "Labelme video: Building a video database with human annotations," in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 1451–1458, IEEE, 2009.

[173] Zen, G. and Ricci, E., "Earth Movers Prototypes: a Convex Learning Approach for Discovering Activity Patterns in Dynamic Scenes," *CVPR*, 2011.

[174] Zentner, M. and Bates, J. E., "Child temperament: An integrative review of concepts, research programs, and measures," *International Journal of Developmental Science*, vol. 2, no. 1, pp. 7–37, 2008.

[175] Zhang, K., Zhang, L., and Yang, M.-H., "Real-time compressive tracking," *ECCV*, pp. 864–877, 2012.

[176] Zhang, Z., "Iterative point matching for registration of free-form curves and surfaces," *IJCV*, vol. 13, no. 2, pp. 119–152, 1994.

[177] Zhang, Z., Wang, C., Xiao, B., Zhou, W., Liu, S., and Shi, C., "Cross-view action recognition via a continuous virtual path," *CVPR*, 2013.

[178] Zhao, T., Aggarwal, M., Kumar, R., and Sawhney, H., "Real-time wide area multi-camera stereo tracking," in *CVPR*, vol. 1, pp. 976–983, IEEE, 2005.

[179] Zhou, F. and de la Torre, F., "Canonical Time Warping for Alignment of Human Behavior," *NIPS*, 2009.

[180] Zhou, S. K., Chellappa, R., and Moghaddam, B., "Visual tracking and recognition using appearance-adaptive models in particle filters," *IEEE Transactions on Image Processing*, vol. 13, no. 11, pp. 1491–1506, 2004.