# ABSTRACT

Title of dissertation: ACTION
COMPOSITIONALITY
WITH FOCUS ON
NEURODEVELOPMENTAL DISORDERS

Leonardo Claudino, Doctor of Philosophy, 2016

Dissertation directed by: Professor Yiannis Aloimonos
Department of Computer Science

A central question in motor neuroscience is how the Central Nervous System (CNS) would handle flexibility at the effector level, that is, how the brain would solve the problem coined by Nikolai Bernstein as the "degrees of freedom problem", or the task of controlling a much larger number of degrees of freedom (dofs) that is often needed to produce behavior. Flexibility is a bless and a curse: while it enables the same body to engage in a virtually infinite number of behaviors, the CNS is left with the job of figuring out the right subset of dofs to use and how to control and coordinate these degrees. Similarly, at the level of perception, the CNS seeks to obtain information pertaining to the action and actors involved based on perceived motion of other people's dofs.

This problem is believed to be solved with a particular dimensionality reduction strategy, where *action production* would consist of tuning only a few parameters that control and coordinate a small number of motor primitives, and *action perception* would take place by applying grouping processes that would solve the inverse problem, that is to identify the motor primitives and the corresponding tuning parameters used by an actor. These parameters can encode not only information on the action per se, but also higher-order cognitive cues like body language or emotion. This compositional view of action representation has an obvious parallel with language: we can think of primitives as words and cognitive systems (motor, perceptual) as different languages.

Little is known, however, about how words/primitives would be formed from low-level signals measured at each dof. Here we introduce the SB-ST method, a bottom-up approach to find full-body postural primitives as a set of key postures, that is, vectors corresponding to key relationships among dofs (such as joint rotations) which we call spatial basis (SB) and second, we impose a parametric model to the spatio-temporal (ST) profiles of each SB vector. We showcase the method by applying SB vectors and ST parameters to study vertical jumps of young adults (YAD) typically developing (TD) children and children with Developmental Coordination Disorder (DCD) obtained with motion capture. We also go over a number of other topics related with compositionality: we introduce a top-down system of tool-use primitives based on kinematic events between body parts and objects. The kinematic basis of these events is inspired by the hand-to-object velocity signature reported by movement psychologists in the 1980's. We discuss the need for custom-

made movement measurement strategies to study action primitives on some target populations, for example infants. Having the right tools to record infant movement would be of help, for example, to research in Autism Spectrum Disorder (ASD) where early sensorimotor abnormalities were shown to be linked to future diagnoses of ASD and the development of the typical social traits ASD is mostly known for. We continue the discussion on infant movement measurement where we present an alternative way of processing movement data by using textual descriptions as replacements to the actual movement signals observed in infant behavioral trials. We explore the fact that these clinical descriptions are freely available as a byproduct of the diagnosis process itself. A typical/atypical classification experiment shows that, at the level of sentences, traditionally used text features in Natural Language Processing such as term frequencies and TF-IDF computed from unigrams and bigrams can be potentially helpful.

In the end, we sketch a conceptual, compositional model of action generation based on exploratory results on the jump data, according to which the presence of disorders would be related not to differences in key postures, but in how they are controlled throughout execution. We next discuss the nature of action and actor information representation by analyzing a second dataset with arm-only data (bimanual coordination and object manipulations) with more target populations than in the jump dataset: TD and DCD children, YAD and seniors with and without Parkinson's Disease (PD). Multiple group analyses on dofs coupled with explained variances at SB representations suggest that the cost of representing a task as performed by an actor may be equivalent to the cost of representing the task alone.

Plus, group discriminating information appears to be more compressed than task-only discriminating information, and because this compression happens at the top spatial bases, we conjecture that groups may be recognized faster than tasks.

# ACTION COMPOSITIONALITY WITH FOCUS ON NEURODEVELOPMENTAL DISORDERS

by

Leonardo Claudino

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2016

Advisory Committee:
Professor Yiannis Aloimonos, Chair/Advisor
Professor Jane Clark
Dr. Cornelia Fermuller
Professor James Reggia
Professor Nicholas Roussopoulos

# Dedication

To my Wife Aninha, my love, my friend and my partner in everything.

To my Son Eric, who filled my life with joy and meaning.

To Mom and Dad for all their love and hard work.

To my Brother Rafael, for the friendship and support.

# Acknowledgments

First and foremost comes my family: I would like to thank my Wife Aninha, for holding my hand through this entire process and for making this enterprise hers as well. I want to thank my Mother for the love, friendship and for lessons of strength and resilience. Thank you, my dear Son Eric, for being my everyday's sunshine. And to Luis and Irani Bravo, our family in the USA.

Next, I would like to thank Yiannis and Cornelia for the opportunity, as well as the Members of my Committee for their participation and comments. Thank you Jane, for sharing your views and expertise, and for the many pleasant meetings.

I would like to thank Professors Philip Resnik and Jordan Boyd-Graber for the numerous work opportunities and for serving as my role models in academia.

Special thanks to my landlords and friends Allen and Liza Linder. And to these great ladies in the CS Department and UMIACS: Jenny Story, Fatima Bangura, Jodie Gray, Janice Perrone, and Dr. Michelle Hugue. Thank you for having my back so many times, Meesh! You too, Jenny!

Finally, I want to thank the CS and UMIACS staff members for their availability and patience. And to my friends at Starbucks at the University of Maryland, where I did most of my work in the last 3 years and met the most incredible characters.

# Table of Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| ACSM | Articulated Cloud System Model |
| AD | Adult or Autism Disorder |
| AIP | Anterior Interparietal Area |
| AOSI | Autism Observation Scale for Infants |
| APA | American Psychiatric Association |
| AS | Asperger's Syndrome or Activity Score |
| ATD | Atypically developing |
| ASD | Autism Spectrum Disorder |
| BVH | Biovision Hierarchy (format) |
| CDC | Center for Disease Control |
| CDD | Childhood Disintegrative Disorder |
| CNS | Central Nervous System |
| CRCNS | Clinical Research Centers for NeoNatal Seizures |
| DAG | Dependency Acyclic Graph |
| DCD | Developmental Coordination Disorder |
| DOF or dof | Degrees of freedom |
| DS | Dynamic symmetry |
| EMG | Electromyography |
| FFT | Fast Fourier Transform |
| fMRI | Functional Magnetic Resonance Imaging |
| GPDM | Gaussian Process Dynamic Model |
| GPLVM | Gaussian Process Latent Variable Model |
| HINE | Hammersmith Infant Neurological Examination |
| IFFT | Inverse Fast Fourier Transform |

| | |
|---|---|
| LIWC | Linguistic Inquiry and Word Count |
| LLS | Linear Least Squares |
| MRI | Magnetic Resonance Imaging |
| MABC | Movement Assessment Battery for Children |
| NLLS | Nonlinear Least Squares |
| NLP | Natural Language Processing |
| PCA | Principal Component Analysis |
| PDD-NOS | Pervasive Developmental Disorder Not Otherwise Specified |
| PLD or pld | Point-Light Display |
| PTSD | Post-Traumatic Stress Disorder |
| RS | Rett Syndrome |
| SB | Spatial basis |
| SC | Shape Context |
| SIDS | Sudden Infant Death Syndrome |
| SS | Static Symmetry |
| ST | Spatio-temporal |
| STS | Superior Temporal Sulcus |
| SVD | Singular Value Decomposition |
| SVM | Support Vector Machine |
| TD | Typically developing |
| TF-IDF | Term Frequency-Inverse Document Frequency |
| TVMS | Time-Varying Muscle Synergies |
| UCM | Uncontrolled Manifold (analysis) |
| VA | Vector Analysis |
| VARPRO | Variable Projection |

# Chapter 1: Introduction

## 1.1 Action primitives

In the context of movement generation, the term *action primitives* denotes a hypothetical set of pre-existing modules of effector activation that would be controlled and co-ordinated by the central nervous system (CNS) to produce action. Many believe this is the way the brain cuts down dimensionality when dealing with multiple degrees of freedom in space and time, the so called "Degrees of freedom (DOF) problem" and it came out of the first round of investigations of Bernstein's work in control and co-ordination, as once posed by Turvey [3]. This problem has been recently revisited by Latash et al. [4, 5] who discuss the related "principle of abundance", referring to the fact that a task demands less degrees of freedom than what is available to be controlled. See Flash et al. [6] for a summary of findings around the nature of motor primitives at behavioral, muscle, neural, and computational levels.

Previous electrophysiological experiments in spinalized animals triggered great excitement when they presented strong evidence supporting the existence of basic modules of movement – also referred to as *motor synergies* – that would be additively combined to produce behavior [7]. As a result, many linear models were

proposed on different levels of representations in the motor space and used to reproduce and analyze experimental data from vertebrate and invertebrate data, namely: spinal force-fields [8, 9], time-varying muscle forces [10–12], or joint-angle configurations [13].

In particular, compact representations of movement have also been pursued by vision psychologists while trying to computationally model the visual phenomenon referred to as *biological motion perception* – a term coined to express the ability of humans to perceive moving dots from point-light displays as coherent articulated rigid bodies that give rise to the perception of classes of activities [14–19]. Of particular relevance, Troje [20] has offered a computational method that produces walking patterns and it is able to discriminate between male and female walks from point-light displays coming from 3-D motion capture positional data. He modeled the temporal occurrences of 4 walking eigenpostures with a family of sine functions, for each he determined the a single fundamental frequency and relative phases.

It is not surprising that some of the mathematical models used to find primitives from motor and visual signals are very similar: in [3], Turvey brought up the issue of "simultaneous organization of afferentiation and efferentiation", sugesting that we should perhaps think of action primitives to lie somewhere between vision and movement. Also supporting this view, in [21], Jeannerod argues in favor of a simulation hypothesis to explain action representations, based on several results in motor psychology and neurophysiology in the last 20-30 years. According to his account [22], perceived actions would be slight variations[1] of executed ones, in

[1]During simulation, activations of most motor areas are weaker, motor output is inhibited and

they would share the same temporal, programmatic and biomechanic constraints but would be suppressed: "If motor cognition is based on the simulation of our own actions (...) then we can develop the idea that perceiving and producing actions are the two faces of the same process".

This thinking is in line with the so-called *direct matching hypothesis*, a product of a number of findings in experimental neuroscience that described areas in the brain that would link visual stimuli to purposive movement. For instance, in [23] authors commented on previous findings of certain neurons in F5 (ventral pre-motor cortex) of macaque monkeys that would fire to visual stimuli of hand-object interactions, and spoke of visuomotor transformations mediated by these neurons. According to their account, when understanding a manipulation, properties of the object should be extracted (size, orientation, graspability) and motor schemas (sensorimotor control plans) supposedly encoded in F5 would be retrieved. Stronger evidence to direct matching came later on, when a subset of neurons recorded in the same area F5 was also found to be connected to goal-directed movements, but these would fire both when the primates experienced and when they performed actions involving a food object, and the usage of hands with various grip types, or the mouth, the reason why they were called *mirror neurons*. More specifically, 92 of the 532 units recorded presented such properties, out of which 29 were found to be "strictly congruent", that is, they would fire only when the action observed/performed in the exact same manner (grip). Authors discussed striking similarities of these units with neurons in the superior temporal sulcus (STS) and, among other things, speculated that

there is lack sensory reafferences.

F5 and STS could belong to independent but complementary roles, as if "... the superior temporal sulcus is the semantic representation of hand-object interactions, while F5 is the pragmatic one", in-line with the Goodale and Milner's former view of separate pathways of perception and action [24] (recently reviewed in [25]). Authors also acknowledged the possibility of F5 carrying some sort of "motor vocabulary" and being part of a visuomotor matching process, as earlier suggested.

According to the FARS model [26], in the context of tool use, these visuo-motor transformations would be mostly centered in parsing object attachments or affordances. Neural networks of the anterior intra-parietal area (AIP) in the parietal cortex were hypothesized to extract the attachment regions out of the shape, size and orientation cues that come from the visual cortex. Properties of attachments are further converted into potential grasp plans, which are forwarded to the pre-motor cortex, where F5 neurons select the most appropriate motor programs and connects to the primary motor cortex (F1) that will recruit the proper motor synergies to produce an overt grasp movement. This view that pictorial representation of object parts would translate into primitive or canonical movements served as inspiration to a number of systems in the fields of Robotics and Computer Vision [27–31].

From a psychophysical perspective, Flanagan and Johansson [32] have also provided evidence of the visuomotor nature of action representation: first, they had a set of subjects to both stack up 3 blocks from one side to another of a horizontal work surface and to watch it being done by others. Then, additional subjects went through the same experiment, but with no visual feedback of the actor's hands when observing the action. From the first round, they have noticed that actors

and observers tend to fixate at the same spatial locations while performing and observing the same task being performed, respectively, and these locations are most often the contact zones rather than the hands and the blocks. Moreover, both their gaze seemed "proactive", in the sense that their eyes would land at those zones before the hand would, and this trend was shown to increase with trial duration. In contrast, subjects of the second round, deprived of hand feedback, relied on tracking the objects to follow the task, resulting in a reactive behavior instead of an anticipatory one. To give an idea, in round 1, the eyes would exit contact sites on average 72 ms ahead of the hand, while in round 2 they would leave these sites about 200 ms after the hand had left. According to the authors, the focus on the contact zones rather than hands and blocks, the increasing predictive behavior and the fact that the observer's oculomotor system seems to tightly reflect the actor's would all support a direct matching view of action, with the mirror system working as a living vocabulary of primitives.

The direct matching and the mirror system hypotheses have been believed to be the basis of imitation, a basic social feature of primate behavior. However, imitation, in the sense of what Jeannerod calls "true imitation", as opposed to bare mimicry (observed in humans as early as 42 minutes after their birth [33]), is a very complex behavior that mingles perception, action and memory. Indeed, imitation demands the individual to (1) properly grasp the goal of an action, (2) judge the used form, (3) eventually figure the actor's intention, and also to (4) reenact it with whatever degrees-of-freedom it has available. This becomes even more complex when we note that these processes can take place both on-line and off-line (with

the use of memory but no visual cues). It is therefore unlikely that direct matching alone would be able to explain such sophisticated cognition, with so many levels of "analogy".

The quest for action primitives poses a series of scientific challenges. First, there is the question of what is the *right domain of investigation*: should one focus on full-body or manipulation tasks? Either way, would the dimensionality reduction principles that give rise to full-body primitives be the same as the one who results in tool use primitives? The second (and related) difficulty is to choose the *proper raw data* to work with: the studies described in the previous section went from single neuron spike rates to EMG and 2D and 3D joint angle rotations. Should tool/object data be included? Alternatively, what is the *right level to probe*: neurons (brain cells, brain tissue, blood oxygenation levels), muscles, joints, gaze, objects? Next, should we hypothesize and test a certain set of primitives (top-down) or should we "mine it out of the data" (bottom-up)? Plus, what are the *right computational techniques* that should be used to group these high-dimensional action data? How much do tasks share primitives? To what extent do neurotypicals and atypicals differ in terms of how they recruit, control and coordinate action primitives?

This research tried to address a subset of these challenges through a number of case studies: in Chapter 2 we discuss a bottom-up approach to find full-body postural primitives as a set of key postures, that is, vectors corresponding to key relationships among degrees of freedom (like angles between body parts) which we call spatial basis (SB) and second, we impose a parametric model to the spatio-temporal (ST) profiles of each SB vector. These two steps constitute the SB-ST

decomposition of an action: SB vectors represent the key postures, their ST profiles represent trajectories of these postures and ST parameters express how these postures are being controlled and coordinated. SB-ST shares elements in common with computational models of motor synergies and biological motion perception, and it relates to human manifold models that are popular in machine learning. We showcase the method by applying SB vectors and ST parameters to study vertical jumps of adults, typically developing children and children with Developmental Coordination Disorder obtained with motion capture. We will come back to SB-ST in Chapter 6 where we sketch an action generation model based on SB and ST estimated parameters, and on the insights obtained from the jump experiment. In that chapter, we also look at how spatial bases seem to be encoding information needed to recognize populations and tasks, this time using data from tasks involving bimanual coordination and object manipulation. In Chapter 3, we introduce a top-down system of tool-use primitives based on kinematic events between body parts and objects. The kinematic basis of these events is inspired by the velocity signature of hand-to-object transportation curves.

## 1.2 Measuring and analyzing primitives: the study of neurodevelopmental disorders

The discussion in the first two chapters assume the existence of the proper means to obtain raw movement data. However, some populations might require custom-made measurement strategies; we support this view by exposing the problem of motion

capture of infants. Our interest in infant populations arise from the fact that these individuals are minimally affected by cultural background and display the fastest rates of evolving cognition and physique, opening possibilities to longitudinal but relatively short-term research. Having the right tools to record infant movement would be of help, for example, to research in Autism Spectrum Disorder (ASD) where early sensorimotor abnormalities were shown to be linked to a future diagnosis of ASD and the development of the typical social traits ASD is mostly known for. That said, in Chapter 4 we provide evidence that, as opposed to the current practice, studies on infant behavior would demand non-invasive instrumentation to measure movement, so the right paradigm to obtain the data will most likely depend on computer vision based pose estimation. We propose the use of canonical postures as an implementation of the principle of stability noted by developmental psychologists, and exemplify how these postures and age-related data could be used to potentially improve existing pose estimation systems. We also show preliminary results suggesting that canonical postures may be recognized using global, low-level contour features augmented by mid-level features like elongatedeness; these results are consistent with previous work in infant pose estimation using pressure-based sensors. We continue the discussion on infant movement measurement in Chapter 5 where we present an alternative way of processing movement by using textual descriptions as replacements to the actual movement signals observed in infant behavioral trials, by noting that these descriptions are freely available as a byproduct of the diagnosis process itself. A typical/atypical classification experiment shows that, at the level of sentences, traditionally used text features in Natural Language Processing

such as term frequencies and TF-IDF computed from unigrams and bigrams can be potentially helpful.

# Chapter 2: The SB-ST Decomposition in the Study of Developmental Coordination Disorder

## 2.1 Introduction

The means to obtain movement data are getting cheaper, more diverse and achieving higher throughput. These data are high-dimensional and highly redundant, both at the level of degrees of freedom (or dofs, for example, angles between body parts) and in terms of how often spatial arrangements of these dofs (postures) are recruited in the timeline of the action. It is thus very hard to analyze raw movement data, and in practice movement analysts will discard dofs, look at a single dof at a time, or assume the existence of a single external variable (an unknown direct or indirect function of dofs) being controlled during the action. A typical approach is uncontrolled manifold analysis (UCM) [4, 5], a framework designed to investigate whether a certain performance variable is being controlled during movement by factoring the variance (or covariance) of one (or more) elemental variables (or dofs) at different instants of a task performance into two manifolds: one that is tangent to the trajectory ($V_{UCM}$) and another that is orthogonal to it ($V_{ORT}$). When most of the variance projects onto $V_{UCM}$, the performance variable is expected to change little in face of

flexible configurations of the considered dofs. For example, when studying vertical jumps, one could use the center of mass trajectory as the performance variable and, within the UCM framework, verify how stable or relatively invariant that variable is when typically (TD) or atypically (ATD) developing children perform, given a number of jumps obtained from both populations.

Despite being a great tool to study the stability of performance variables, UCM was not designed to identify ensembles of dofs and/or parametrize its relative timings, plus it will often rely on multiple trials to calculate manifolds. With that in mind we propose an alternative representation obtained by decomposing a single trial action matrix $\boldsymbol{Y}_{T\times J}$ ($T$ = time instants, $J$ = dofs) in two decoupled steps: first, we discover a set of vectors spanning the $J$ space of $\boldsymbol{Y}$, which we call spatial basis (SB) because they are supposed to represent key relationships between dofs, or key postures. Second, we impose a parametric model to the spatio-temporal (ST) profiles of each SB vector. Spatio-temporal profiles of SB vectors are 1-D signals expressing their temporal correlation with $\boldsymbol{Y}$; a high correlation of a vector at time $t$ indicates strong recruitment or activation of the vector at that time. These two steps constitute the SB-ST decomposition of an action: SB vectors represent the key postures, their ST profiles represent trajectories of these postures, and ST parameters express how much (control) and when (coordination) these postures are being recruited in each case. Going back to the jump example, we can now use SB-ST to compare jumps of TD and ATD children simultaneously in terms of dof recruitment, trajectories, control and coordination.

Dimensionality reduction of movement data has been studied in the context

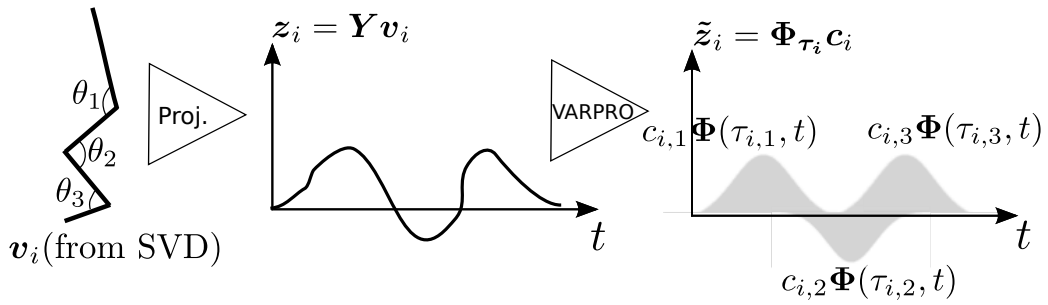



**Figure 2.1:** SB-ST action decomposition. In this example, $J$-dimensional spatial basis vector $\boldsymbol{v}_i$ encodes a linear combination of joint angles $\theta_1$, $\theta_2$ and $\theta_3$ computed with SVD, as shown by the leftmost figure. The projection $\boldsymbol{z}_i = \boldsymbol{Y}\boldsymbol{v}_i$ of action matrix $\boldsymbol{Y}_{T\times J}$ onto $\boldsymbol{v}_i$ results into an often smooth temporal series of correlations that represents the activity of that particular spatial arrangement (posture) along the timeline of the action (center figure). We use VARPRO to produce a compact parametric representation for this temporal behavior by fitting a mixture of $\tilde{\boldsymbol{z}}_i = \boldsymbol{\Phi}_{\boldsymbol{\tau_i}}\boldsymbol{c}_i$ to $\boldsymbol{z}_i$ (right figure) which results in parameter vectors $\boldsymbol{\tau}_i = \{\tau_{i,1}, \tau_{i,2}, \tau_{i,3}\}$ and $\boldsymbol{c}_i = \{c_{i,1}, c_{i,2}, c_{i,3}\}$. An action matrix is therefore fully characterized by each spatial basis vector $\boldsymbol{v}_i$ and corresponding set of spatio-temporal parameter vectors $\boldsymbol{\tau}_i$ and $\boldsymbol{c}_i$.

of different disciplines: for example, in motor neuroscience, the time-varying muscle synergy (TVMS) model was originally designed to study laboratory data from frog jumps [12] and walking data from humans [34]. In psychology of vision, the locomotory model of Troje [20] was used to characterize point-light displays[1] of walkers, and for the synthesis of new walking displays. In machine learning, human manifold models like the GPLVM family [35, 36] were shown to perform very well in tasks such as tracking and pose estimation. Like SB-ST, the first two approaches produce representations that decouple space and time. The latter reduces dimensionality in both spaces simultaneously.

Our contributions are threefold: (1) we present a very unique application of dimensionality reduction: the analysis of motion capture data of vertical jumps performed by adults, TD children and children with Developmental Coordination Disorder (DCD); there is an increasing demand for this kind of study, in response

[1] A typical point-light display is a video with an actor dressed up with dark clothes and white spherical markers in a way that only the markers are visible. The result is a moving point cloud.

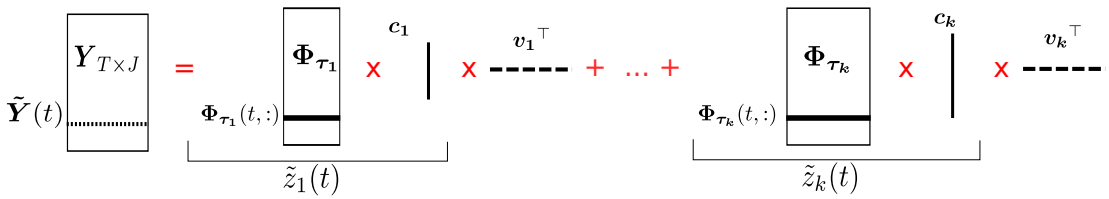



**Figure 2.2:** Generating postures with SB-ST. A posture $\boldsymbol{Y}(t)$ results of a linear combination of spatial basis vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \ \ldots \ \boldsymbol{v}_k$ (dashed lines), as in Equation 2.3. Coefficients $\tilde{z}_i(t)$ of each vector $\boldsymbol{v}_i$ are the product of the $t$-th time row of its spatio-temporal matrix $\boldsymbol{\Phi}_{\boldsymbol{\tau}_i}$ and respective linear parameter vector $\boldsymbol{c}_i$ (solid lines).

to *recent scientific findings correlating movement abnormalities in childhood and the later development of neuro-developmental disorders* [37]. Using the jump data we: (2) introduce a framework to study actions and actors based on SB vectors and ST parameters and present evidence that the major differences between TD and DCD jumps are more likely to reside in the spatio-temporal facet of the behavior, plus (3) evaluate and compare SB-ST with alternative techniques. For example, as opposed to SB-ST, TVMS does not work well on individual trials, and both Troje's method and GPLVM miss local temporal features that are crucial to the study of behavior [34].

This paper is organized as follows: we begin by introducing SB-ST (Sec. 2.2) followed by previous work (Sec. 2.3), to facilitate comparing the proposed method with alternative techniques by having presented its structure first. Next, we discuss our experiments and conclusions.

## 2.2 The SB-ST action decomposition

SB-ST is computed in 2 major steps: (1) given an input action matrix, we first we extract spatial basis (SB) vectors and compute their spatio-temporal (ST) profiles,

and (2) we fit a parametric model to the ST profiles of each vector, as follows. See Fig. 2.1 for an overview of the method.

### 2.2.1 Spatial basis SB and spatio-temporal profiles

Let $\boldsymbol{Y}_{T\times J}$ be a multi-dimensional action signal, for example, a $T$-length sequence of $J$ degrees of freedom (dofs). The $k$-th order approximation of that signal by SVD, in matrix notation is:

$$\hat{\boldsymbol{Y}}_{T\times J} = \boldsymbol{z}_1\boldsymbol{v}_1^\top + \boldsymbol{z}_2\boldsymbol{v}_2^\top + \ldots + \boldsymbol{z}_k\boldsymbol{v}_k^\top, \tag{2.1}$$

where $\boldsymbol{v}_i$ is one of the top $k$ right singular vectors of $\boldsymbol{Y}$, therefore spanning the column space of that matrix, and projection $\boldsymbol{z}_i = \boldsymbol{Y}\boldsymbol{v}_i$ corresponds to the *spatio-temporal profile* of $\boldsymbol{v}_i$, that is a one-dimensional time series that expresses the correlations of the particular spatial configuration represented by $\boldsymbol{v}_i$ along the timeline of the action[2]. For each $i$, let $\{\boldsymbol{\Phi}(\tau_{i,j}, t) : j = 1 \ldots N_i\}$ be a family of $N_i$ Gaussians with fixed standard deviations and $\boldsymbol{\Phi}_{\boldsymbol{\tau_i}}$ to be the corresponding $T \times N_i$ matrix such that each function is evaluated at $T$ instants and it becomes a column of that matrix. *We will parametrize* $\boldsymbol{z}_i$ *by fitting a linear combination of the columns of* $\boldsymbol{\Phi}_{\boldsymbol{\tau_i}}$ *with linear parameters* $\boldsymbol{c}_i = \{c_{i,1}, c_{i,2} \ \ldots \ c_{i,N_i}\}$:

$$\tilde{\boldsymbol{Y}}_{T\times J} = \underbrace{(\boldsymbol{\Phi}_{\boldsymbol{\tau_1}}\boldsymbol{c}_1)}_{\tilde{\boldsymbol{z}}_1}\boldsymbol{v}_1^\top + \underbrace{(\boldsymbol{\Phi}_{\boldsymbol{\tau_2}}\boldsymbol{c}_2)}_{\tilde{\boldsymbol{z}}_2}\boldsymbol{v}_2^\top + \ldots + \underbrace{(\boldsymbol{\Phi}_{\boldsymbol{\tau_k}}\boldsymbol{c}_k)}_{\tilde{\boldsymbol{z}}_k}\boldsymbol{v}_k^\top, \tag{2.2}$$

[2]Note that, for right singular vector $\boldsymbol{v}_i$, $\boldsymbol{z}_i = \boldsymbol{Y}\boldsymbol{v}_i = \sigma_i\boldsymbol{u}_i$, with $\sigma_i$ being the $i$-th singular value and $\boldsymbol{u}_i$ the $i$-th left singular vector. We chose to use $\boldsymbol{Y}\boldsymbol{v}_i$ rather than $\boldsymbol{u}_i$ just to emphasize that vector $\boldsymbol{z}_i$ expresses a time series of correlations between the data matrix $\boldsymbol{Y}$ and the particular $\boldsymbol{v}_i$.

**Algorithm 1:** SB_ST($\boldsymbol{Y}$, $k$, $[N_{1\ldots k}]$)

Compute $[\mathbf{U}, \boldsymbol{\Sigma}, \mathbf{V}]$ = SVD of $\boldsymbol{Y}$
**for** $i = 1$ to the first $k$ columns $\boldsymbol{v}_i$ of $\mathbf{V}$ **do**
  Form $\boldsymbol{z}_i = \boldsymbol{Y}\boldsymbol{v}_i$
  Form approximation $\tilde{\boldsymbol{z}}_{\boldsymbol{i}}$ by:
  1. running NLLS solver that calls $[\boldsymbol{\tau}_i, \tilde{\boldsymbol{c}}_i, \boldsymbol{r}, \boldsymbol{J}]$ = VARPRO_loop($\boldsymbol{\tau}_i$, $\boldsymbol{z}_i$, $[N_{1\ldots k}]$) w/random initial $\boldsymbol{\tau}_i$ (solver minim. $\boldsymbol{r}^2$ using Jacob. $\boldsymbol{J}$),
  2. recalculating $\boldsymbol{\Phi}_{\boldsymbol{\tau_i}}$ from optimal $\boldsymbol{\tau}_i$ and fixed stds,
  3. making $\tilde{\boldsymbol{z}}_i = \boldsymbol{\Phi}_{\boldsymbol{\tau_i}}\boldsymbol{c}_i$ using optimal $\boldsymbol{c}_i$
  Update approximation $\tilde{\boldsymbol{Y}}_{T\times J} \leftarrow \tilde{\boldsymbol{Y}}_{T\times J} + \tilde{\boldsymbol{z}}_i\boldsymbol{v}_i^\top$
**end for**
Return $\boldsymbol{v}_i$, $\boldsymbol{c}_i$, $\boldsymbol{\tau}_i$ ($i = 1\ \ldots\ k$) and $\tilde{\boldsymbol{Y}}$

and we have $\tilde{\boldsymbol{z}}_i = \boldsymbol{\Phi}_{\boldsymbol{\tau_i}}\boldsymbol{c}_i$. Equivalently, the posture produced by the model at time $t$ is:

$$\tilde{\boldsymbol{Y}}(t) = \tilde{z}_1(t)\boldsymbol{v}_1^\top + \tilde{z}_2(t)\boldsymbol{v}_2^\top + \ldots + \tilde{z}_k(t)\boldsymbol{v}_k^\top, \tag{2.3}$$

where:

$$\tilde{z}_i(t) = c_{i,1}\boldsymbol{\Phi}(\tau_{i,1}, t) + c_{i,2}\boldsymbol{\Phi}(\tau_{i,2}, t) + \ldots + c_{i,N_i}\boldsymbol{\Phi}(\tau_{i,N_i}, t). \tag{2.4}$$

See Fig. 2.2 for a schematic view. Vector $\boldsymbol{v}_i$ corresponds to the $i$-th *spatial basis* (SB) vector of $\boldsymbol{Y}$ or SB-$i$. Each $\boldsymbol{v}_i$ expresses relationships between dofs (principal postures). Basis functions $\boldsymbol{\Phi}(\tau_{i,j}, t)$ (and, equivalently, its matrix version $\boldsymbol{\Phi}_{\boldsymbol{\tau_i}}$) together with the mean vector $\boldsymbol{\tau}_i$ and the linear parameter vector $\boldsymbol{c}_i$ constitute what we call the $i$-th *spatio-temporal representation* (ST) of $\boldsymbol{Y}$ or ST-$i$. These parameters map local temporal patterns and describe how a spatial vector $\boldsymbol{v}_i$ is controlled and coordinated.

### 2.2.2 Spatio-temporal representations ST

Because we made $\boldsymbol{\Phi}(\tau_{i,j}, t)$ a family of single-parameter Gaussians, this problem turns out to be a separable least-squares regression problem, which allows us to

**Algorithm 2:** VARPRO_loop($\boldsymbol{\tau}_i$, $\boldsymbol{z}_i$, $[N_{1\ldots k}]$)

Compute matrix $\boldsymbol{\Phi}_{\boldsymbol{\tau_i}}$ from $\boldsymbol{\tau}_i$ and fixed stds
Compute $[\mathbf{U}, \boldsymbol{\Sigma}, \mathbf{V}]$ = SVD of $\boldsymbol{\Phi}_{\boldsymbol{\tau_i}}$
Make $\tilde{\boldsymbol{c}}_i = \boldsymbol{V}\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{U}^\top \boldsymbol{z}_i$
Compute current $\tilde{\boldsymbol{z}}_{\boldsymbol{i}} = \boldsymbol{\Phi}_{\boldsymbol{\tau_i}}\tilde{\boldsymbol{c}}_i$ and residual $\boldsymbol{r} = \boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i$
**for** $j = 1$ to $N_i$ Gaussians of $\boldsymbol{\Phi}_{\boldsymbol{\tau_i}}$ **do**
  Form matrix with partial derivatives $\boldsymbol{D}_j = \frac{\partial \boldsymbol{\Phi}(\tau_{i,j}, t)}{\partial \tau_{i,j}}$
  Make $\boldsymbol{a}_j = \boldsymbol{D}_j\tilde{\boldsymbol{c}}_i - \boldsymbol{U}(\boldsymbol{U}^\top(\boldsymbol{D}_j\tilde{\boldsymbol{c}}_i))$ and $\boldsymbol{b}_j = \boldsymbol{U}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{V}^\top(\boldsymbol{D}_j^\top \boldsymbol{r})))$
  Add $\boldsymbol{a}_j$ and $\boldsymbol{b}_j$ and form the $j$-th column of $\boldsymbol{J}$ as in Eqs. 2.9 to 2.11
**end for**
Return $\boldsymbol{\tau}_i$, $\tilde{\boldsymbol{c}}_i$, $\boldsymbol{r}$, $\boldsymbol{J}$

solve for $\boldsymbol{\tau}_i$ and $\boldsymbol{c}_i$ using variable projection (VARPRO) [38]. The method exploits the linear substructure of this particular case of nonlinear least squares (NLLS) regression: if you fix the set of non-linear parameters $\boldsymbol{\tau}_i$, the problem turns out to be linear in $\boldsymbol{c}_i$ and can be solved for the latter using linear least squares (LLS). In other words, parameter $\boldsymbol{c}_i$ becomes a function of parameters $\boldsymbol{\tau}_i$ and so, instead of solving:

$$\min_{\boldsymbol{\tau}_i, \boldsymbol{c}_i} ||\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i(\boldsymbol{\tau}_i, \boldsymbol{c}_i)||, \tag{2.5}$$

we solve a less parametrized problem:

$$\min_{\boldsymbol{\tau}_i} ||\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i(\boldsymbol{c}_i(\boldsymbol{\tau}_i))||. \tag{2.6}$$

In the LLS stage, the pseudo-inverse solution for $\boldsymbol{c}_i$ is:

$$\tilde{\boldsymbol{c}}_i = [\boldsymbol{\Phi}_{\boldsymbol{\tau_i}}]^\dagger \boldsymbol{z}_i. \tag{2.7}$$

where $\tilde{\boldsymbol{z}}_i$ is VARPRO's approximation to $\boldsymbol{z}_i = \boldsymbol{Y}\boldsymbol{v}_i$. The solution can be expressed in terms of the SVD of $\boldsymbol{\Phi}_{\boldsymbol{\tau_i}}$:

$$\tilde{\boldsymbol{c}}_i = \boldsymbol{V}\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{-1}}\boldsymbol{U}^\top \boldsymbol{z}_i. \tag{2.8}$$

The LLS solution is then directly embedded in the calculation of the Jacobian of $\tilde{\boldsymbol{z}}_{\boldsymbol{i}}(\boldsymbol{c}_i(\boldsymbol{\tau}_i))$ for the NLLS part of the optimization. The Jacobian can be expressed as a sum of two matrices [39]:

$$\boldsymbol{J} = -(\boldsymbol{A} + \boldsymbol{B}), \tag{2.9}$$

where each of their $N_i$ columns are:

$$\boldsymbol{a}_j = \boldsymbol{D}_j\tilde{\boldsymbol{c}}_i - \boldsymbol{U}(\boldsymbol{U}^\top(\boldsymbol{D}_j\tilde{\boldsymbol{c}}_i)), \tag{2.10}$$

$$\boldsymbol{b}_j = \boldsymbol{U}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{V}^\top(\boldsymbol{D}_j^\top \boldsymbol{r}))). \tag{2.11}$$

where $\boldsymbol{D}_j$ has zeros at all columns but $j$, which will have the partial derivatives of the $j$-th Gaussian $\boldsymbol{\Phi}(\tau_{i,j}, t)$ (or the $j$-th column of matrix $\boldsymbol{\Phi}_{\boldsymbol{\tau_i}}$) with respect to $\tau_{i,j}$, evaluated at all $t$. $\boldsymbol{U}$, $\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{-1}}$ and $\boldsymbol{V}$ are the SVD factors of $\boldsymbol{\Phi}_{\boldsymbol{\tau_i}}$ (Eq. 2.8), and $\boldsymbol{r}$ is the residual $\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i$. Operations were grouped so that only matrix-vector products are required. The full SB-ST decomposition is summarized in Algorithms 1 and 2.

## 2.3 Related Work

### 2.3.1 Motor synergies

In the field of motor neuroscience, many agree that the central nervous system (CNS) organizes behavior by solving a dimensionality reduction problem known as *Bernstein's degrees of freedom (dofs) problem* [3] or how to manage multiple dofs in space and time. One hypothesis is that the CNS controls dofs synergistically as opposed to individually, and that a small number of such *motor synergies* is sufficient [6, 9, 12, 13, 40]. There are various theories around the nature of motor synergies; SB-ST has more aspects in common with computational models involving matrix factorizations, in particular the *time-varying muscle synergies model* (TVMS). Like SB-ST, TVMS also approximates the temporal evolution of a multi-dimensional action vector with $k$ components, which according to our notation would be:

$$\tilde{\boldsymbol{Y}}(t) = z_1 \boldsymbol{v}_1 (t - \tau_1)^\top + z_2 \boldsymbol{v}_2 (t - \tau_2)^\top + \ldots + z_k \boldsymbol{v}_k (t - \tau_k)^\top, \tag{2.12}$$

where the synergy vectors $\boldsymbol{v}_i(t - \tau_i)$ are columns of synergy matrices like the $\boldsymbol{V}_i$'s of Fig.2.3a. These matrices correspond to short-length sequences of postures that are time-shifted by $\tau_i$ and scaled by a fixed value $z_i$ (Fig.2.3b). In contrast, SB vectors $\boldsymbol{v}_i$ correspond to individual postures with *time-varying scaling magnitudes* $z_i(t)$ (compare with Eq. 2.3).

Both methods have in common the use of explicit local parametrization for spatio-temporal profiles; the importance of this choice can be illustrated by the

studies of Ivanenko et al. who use EMG data of human locomotion to look for compositional differences between walking alone and walking combined with voluntary behaviors, such as kicking a ball or overcoming an obstacle [34, 41]. Their results showed that all behaviors agreed upon the same five first profiles – which happened to be very similar to walking – but not upon the sixth, whose synergy activation times varied across behaviors. They proposed an additive model tailored to their observations, where each profile was parametrized by a single Gaussian with standard deviation fixed at 6% of the walking cycle duration. SB-ST, on the other hand, represents these profiles with mixtures and thus allows for more than one activation in the timeline of the action.

There are various theories around the nature of motor synergies, making it an active research topic across many different communities, namely cognitive and humanoid robotics, kinesiology and movement psychology. Models and theories around the nature of synergies have been proposed in terms of spinal force-fields [7–9], time-varying synergies of muscle forces (TVMS) [10–12], joint-angle configurations [13], uncontrolled manifolds [4, 5, 42], nonlinear dynamical systems [43] among others.

### 2.3.2 Biological motion perception

The perception of movement is also believed to be founded on compact representations. In the pioneer experiment of Gunnar Johansson [14] point-light displays[1](pld) of moving actors were presented to completely naïve observers who all reported seeing a walking human, despite the lack of form information in the visual stimuli. He

then proposed *vector analysis* (VA) as a model to explain the phenomenon, in in which a body part is modeled as a pendulum fixed at the body part it attaches to, and the whole stimuli results in a hierarchy of moving pendulums perceived as a single gestalt unit. This study is considered to have started the *biological motion perception* research framework, and the same pld setting has been used to study more complex classes of activities [18]. Of particular interest, Troje [20] proposed a computational method to create and manipulate synthetic plds of walking data. His *eigenpostures*, or the 4 first principal components of a single-walker data matrix, are equivalent to the SB described in the previous section[3]. He modeled the temporal occurrences of the eigenpostures with a family of sine functions. His sine functions are thus a special case of our spatio-temporal representation, because it will only pick up patterns that are global to the whole timeline of the action, and will miss local events that can reveal control and coordination differences across populations.

### 2.3.3 Human manifold models

SB-ST parametrizes trajectories projected on a low-dimensional space, so it also relates to human manifold models. Especially, *Gaussian process latent variable models* (GPLVM) are a family of models that map low-dimensional latent points $\boldsymbol{X}_{T\times\hat{J}}$ to observed data $\boldsymbol{Y}_{T\times J}$ by maximizing the likelihood of $\boldsymbol{Y}$ given $\boldsymbol{X}$ [44], where $\hat{J}$ is the number of latent dimensions. GPLVM extends principal component analysis (PCA) and probabilistic PCA in it allows for non-linear mappings by kernelizing the

[3] Although we have used SVD to create our SB vectors, other factorizations that are not PCA-like could have been used, i. e. eigenpostures are just one possible set of SB vectors.

process covariance function. Faster extensions to GPVLM were shown to improve sparsification in the latent space [35] and to model time-dependency in $\boldsymbol{X}$, like Gaussian Process Dynamic Models (GPDM) [36]. Conceptually, the columns of $\boldsymbol{X}$ produced by GPLVM and the like are analogous to SB-ST's spatio-temporal profiles $\boldsymbol{z}_i = \boldsymbol{Y}\boldsymbol{v}_i$ obtained with SVD (note the similarity of $\boldsymbol{x}_1$ and $\boldsymbol{z}_1$ in Fig. 2.5(top-I)). Regarding GPDM, although it models dynamics, it still produces the same one-to-one $\boldsymbol{X} \rightarrow \boldsymbol{Y}$ kind of mapping as GPLVM, because the model marginalizes out the basis functions $f(\cdot)$ that relate one latent posture to its preceding ones and $g(\cdot)$ that models how latent variables relate to observed postures (Eqs. 1, 2 in [44]). In the end, the method creates a representation that merges space and time within the same manifold, and although this unifying approach has proven adequate for various human movement tasks [45–47], explicit local parametrization of dynamics – in contrast, present in SB-ST – is key to uncover aspects of control and coordination that are not otherwise accessible (see *Motor synergies*, Sec. 2.3).

## 2.4 Experiments and Results

The first goal of our experiments was to examine data reconstruction performance of SB-ST alone and in comparison with (1) methods that, as SB-ST, decouple space and time (Troje and TVMS) and (2) a method that does not (GPLVM). Our second goal was to illustrate how SB-ST can used to provide insights to both actions and actors involved.

Although any kind of action could have been chosen, we looked at *vertical*

*jumps*, a non-trivial behavior that requires strength, coordination and balance. Our 39 participants were first setup with 34 infrared markers and next told[4] to jump as high as possible and try to reach for a visual target, while being recorded by an Optitrack (NaturalPoint Inc.) motion capture system with ten V100 and V100:R2 Flex cameras. The Arena software (included) was used to export its proprietary data format to BVH (Biovision Hierarchy). BVH data were later processed by code[5] written in MATLAB® (versions R2010b and R2011a).

We were able to collect a total of 358 jumps: 9 typically developing female children (TD-F, 98) 6 adult females (AD-F, 61) 10 TD male children (TD-M, 88) 5 adult males (AD-M, 52) and 7 children diagnosed with Developmental Coordination Disorder (DCD, 59) [48]. DCD data were collapsed across gender to make the sample bigger. Children were in the broad age range of 5.1 to 14 years old. Adults (AD) were in their early 20's.TD and DCD groups were both assessed with the MABC (Movement Assessment Battery for Children) test [49], with scores $< 5^{th}$ percentile and $> 29^{th}$ percentile, respectively.

All jump trials were decomposed into a spatial basis of 3 vectors SB-1, SB-2 and SB-3. Regarding ST basis functions, standard deviations were fixed to $\boldsymbol{\sigma}_i = \{1/(2\cdot 1), 1/(2\cdot 2) \ \ldots \ 1/(2\cdot N)\} \times T$, with $T \approx 80$ rows (about .8 seconds) and $J = 6$ columns: left and right hips, knees and ankles. We only used the flexion/extension intersegmental joint angles. Each individual trial was manually segmented by an expert in the vertical jump movement, so that they span the same postural range:

[4]Written informed consent was obtained from all subjects/parents/legal representatives.

[5]Most of the code we used to parse the BVH files is part of Prof. Neil Lawrence's motion capture toolbox, which can be downloaded for free by registering at the author's website. The toolbox is currently hosted at: http://staffwww.dcs.shef.ac.uk/people/N.Lawrence/mocap/.

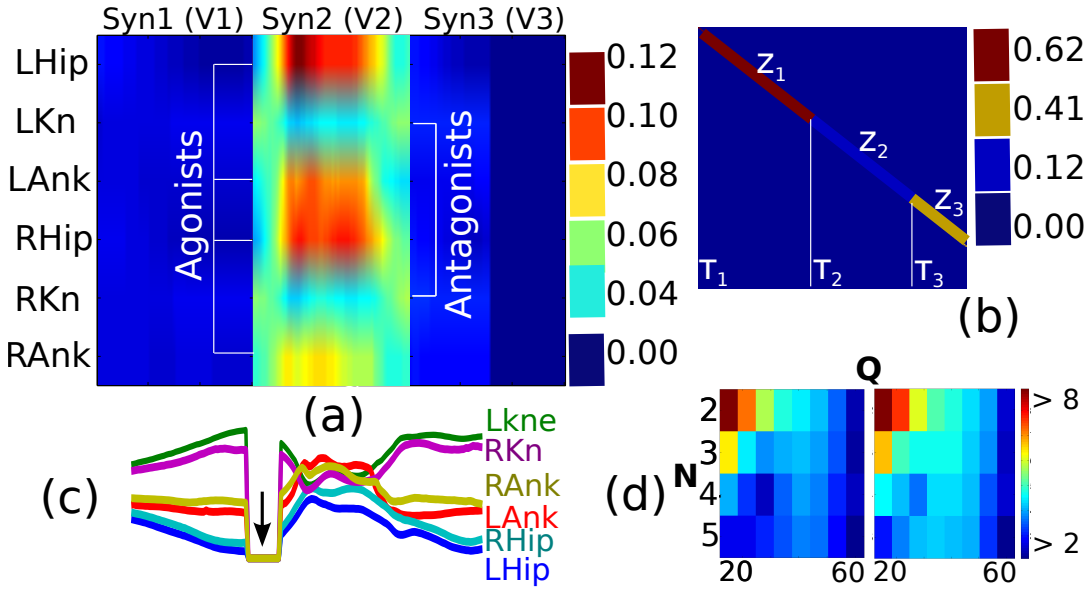



**Figure 2.3:** Single-trial TVMS ($N = 3$ synergies of length $Q = 60$ time units, stopped at 100 iterations or $R^2 \geq 10^{-5}$). (a) $\boldsymbol{V} = [\boldsymbol{V}_1|\boldsymbol{V}_2|\boldsymbol{V}_3]$ is the *synergy matrix*. (b) $\boldsymbol{H}$ (not to scale) shifts and scales all $\boldsymbol{V}_i$ by $\tau_i$ and $z_i$, respectively. (c) Arrow shows zeroed part of signal after reconstruction. (d) As a result, mean and std $R^2$'s for different values of $N$ and $Q$ appear off the usual $[0, 1]$ range.

all poses captured within the initial and final peak knee flexions. Prior to parameter estimation, each $\boldsymbol{z}_i$ was normalized into a unit vector. When using VARPRO, $\boldsymbol{\tau}_i$ was constrained to $[0, 1]$, and no constraints were applied to $\boldsymbol{c}_i$.

Overall, SB-ST achieved an average reconstruction accuracy of $R^2 \geq 0.95$ for all $N$ tried, where $R^2$ is the coefficient of determination (Fig. 2.4 (top)). We will next discuss how SB-ST performed against competing models.

### 2.4.1 Reconstruction: comparing with TVMS

From TVMS results, it appears that a single synergy matrix $\boldsymbol{V}_2$, with hips-ankles as agonists and knees as antagonists would explain most of the jump trial ($\boldsymbol{V}_1$ and $\boldsymbol{V}_3$ are mostly all zeros in Fig.2.3). However, we were often unable to get satisfactory reconstruction of our data using TVMS, as shown in Fig. 2.3c: a significant part of

the signal is not covered by the resulting synergies, resulting in a very low $R^2$. To rule out the cause of the problem to be the poor selections of $N, Q$, we ran TVMS on the whole data using different combinations of these quantities, but the low $R^2$ still persisted, as illustrated by the statistics of $R^2$ shown in Fig. 2.3d. As a result, we discontinued the analysis based on that method. We then conjecture that *the reconstruction problems of TVMS on our data should result from not using more than a single trial to compute synergies and other parameters.* TVMS was designed under the assumption that there exists latent repertoires of synergies/control and coordination parameters that span both multiple behaviors [11] and others that are behavior-dependent [12] and thus constrained their optimization to obtain factors that are faithful to these assumptions; synergies and parameters are supposed to be obtained from minimizing reconstruction errors across several trials. Because SB-ST operates on a per-trial basis, to be able to properly compare the two methods, we had to run TVMS on a single-trial basis.

### 2.4.2 Reconstruction: comparing with Troje-inspired

Fig. 2.4 (top) shows the performances of SB-ST against a *Troje-inspired* decomposition. To clarify: Troje [20] fits the time series of his eigenpostures with a single fundamental harmonic, which he finds sufficient to model locomotion. A natural extension to non-periodic actions like jumps is to select as many harmonics as needed to obtain good approximations. This is what we call a Troje-inspired decomposition. For a certain $N$, the decomposition consisted in selecting the top-$N$ responding

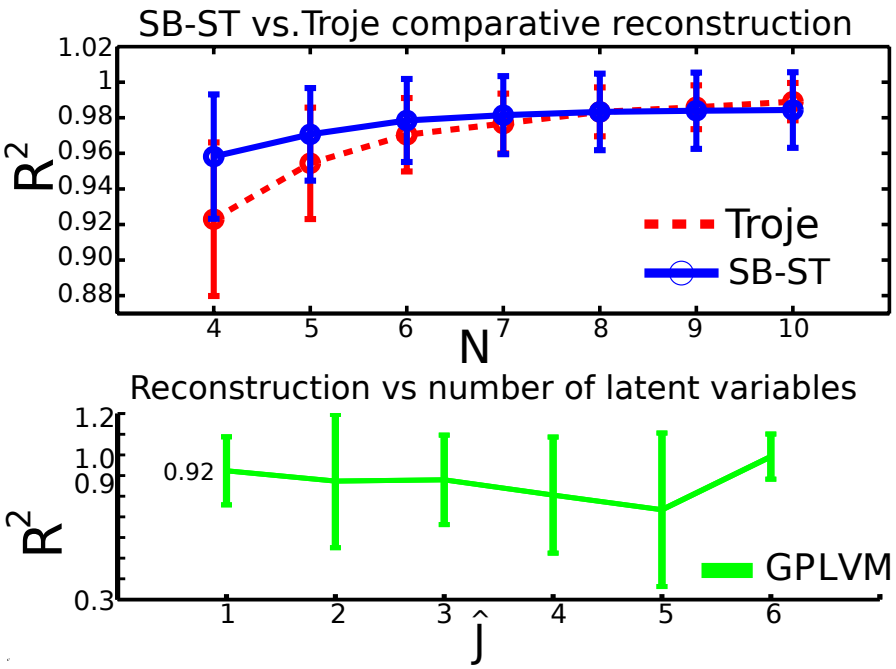



**Figure 2.4:** Quantitative comparison. (Top) $R^2$ versus $N$ for Troje-inspired and SB-ST. Each point = mean $R^2$ $\pm$ std. and the $R^2$ of a trial = average of the per-joint $R^2$'s. SB-ST tops Troje when $N \leq 7$. (Bottom) corresponding $R^2$ of GPLVM for latent vectors $\hat{J} = \{1 \dots J = 6 =$ number of joints$\}$. The number of active points was set to the length of the trial. SB-ST's lowest performance (top, $N = 4$, 42 parameters) tops the best GPLVM ($\hat{J} = 1, 83$ parameters).

Fourier harmonics via FFT of $\boldsymbol{z}_i$ and using only these harmonics to reconstruct the original $\boldsymbol{z}_i$ via IFFT. Note that, for a certain $N$, the number of parameters needed to reconstruct $\boldsymbol{z}_i$ is the same in both cases making these methods comparable: SB-ST fits a mixture of $N$ Gaussians of fixed scales, therefore resulting in $N$ pairs $\boldsymbol{\tau_i}, \boldsymbol{c}_i$ (ST-$i$ parameters) while Troje uses $N$ pairs of Fourier harmonics along with respective responses. Our results show that *SB-ST outperformed Troje-inspired approximation of* $\boldsymbol{z}_i$ *when* $4 \leq N \leq 7$, *which could be considered the range with the best trade-off between number of parameters and reconstruction error* (note the change of slope in both methods when $N$ moves from 3 to 4, as well as the dramatic decrease in $R^2$ variances). Fig. 2.5(top-I) also shows superior qualitative performance: for the same $N = 8$, SB-ST fits the local details of $\boldsymbol{z}_i$ better than its competitor.

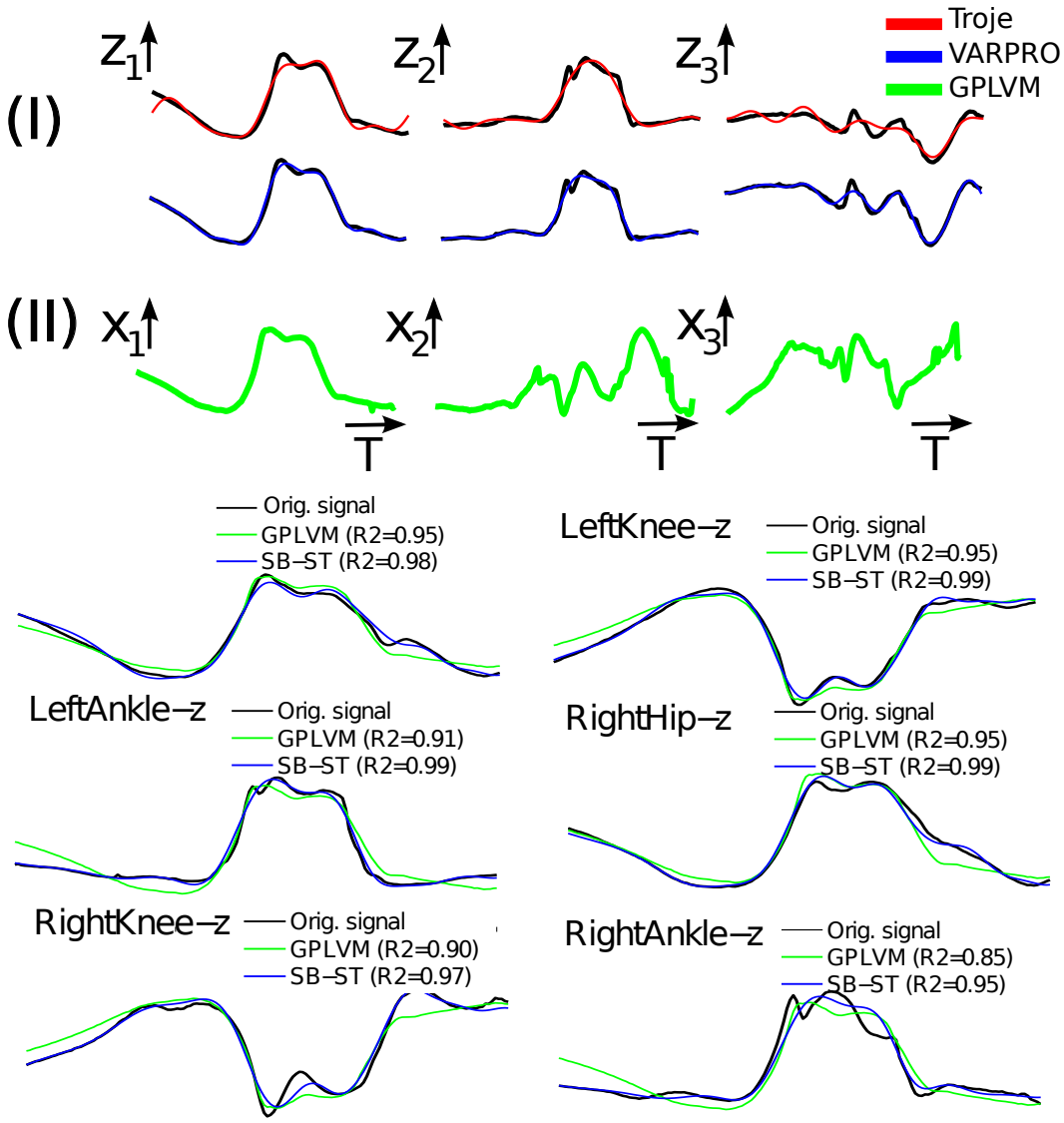



**Figure 2.5:** Qualitative comparison. (Top-I) Fits to $\boldsymbol{z}_1$, $\boldsymbol{z}_2$ and $\boldsymbol{z}_3$ by Troje-inspired and VARPRO for one trial. (Top-II) Corresponding $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ produced by GPLVM. (Bottom) Comparative reconstruction of joint signals with $R^2$.

### 2.4.3 Reconstruction: comparing with GPLVM

To evaluate a GPLVM computed for an action matrix $\boldsymbol{Y}$, we used two steps. (1) With the resulting set of latent vectors $\boldsymbol{X}$ (see Sec. 2.4.3) we pseudo-inverted the

$5^{th}$ Equation of [44] to get the approximation $\tilde{\boldsymbol{Y}}$:

$$\tilde{\boldsymbol{Y}} = \hat{J} \cdot \boldsymbol{X}(\boldsymbol{Y}^\top \boldsymbol{K}_{\boldsymbol{X}}^{-\mathbf{1}} \boldsymbol{X})^\top, \tag{2.13}$$

where, $\boldsymbol{K}_{\boldsymbol{X}}$ is the kernelized covariance matrix, and $\hat{J}$ is the number of columns of $\boldsymbol{X}$ used in the approximation. (2) We computed $R^2$ from $\tilde{\boldsymbol{Y}}$ and $\boldsymbol{Y}$.

Fig. 2.4 (bottom) shows statistics of $R^2$ on the full jump data: from left to right, more parameters are being used to to compute $\tilde{\boldsymbol{Y}}$, that is, the larger $\hat{J}$ the more columns of $\boldsymbol{X}$ are being used to compute $\tilde{\boldsymbol{Y}}$. Note that the best result $R^2 = 0.92$ is still lower than any of the SB-ST scores in Fig. 2.4 (top). Moreover, we note that a GPLVM setup with $\hat{J} = 1$ will result in $\hat{J} \cdot \bar{T} + 3 = 83$ parameters ($\bar{T} = 80$ is approximately the average length of $\boldsymbol{X}$ obtained from our jumps) while an SB-ST configuration with $N = 4$ scoring $R^2 > 0.95$ has exactly $k(J + 2N) = 42$ parameters ($k$ is the number of SB vectors and $2N$ is the number of pairs of ST parameters): *with half as many parameters, SB-ST performs better than GPLVM*, which is also visible from the qualitative example of Fig. 2.4 (bottom) where a SB-ST configuration with $k = 3$ and $N = 10$ (78 parameters) fits the local details of the joint signals better than its competing one-latent vector GPLVM (154 parameters).

But more interesting than the $R^2$ differences between the two methods is that the best GPLVM configuration (other than the full-dimensional, $\hat{J} = J = 6$) is the one with a single latent vector ($\hat{J} = 1$) and that increasing $\hat{J}$ from 2 to 5 (except for $\hat{J} = 3$) makes $R^2$ decrease, which we found somewhat counterintuitive. A possible explanation would be that *one major GPLVM latent variable is enough to represent*

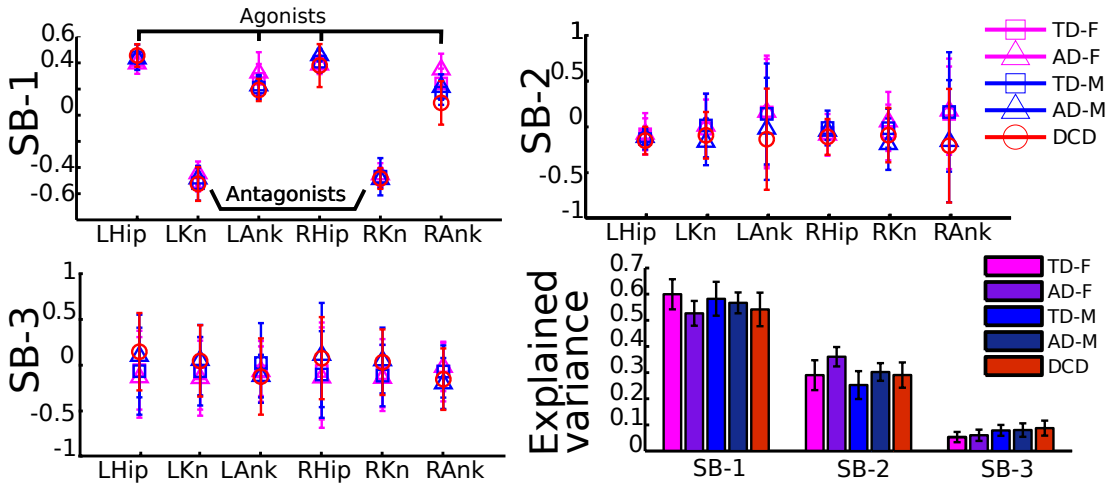



**Figure 2.6:** Mean ± std of SB-1, SB-2 and SB-3 and mean ± std of explained variances per SB vector. Scales were set to accommodate the biggest variances.

*the major features of the jumps* as seen with synergy matrix $\boldsymbol{V}_2$ of TVMS and as will be seen next with SB-1.

### 2.4.4 Data exploration: looking at jumps and jumpers based on the SB-ST parameters

In our second experiment, we used SB-ST to explore our jump data. As in Fig. 2.6, SB-1 coefficient statistics demonstrate that over 50% of the explained variances in the vertical jump come from 2 main groups of rotations: hips-ankles and knees. *SB-1 thus works by clustering leg joints into groups of agonist and antagonist motions, and these distributions seem to generalize across all populations, given the tight clusters.* Fig. 2.6 also reveals that both SB-2 and SB-3 coefficients are almost zero-centered and have high variances, meaning they provide no clear interpretation of the action, so the remaining of the analysis focus on spatio-temporal aspects of SB-1 alone, that is, the statistics of ST-1's $\boldsymbol{\tau}_1 = \{\tau_{1,1} \ \ldots \ \tau_{1,N}\}$ and $\boldsymbol{c}_1 = \{c_{1,1} \ \ldots \ c_{1,N}\}$. We may

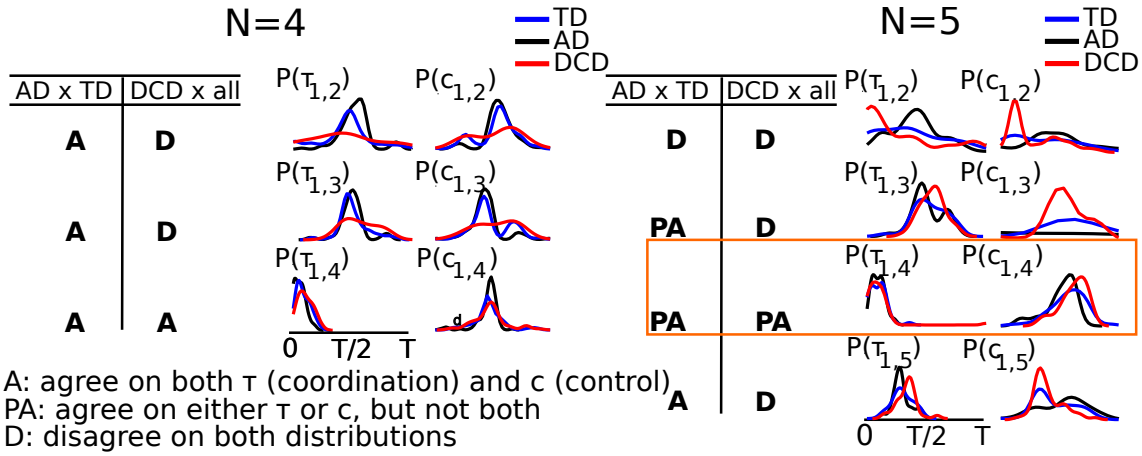



**Figure 2.7:** ST-1 statistics for TD, DCD and AD. Each row displays distributions of $\boldsymbol{\tau}_1$ and $\boldsymbol{c}_1$ for the $2^{nd}$ to $4^{th}$ Gaussians (left, $N = 4$) or $2^{nd}$ to $5^{th}$ Gaussians (right, $N = 5$). Tables point out if populations agree (A), partially agree (PA) or disagree (D) based on the overlap of their curves. Distributions were approximated with MATLAB® `ksdensity()` function, which was set to sample the data range at 50 points and to use a Gaussian kernel for smoothing. Bandwidths were automatically computed by that function, and varied across parameter distributions. The orange selection shows a scenario where all $\tau_{1,4}$ peak at about the same time for all populations, while $c_{1,4}$ do not (see text for details).

also call $\boldsymbol{\tau}_1$ and $\boldsymbol{c}_1$ *coordination* and *control* parameters respectively, because the former "places" each of the Gaussians along the timeline, so they match the local features of the spatio-temporal profile of SB-1, while the latter scales these Gaussians in accordance to the intensities of SB-1 activation. We ran VARPRO with two settings ($N = 4,\ 5$) just to illustrate how the choice of $N$ can affect parameter distributions. After smoothing all distributions, we looked at how jumpers at different developmental stages agreed on ST-1. Data were collapsed across gender to increase the number of subjects per population of interest.

To be considered to agree, two distributions must have similar shape and/or about the same peak abscissa. We judged that, for the present purposes, visual inspection was enough to assess agreement. As seen from Fig. 2.7, frequent agreements between adults (AD) and typically developing children (TD) plus partial

and full disagreements with DCD children suggest that the 3 populations may be controlling and coordinating SB-1 distinctly: for example, when $N = 5$, all populations seem to be recruiting the fourth Gaussian early in the timeline, since all $\tau_{1,4}$ peak at about the same time but $c_{1,4}$ do not (orange selection, Fig. 2.7) so we can hypothesize that (1) *there may be inter-population discrepancies related to spatial configuration SB*-1 *taking place somewhat early in the jump* and that (2) *the more you move to the right on* $c_{1,4}$*, the less mature the jump is, since the sequence of peaks is AD→TD→DCD*. The movement analyst could then manipulate $c_{1,4}$, reconstruct the jumps and inspect the effects near $\tau_{1,4}$.

## 2.5 Conclusions and future work

This paper describes the SB-ST decomposition and how it factors action matrices. Conceptually speaking, *SB-ST can be seen as a synergy model of single postures with time-varying scaling magnitudes*, and it generalizes spatio-temporal profiles proposed to explain locomotory data in motor neuroscience [34] and psychology of vision [20]. Local parametrization of spatio-temporal profiles, although proven critical in the study of actions [34, 41], is not present in human manifold models like GPLVM and GPDM, but it is a feature of SB-ST.

Comparative reconstruction of vertical jumps suggested that: (1) SB-ST can be more adequate than TVMS to factor single-trials, (2) SB-ST can outperform Troje-inspired at the best the trade-off between number of basis functions and $R^2$, (3) it do as well or better than GPLVM with half the representation size. In a second

experiment, we showed that SB-ST can be a good tool to study actions and actors, and results revealed that (1) despite conceptual differences, TVMS, GPLVM and SB-ST all agreed that jumps are mostly loading on a single factor. (2) SB-1 coefficients were consistent among all populations, suggesting jumpers are recruiting the same major synergy regardless of jump maturity (age, presence of disorder) or gender. By inspecting ST-1 statistics, we saw that (3) one of the Gaussians is consistently *coordinated* by all populations to be at the beginning of the trial, but it is *controlled* differently. We note that to discern what exactly these differences mean as well as their significance would require a more thorough analysis and rigorous statistical testing, which surpasses the scope of this discussion (but see Chapter 6).

In ST parameter estimation, we use a family of $N$ Gaussians with fixed standard deviations (stds) to facilitate the comparison among populations, because we could establish correspondences between Gaussians based on corresponding stds (as we did in Sec. 2.4.4 when we fixed the fourth Gaussian and looked at differences in $c_{1,4}$ and $\tau_{1,4}$). Therefore, reconstruction results could improve further if we also optimized for stds; a future development would be to add std optimization, discretize these stds into bins and correspond Gaussians based on the bins. Another interesting future experiment would be to compare the performance of our VARPRO-based ST representation with an ST learned with the dynamic primitives proposed by Ijspeert et al. [43].

# Chapter 3: Alternative representations of action primitives: the traveler-target framework

## 3.1 Introduction

The SB-ST decomposition and related synergy models presented in the previous chapter try to discover primitives from the data with no supervision. Although the methods produce high compression of movement signals, they rely on optimizing criteria that do not express explicit knowledge of actions or actors. This means the assumptions in these models are weakly connected to scientific findings and hypotheses on action primitives, except perhaps for dimensionality reduction (see Chapter 1 for details). For instance, there is no explicit differentiation between body parts and objects; if one wanted to include objected, he or she would have to add extra columns to the action matrix, so objects will be semantically equivalent to any other degree of freedom. To explore a different direction, in the next section, we introduce a top-down system of primitives that incorporates some of the described scientific evidence and tries to design more semantically relevant primitives. As it will be shown, it is founded on *special kinematic events between entities of interest*, that is, body parts and objects. These events reflect relative motions between these

entities, so that primitives will encode (1) target entities, (2) travelers that move with respect to these targets and (3) a description of the motion between the two.

The proposed system is consistent with the evidence surrounding motor synergies: instead of factoring $x$, $y$ and $z$ time series into linear combinations of all body parts (and objects) at once, it will *select a few pairs* of dofs that are meaningful in the context of certain cognitive tasks such as tool use, although linearity in the way dofs are combined is not assumed. The very fact that *it works by pairing* face, hands, objects (that can be later grouped into pairs of sets) rather than representing each of these individually obviously contributes with dimensionality reduction as well. In fact, *the system is based on a grammar* that defines which entities can be travelers and targets, and which cannot, and what types of motion events are considered relevant. This grammar supports the belief of an existing vocabulary in F5/pre-motor cortex of macaques and in the mirror neuron system of humans. Note that working with pairs and their interaction rather than each dof alone is also consistent with the fact that mirror neurons were observed not to fire for mimicry or objects alone, but for both manual and oral grasps. Although the proposed system is discussed in the context of object manipulations and tool use, we believe it can be extended to support full-body actions.

## 3.2 Velocity-based transportation events and the traveler-target framework

In the previous section, we mentioned that the system introduced in this chapter will be "based on special kinematics events", and that these events arise from processing relative velocities between entities. We note, though, that not all relative velocities are of interest; we are particularly interested in the ones that can inform something on the *purpose of the action*, in special, manipulations. Conveniently, purposive manual motions will have a signature, as observed by Marc Jeannerod in 1984, when studying *velocity-based transportation curves* involved in reach-to-grasp velocities [50]. From 7 testing subjects, he saw that the time of peak velocity correlated with the initial distance from the hand to the object, as well as with the amplitude of the movement, rather than its duration. He also reported the onset of the low-velocity phase (a re-acceleration typical when the hand is close to grasp the target) to be highly correlated (in fact, almost equivalent) to the time of the maximum grip aperture.

We have verified qualitatively that the shape of such *transportation curves* can also vary significantly with the level of planning during action execution. For example, in one of our recordings, a test subject pretended to drink from a mug 3 times in a row. Let us consider the relative velocities between left and right hands. In the first trial, he was not immediately sure of whether he should use the left hand as a "helper hand" to carry the mug, and when he decided to do so,

he quickly moved the left hand towards the right hand to catch up with it, and together both hands brought the mug first to the subject's mouth and next back to the table. This "unsure/unplanned" behavior produced an asymmetric profile consistent with Jeannerod's reported curve[1] but much sharper than, for example, the curve resulting from moving the right hand towards the head, that was clearly planned out well before the action took place (compare Fig. 3.1, bottom, #1 and Fig. 3.3, "towards"). In the second trial, he began moving the left hand to help, had a quick moment of doubt and withdrew the hand, causing a weak and jittery velocity pattern (Fig. 3.1, bottom, #2). In the third trial, he brought up the left hand almost as a reflex and pulled it back, resulting in a sharp, low-amplitude velocity profile resembling a short burst (Fig. 3.1, bottom, #3). We would expect that, in a fourth trial, he would not move his left hand at all.

Founded on such observations, one hypothesis is that transportation events could form a visuomotor basis for partially understanding intentional behavior and as a consequence should guide the choice for action primitives. There is plenty of evidence that these relative kinematic cues are being extracted from visual images of moving body parts and objects during action interpretation, as we saw with the proactive attention shifts reported by Flanagan and Johansson [32]. These are signals that somewhat reflect the state-of-mind of the actor (an "honest signal", like termed by Pentland [51]), and the experience of one such curve clearly evokes the notion of a traveler and a target within the visual field of the observer, being therefore essential in describing the action. Next, we describe the traveler-target

[1]But without the low-velocity phase, since there was no grasping.

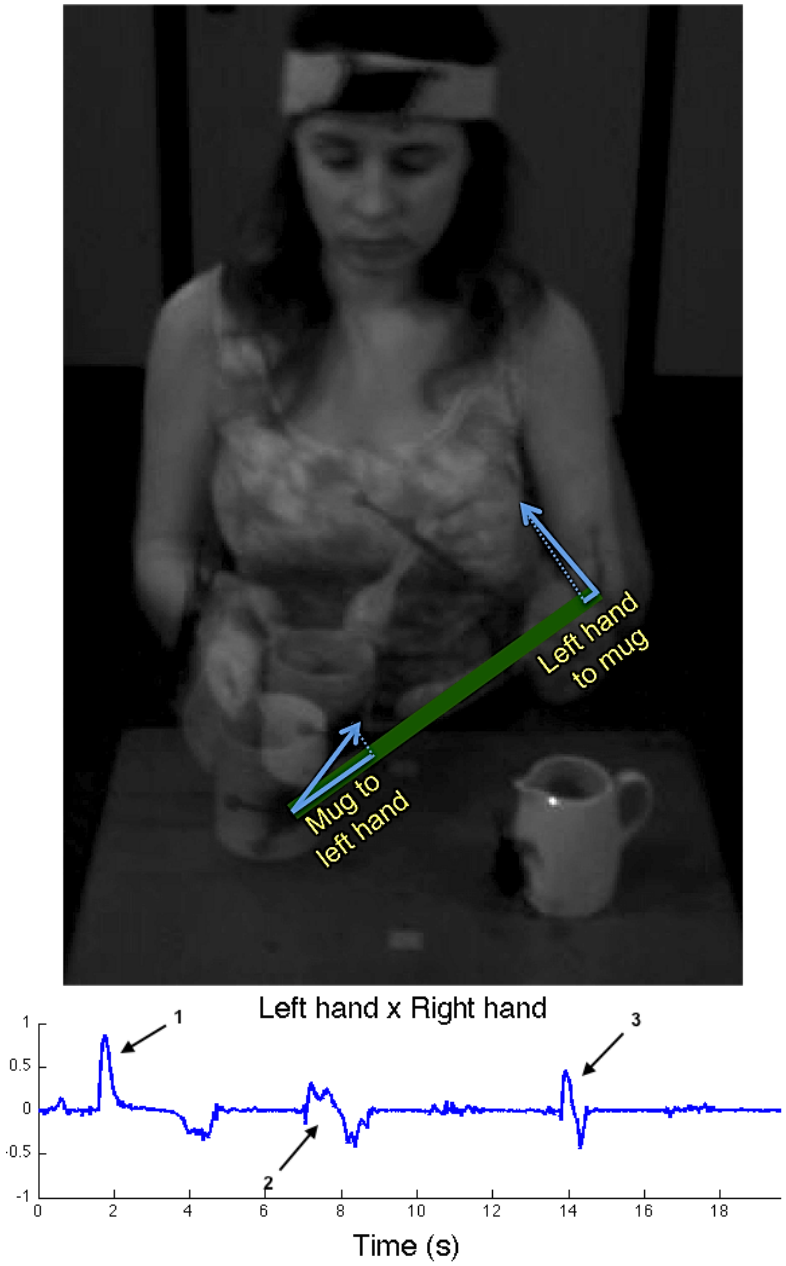



**Figure 3.1:** Top: two overlaid frames. We compute traveler-target velocities by (1) projecting the traveler's displacement vector (blue arrows) from frame $t$ to frame $t$+1 onto the vector that separates the two at $t$ (green line), (2) computing the norm of the displacement and (3) dividing it by the number of seconds $\Delta t$. Note that this quantity is asymmetric: in the given example, the subject brings the mug towards the left hand faster than the other way around, which can be seen by the different sized projections (blue lines on top of green line). Bottom: data from another subject. The left hand moves towards the right three times, but only once with intention to help the other hand (black arrow 1 versus black arrows 2, 3). See text for details.

framework as a computational approach to extract action primitives, and sketch a testbed that could be used to inspect how plausible these primitives are.

### 3.2.1 Extracting action primitives from manipulation data

Here, the raw data will be the trajectories of a number of relevant entities (time series of its 3-D pivot coordinates) as well as a front-to-parallel registered video, such as the actor's head, the hands, tools involved, to-be-manipulated objects and distractors. These data are acquired with an optical motion capture system with a point cloud tracking software and custom-designed trackers. Figure 3.1 shows an example of one frame of our data and illustrates the subject and objects' setup. The resulting time series will be processed off-line, so our algorithms will only have access to the full data of an activity after it is completed. Moreover, we will know which time series correspond to certain body parts and objects, meaning that video, although collected, will only be important in later stages of our work, when more sophisticated vision-based object recognition can be applied to keep track of the identities of entities during the course of an action trial.

To extract primitives, we process the time series of all body parts and objects, find out which targets and travelers are available at each instant and produce a description of what their relative motions look like, according to the following steps.

1. By knowing what entities are objects and what entities are body parts, form sets of candidate travelers and targets by respecting the following rules:

    - Travelers: hands ‖ hands with objects.
    - Targets: hands ‖ head ‖ hands and objects ‖ head and object ‖ just object ‖ a site of interest

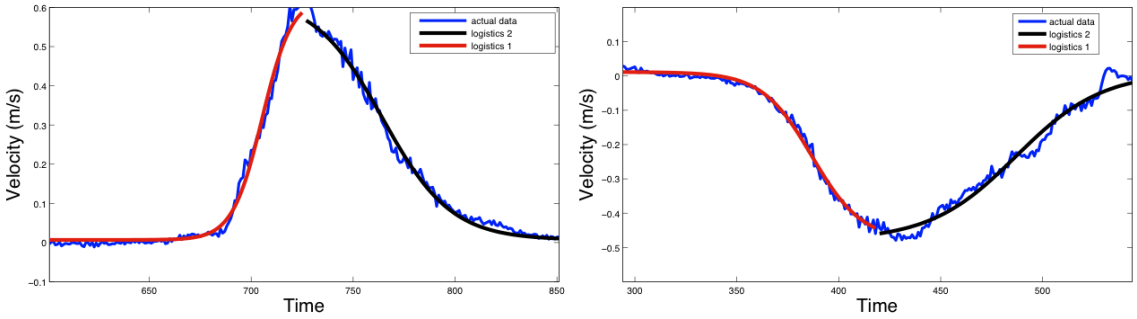



**Figure 3.2:** We fit a composite logistics model to the relative velocity trajectories. On the left, a typical fit for a true transportation curve (*towards*). On the right, the fit for an inverse transportation (*away from*). The red piece corresponds to the first logistics component and explains transportation up to peak velocity. The blue curve explains the piece of the action up to reaching the target, and may include grasping.

2. Compute relative velocity time series between candidate traveler entity $e_1$ and candidate target $e_2$ in the data (like those in Figure 3.1, bottom and Figure 3.3).

3. For each velocity time series, test the hypothesis that $e_2$ is the target of $e_1$ at every instant $t_0$ (i. e. search all transportation events). We do that by first locally fitting a composite logistic model $y = a + (c - a)/(1 + \exp k(t - t_\theta))$ to the data (Figure 3.2) and classifying the optimal parameters as belonging to *a true transportation event* ($e_1$ moves *towards* $e_2$), *an inverse transportation event* ($e_1$ moves *away from* $e_2$, in which case $e_2$ behaves as an anti-target) or *any other type of event*. One logistic models the trajectory anterior to $t_0$, that is, $t < 0$ and $k < 0$ while the second models the posterior section, in that case $t > 0$ and $k > 0$. They are fit independently to allow for large asymmetry before and after $t_0$, and the process produces two sets of optimal parameters. A classifier can then be trained with examples of true transportations and other

motions manually extracted from a few trials. This step turns the velocity profiles into events that flag true and inverse transportations (*towards* and *away from* motions, respectively) or no relevant motion per instant.

4. Apply simple rules (or even classify) combined occurrences *towards* and *away from* into finer motor descriptions, by looking for temporal and spatial regularities, for example:

   - $e_1$ moves *back and forth* w.r.t. $e_2$: when multiple *towards* and *away from* events between $e_1$ and $e_2$ are signaled within a time window.
   - $e_1$ moves *in circles* near $e_2$: when multiple *back and forth* events spanning the same 2-D plane and within a certain time window and range.

   Note that the model leaves room for creating other types of behaviors, and this could go as fine-detail as desired. Computer vision could also be used to supply information about the appearance of hands and objects. Adding the direction of motion should also enrich these descriptions, e. g. $e_1$ moves *back and forth, up-down* near $e_2$. The result of this step is a number of traveler-target pairs per instant of the trial, and the type of motion.

5. For all instants, group travelers that go to the same target (Figure 3.3). A new traveler that resulted from the union of $e_1$ and $e_3$ will then be referred to as $e_1$ *with* $e_3$. This should be analogous to perceptual grouping by common fate.

   As a result of the previous processing steps, and with the addition of the

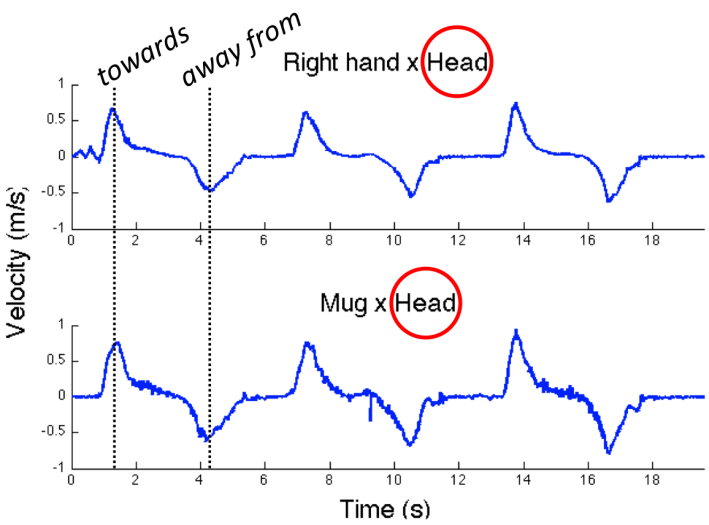



**Figure 3.3:** The right hand (top plot) and the mug (bottom plot) travel in-phase w.r.t. head (see red circles), first towards than away from (dashed lines).This will make right hand and mug to merge into a single traveler (at the marked times) to represent the fact that they have common fate (head).

proper language constituents, every instant of a trial can be expressed as a set of sentences relating travelers and targets, such as:

$(t)$ {*Left hand* moves *towards mug*, *Right hand* moves *away from left hand*}

$(t + \Delta t)$ {*Left hand and mug* move *towards head*}

$(t + \Delta t)$ {*Left hand and mug* move *away from head*, *Right hand* moves *towards pitcher*}

$\vdots$

$(t + k \cdot \Delta t)$ {*Left hand* moves *away from mug*}

From now on, we will talk about primitives in terms of these sentences.

6. Apply two simple rules to reduce the number of sentences:

- The *traveler-unity constraint*: at time $t$, if two entities were once detected as a combined traveler, they cannot be traveler or target of each other. For example, if right hand and mug are a traveler of some target (say, the head) then the sentence *Right hand* moves *towards mug* is forbidden at time $t$, since right hand and mug are grouped, and will then be discarded.

- The *target-unity constraint*: At time $t$, if two travelers move in the same way towards different targets, these targets are grouped into a single target w.r.t. those travelers. The rationale behind this constraint is that, since the transportation curve reflects a position in space, it is likely that travelers going to different targets are in fact going to a single system of targets close together, and we want to reflect that in the representation. For example, sentences: *Left hand* moves *towards right hand* and *Left hand* moves *towards mug* are replaced by *Left hand* moves *towards right hand and mug*.

This concludes the extraction of action primitives according to the traveler-target framework: in short, we produce sets of sentences, each describing a time slice of the action from the viewpoint of two meaningful entities involved and their spatio-temporal relationship. Alternatively, one could see this whole process as processing each time sample of an action data through a grammar like the one in Figure 3.4, where the occurrences of motor-related terminal symbols depend on detecting and processing transportation events, and the object-related ones would rely on visual tracking of entities (body parts and objects). The framework can be extended to

S ⇒ TRAVELER moves IN A CERTAIN WAY TARGET ‖ $\epsilon$

TRAVELER ⇒ ACTING BODY PART ‖ ACTING BODY PART *with* OBJECT

IN A CERTAIN WAY ⇒ DIRECTION *towards* ‖ DIRECTION *away from* ‖ *back and forth*, CYCLIC DIRECTION near ‖ *in circles* near

DIRECTION ⇒ *down to* ‖ *up to* ‖ $\epsilon$

CYCLIC DIRECTION ⇒ *up-down* ‖ *along-across* ‖ *in-out*

ACTING BODY PART ⇒ *left hand* ‖ *right hand*

OBJECT ⇒ *object 1* ‖ $\cdots$ ‖ *object o*

SITE ⇒ *relevant location 1* ‖ $\cdots$ ‖ *relevant location r*

TARGET ⇒ BODY PART ACTED UPON ‖ OBJECT ‖ SITE ‖ BODY PART ACTED UPON *and* OBJECT

BODY PART ACTED UPON ⇒ *left hand* ‖ *right hand* ‖ *head*

**Figure 3.4:** The main production rule S will parse the action data at every instant, and output a sentence describing an action or an empty string. Note that all symbols have clear meanings: nouns refer to objects, body parts or relevant locations. Symbols *towards*, *away from*, *back and forth*, *in circles* together with some direction modifiers express the manner of the motion, and are based on the occurrences of transportation events. The *with* symbol expresses common fate, that is, redundancy in the traveler space. Finally, the *and* symbol reflects redundancies in the target space.

accommodate different entities and type of motions.

## 3.3 Discussion

Previous attempts of designing vocabularies of action primitives concentrated on forming symbols based on absolute kinematics (i. e. different than the traveler-target idea) hand-to-tool events or sub-actions like hand motion to one side or an arm reaching out. These may be useful to ad-hoc tasks but will lack generality, simply because they do not absorb enough semantics. See, for instance, the repre-

sentative work of Inamura et al. [52]: their *proto-symbols* are instances of continuous Hidden Markov Models that produce sequences of low-level based motion elements based on joint-angles (although their model would accommodate other physical descriptions). These proto-symbols are very sophisticated and are able to learn and generate motion patterns, but lack clear semantic meaning: they cannot convey intentions or describe an action just like the sentences that we described in the previous section. This ability to express the action on linguistic grounds is crucial because language is exactly what links the motor and visual facets of the action.

We are rather looking for a system of primitives that could convey purposiveness or intentionality, along the lines discussed by Justine Cassel [53], but in the realm of day-to-day activities rather than gesture understanding. As she says, "in order for (...) gestures to be accounted for in a theory of lexical choice, the semantics must be of a form that allows knowledge of the world", that is, in the context of manipulations, the code has to reflect not just the kinematics of effectors, but to transform these kinematics into something that helps expressing the actor's intentions while carrying out the action. In a later effort, Cassel et al. [54] proposed an encoding scheme that made use of hand shape, orientation and location within pre-established zones in the actor's workspace (see the Appendix of McNeill et al. [55], pp. 378 for more details). Their goal was to assess the use of gestures when communicating directions, and they did merge hand kinematics with semantics (e.g. hand pointing to building) at utterance generation level, but in the end, the code itself was only based on absolute hand features. The semantic event chain (SEC) proposed by Aksoy et al. [56] shares some ideas with the system of primitives

proposed here, because it encodes the relationship between parts, objects and even object states (liquid moved from one container to another) but without expressing these primitives as language constructs.

# Chapter 4: Actor-aware measurement of movement: the case of infant motion capture and Autism Spectrum Disorder

## 4.1 Introduction

The previous two chapters assumed movement data to be readily available for the discovery and analysis of action primitives. Although this is true when recording from various animals and most adult humans, it is not always the case. In this chapter we will then switch to a discussion on the importance of making these measurement systems aware of the nature of the subject being measured. Our focus will be typically and atypically developing human infants, for reasons that will soon become clear.

From preventive healthcare to developmental robotics, many are the disciplines that can profit from automatic means of acquiring infant movement data. Here we are particularly interested in aspects of Autism Spectrum Disorder (ASD), a neurodevelopmental disorder whose most characterizing traits have been shown to be tightly connected to many of the aspects that make humans different than any other species, such as creativity, language, social engagement and even thinking [57, 58].

As Peter Hobson puts it in *The cradle of thought* [59], p.183:

> "If we are interested in uncovering the foundations of interpersonal relations and creative, flexible symbolic thinking, autism is a good place to start – precisely because it is in autism that we find a unique combination of abnormalities in these two domains of mental functioning. Autism promises to disclose the conditions that make symbolic thinking possible for those of us who are not autistic".

In the context of actions, the recently discovered link between certain early-age movements and later development of typical ASD traits makes it reasonable to consider the use of behavioral assessment tasks empowered by pattern recognition tools as adequate means to the pursuit of the relationship between complex later-in-life obsessive traits, or even social impairments and motor abnormalities that are commonly displayed by infants at high-risk for ASD [37]. An objective, quantitative answer to one such question could perhaps open way for science to trace back to the neural processes involved in these kinds of abnormal motor developments, or even the underpinning genetics. Naturally, these studies will thus depend on the availability of movement data. Although the natural choice of marker-based motion capture technology will appear adequate at first glance, given it has been successfully used to collect data from older children and adults ([60–62] and see Section 2.4 in Chapter 2) it should be expected to fail on infant subjects. For example, in Sec. 4.2 we go over a preliminary experiment we co-mentored in partnership with the Center for Autism and Related Disorders at the Kennedy-Krieger Institute, which

gave us a practical notion of the difficulties and consequences of subjecting infants to standard marker-based motion capture [63]: markers are usually relatively big, bulky and distracting, which can make the capture sessions very uncomfortable for the infants or even contaminate the data and mislead interpretations. Henceforth, the markerless paradigm is not just desirable, but the right way to go about it.

More precisely, by markerless motion capture of infants or *markerless infant mocap*, we refer to the problem of obtaining full-body movement data from children within the age of 0 to 12 months with the use of movement sensors that track the child within a pre-defined volume without depending on any physical markers (or trackers) or wires or special suits to be placed on the child's body. A careful literature review reveals that approaches to the problem split into two main paradigms, depending on if it is *pressure-based* or *optical-based*. Digital pressure sensors were first reported as being part of a markerless infant mocap system in the late 1990's, but are still in use [64, 65]. Typical setups will include one or more lattices of multiple sensors placed under the cushion of a crib station, and these sensors produce maps of simultaneous activities of body parts from time to time. Meanwhile, optical-based setups comprise standard, single-view video acquisition.

Regarding applications, markerless infant mocap architectures have been designed to support infant psychology research and to handle tasks like baby posture and activity recognition, biometrics, general child monitoring, Sudden Infant Death Syndrome (SIDS) prevention, seizure diagnosis and automatic computation of behavioral markers for the early study of neurodevelopmental disorders such as Cerebral Palsy and Autism Spectrum Disorder (ASD) itself. These studies are further

summarized in Sec. 4.3.

In terms of previous achievements and state-of-the art, pressure-based solutions rely on tracking blocks in images that, under very strong assumptions and a lot of luck, will match body parts [1]. In all cases infants will remain in the crib, which limits the number of behaviors that could be recorded and makes it too constraining. However, a major conceptual outcome of studies based on pressure-based architectures was that holistic representations of infant motions may be sufficient to allow for the inference of certain (canonical) postures that are often seen in infant behavior [1]; this is an important result that could even be explored by future computer vision systems, for example through the use of the increasingly popular (and affordable) depth sensors [66]. Speaking of which, current computer vision technology has gone as far as being able to detect epileptic seizures through the tracking of motion blocks [67–73], and when more sophisticated skeletal models were attempted, vision systems were able to achieve good discrimination between normal and abnormal head lags and arm asymmetries [2, 74–76]. Still, the number of published studies to date is still surprisingly low, plus, except for a couple of cases, most of the presented results are either of qualitative nature or reported solely on the basis of the driving application, making it really hard to judge the accuracy of the obtained data, or equivalently, how well their markerless infant mocap solutions are performing: take for example Hashemi and colleagues [75, 76], or Spina and colleagues [2], who presented evidence of good agreement between their system's ability to point out arm asymmetries and the inputs of an expert physician, given the movements of a small population of infants at high-risk for ASD, but did not

show any comparison between their generated motion capture signals and *available* ground-truth. Besides, no learning of thresholds and model parameters or cross-validation of results were reported in their work, so it is also very hard to judge how well the predictions (scores, diagnoses) would generalize unseen data; it could be the case that parameters are just overfitting the clinician assessments.

We believe the main reason why markerless infant mocap has been overlooked by computer vision and artificial intelligence communities would result of a first impression that the problem would be a mere downscaled version of the general markerless human motion capture. On the contrary, it is a very special case and should be treated as such, mostly because infants are pre-language human subjects that have unique physique, a peculiar postural repertoire and fast-evolving physique and cognition. Despite the existing evidence on the importance of modeling these aspects [1], none of the vision systems have chosen to do so, which may result in restrictive performance and/or applicability upper boundaries, and consequent lack of generality. In Sec. 4.4 we reflect on the achievements and setbacks of markerless infant mocap research, and consider ways of exploring the infant features above listed to advance the state-of-the art.

## 4.2 Motion capture and the early assessment of Autism Spectrum Disorder

Autism Spectrum Disorder (ASD) is presently understood as a neurodevelopmental disorder that alters how a person senses and acts towards other people, objects

**(a)** **(b)**

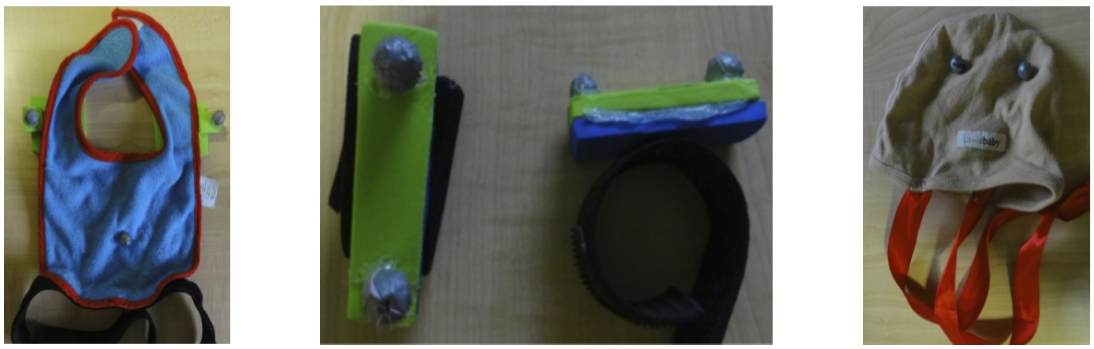
**Figure 4.1:** (a) U. of Miami-UCSD motion capture suit measures interactions between the baby and the caregiver. Reproduced from [77]. (b) U. of Maryland AMIRA team's custom-designed infant mocap suit on a baby dummy: it consists of a bib tracker, two wrist straps and a hat. Reproduced from [63].

or even themselves. It is considered a *spectrum disorder* because it encompasses a variety of symptoms, and these symptoms vary from individual to individual. The American Psychiatric Association's Diagnostic and Statistical Manual (APA manual) has a number of criteria for the diagnosis of ASD, and the manual is revised from time to time. For example, according to the $4^{th}$ edition [57] the individual used to be diagnosed as having Autism Disorder (AD), Asperger's Syndrome (AS), or the catchall Pervasive Developmental Disorder Not Otherwise Specified (PDD-NOS) which included subgroups Rett Syndrome (RS) and Childhood Disintegrative Disorder (CDD). Group and subgroup selection depended on the symptoms, their severity and when in the developmental process they were observed. More recently, the $5^{th}$ edition of the APA manual [58] ended the formal diagnosis of AD, AS and PDD-NOS, and placed them under the single umbrella of *Autism Spectrum Disorder*. Individuals are now supposed to be diagnosed as pertaining to a certain *level* of the ASD continuum rather than to a group or subgroup. Both versions of the manual elaborate on the *action aspect of the disorder*, and establish that ASD individuals

are expected to display a subset of the following: (1) impaired use of non-verbal behaviors that regulate social interactions like eye contact, body postures and gestures or (2) stereotyped manual and full-body motor mannerisms, (3) eventual loss of purposeful manual skills or even (4) problems coordinating trunk and gait.

While scientists are still in the process of figuring out the nature of the disorder, the latest data from Centers for Disease Control (CDC) [78] indicate an increase of ASD incidence: 1 in 54 boys and 1 in 252 girls were identified as having ASD, a growth of 23% compared to the (last recorded) 2006 prevalence ratios. On the positive side, recent results are pointing to a possible early diagnosis of ASD. Bhatt and colleagues [37] compiled a variety of studies and proposed that *the observation of certain sensorimotor abnormalities still in infancy can predict both a future diagnosis of ASD and the development of the typical social traits ASD is mostly known for*. They also reported findings where babies who have siblings with positive diagnosis for the condition are 20% more likely to display certain motor (gross and fine), postural and perceptual delays, among which trouble holding the head or rolling, or to reach for an object, preference for prone playing rather to sitting, and lack of attention in visual tasks. These results are extremely important from a prophylactic viewpoint: infants can be run through behavioral tests that verify the presence of such delays, and depending on the severity of what is observed, these children can undergo preventive therapy much before the 5 years deadline [79]. Although this will certainly not cut down on the CDC incidence numbers, it will definitely ease future social inclusion of ASD individuals and shorten the need for medical treatment, hence minimizing associated costs [80]. From a methodological viewpoint, it

becomes clear that the study of ASD will require the acquisition of movement data from infants while engaged in certain tasks, so that the links proposed in [37] can be better understood.



**(a)** **(b)** **(c)**

**Figure 4.2:** Closer view of the AMIRA suit trackers. (a) The bib with 2 shoulder markers and a chest marker. (b) Velcro wrist straps with two markers on a foam base each. (c) The hat and the 2 frontal markers (the third is not visible). The point-cloud tracking software models the wrist markers in (b) as lines in space, and the ones in (a, c) as planes. Figures reproduced from [63].

### 4.2.1 Marker-based motion capture in ASD studies and related

There are a number of studies based on movement data captured from babies; in three cases we came across, movement data were collected by manually marking trajectories over video frames [81–83], but most often marker-based motion capture was used. In the context of marker-based studies of ASD, Mari et al. [60], focused on reach-to-grasp patterns, Chester and Calhoun [62] observed gait symmetry disparities from full-body motion capture, and Shic and et al. [61] found gross and fine attention differences leading to the appearance of particular developmental trajectories, out of data from six subjects exposed to face stimuli. We note that only a small fraction of works using baby motion capture did it predominantly on infant subjects [84, 85]. Partner groups from the University of Miami and the Univer-

sity of California at San Diego have recently built a custom motion capture suit to acquire data from infants and caregivers (Fig. 4.1a) while engaged in behavioral tasks [77, 86]. Their goal is to both study ASD and replicate baby behaviors in robots. At the same time, project Early Autism Sweden (EASE) is a collective effort between the Karolinksa Institutet and the Uppsala University that is looking into applying eye-tracking and body motion capture to study first-year development of ASD [87].

The thesis project of team AMIRA (Analyzing Movement of Infants at the Risk of ASD), a group of undergraduate students working under our guidance and the support of Dr. Rebecca Landa, founder and director of the Center for Autism and Related Disorders at the Kennedy-Krieger Institute, was an effort to take the considerations of Dr. Landa and her colleagues [37] to the experimental level [63]. The idea was to try marker-based motion capture to measure movements of infants at high-risk and control subjects engaged in behaviors that should give rise to the delays more likely to be displayed by the former group. These behaviors were: pulled-to-sit, postural control/imitation, reach-to-grasp and visual-tracking of an object[1]. Data sessions included the baby, the caregiver (mother) and two testers, all working on a mat, plus one student that operated the computer and another that video-recorded the trials (Figs. 4.3a–4.3d). Because infants are smaller and have different body proportions than older children and adults, markers were often

[1]This study was conducted according to the principles expressed in the Declaration of Helsinki. It was reviewed and approved by the Institutional Research Board of the University of Maryland at College Park (IRB Protocol: 10-0445 – Analyzing the movement of infants at risk for autism spectrum disorders). Written informed consent was obtained from parents after a careful explanation of the testing procedures.

too close to each other, and the available marker-based motion capture software quickly fell apart. Students had to come up with a custom-made baby mocap suit (Fig. 4.1b) and their own tracker setup (Fig. 4.2), and resorted to a point cloud tracking software to read in movement data, that is, a system that only tracks the position of markers without fitting a physical model of the human body to it.

Team AMIRA's study was able to conclude that high-risk participants were significantly slower to grasp than control counterparts, out of 9 samples of high risk grasps (2 participants) and 16 samples of control grasps (4 participants), but still, a lot of data was not useful or lost, most likely because of the system's sensitivity to the people in the volume and surroundings, resulting in occlusions and camera interferences. As a consequence, students had to manually label and/or post-process tracked markers as an attempt to rescue data that got corrupted due to tracking errors. In a few situations, the babies did not seem to feel comfortable wearing the suit or simply became curious and tried to remove the markers, which caused even more problems to the motion capture system and delayed capture sessions (Figs. 4.3e–4.3h). Nonetheless, we believe that the most important take-home lessons of the AMIRA project are that (1) a potential tool for ASD diagnosis would centrally depend on systems capable of obtaining movement data from infants in a minimally invasive fashion and (2) whoever designs such systems has to keep in mind that the target subjects will be pre-language humans with unique biometrics, and these individuals are going to be acting around a number of other people with or without assistance, and will potentially interact with objects. Marker-based motion capture does not appear to be a viable solution: even sophisticated, custom-designed suits

like the one in Fig. 4.1a will still present serious drawbacks: wires may affect the baby's motion, and leds can be distractive. Even more importantly: *these issues could even compromise the reliability of the output movement data in a very subtle and dangerous fashion*: for example, bare distraction caused by flashy lights and hanging wires could be mistaken for abnormal eye-contact and as a consequence prescribe a wrong diagnosis of ASD or another incorrect conclusion.

### 4.2.2 Markerless motion capture of infants in ASD studies and related

The bulk of research we reviewed together with our own practical experience in capturing movement data from infants made us advocates of the markerless approach. In the next section, we will scrutinize different methodologies and problems where markerless infant mocap was attempted along with state-of-the-art achievements and setbacks, including the very recently developed systems of Dogra et al. [74] and Hashemi/Spina and colleagues [2, 75, 76], both pioneers in the use of the markerless paradigm in the computation of behavioral markers towards the early assessment of neurodevelopmental disorders. The first team resulted of a partnership between the Indian Institute of Technology at Kharagpur and the Institute of Post-Graduate Medical Education and Research/Seth Sukhlal Karnani Memorial Hospital at West Bengal, both in India; they were able to reasonably predict pulled-to-sit scores for 43 infants from their video recordings, based on feedback provided by collaborating physicians.

Of major importance, the second group, with researchers from Universidade Estadual de Campinas in Brazil, Duke University and University of Minnesota, have developed computer vision tools to help in the studies of early-age ASD to assess performances on both visual attention and motor tasks. The first tool aims at investigating visual attention patterns, and consists of a tracking software that measures left-right and up-down head motions based on the detection of eyes, nose and ears. Whenever tasks involved objects, their positions on the video had to be manually marked. Left-right measurements were used to approximate the children's performances on *visual tracking* (following an object from one side to another, like in Fig. 4.3c) and *disengagement of attention* tasks (shift attention to a second competing conspicuous stimulus presented along the left-right axis, while attending to another stimulus). Delayed, discontinuous or non-smooth tracking and/or delayed disengagement are regarded as abnormal and point to ASD. Meanwhile, up-down motions were used to approximate performances on *shared interest* tasks, that is, a complex test that verifies whether the child perceives a third-party involved in a task, and seeks to engage with that party. In their version of the task, the experimenter rolled a ball on the table towards the child, and if the child sought eye contact with the experimenter or the caregiver, the behavior would have been considered normal. Infrequent or limiting face seeking would indicate ASD. These test behaviors are a subset of the standard AOSI (Autism Observation Scale for Infants) battery of behavioral assessment [88]. Their tools were experimented on movements of 15 children, 9 of which were infants, recruited for (1) being premature, (2) having an ASD sibling, (3) showing delays or (4) being diagnosed with ASD (1 subject).

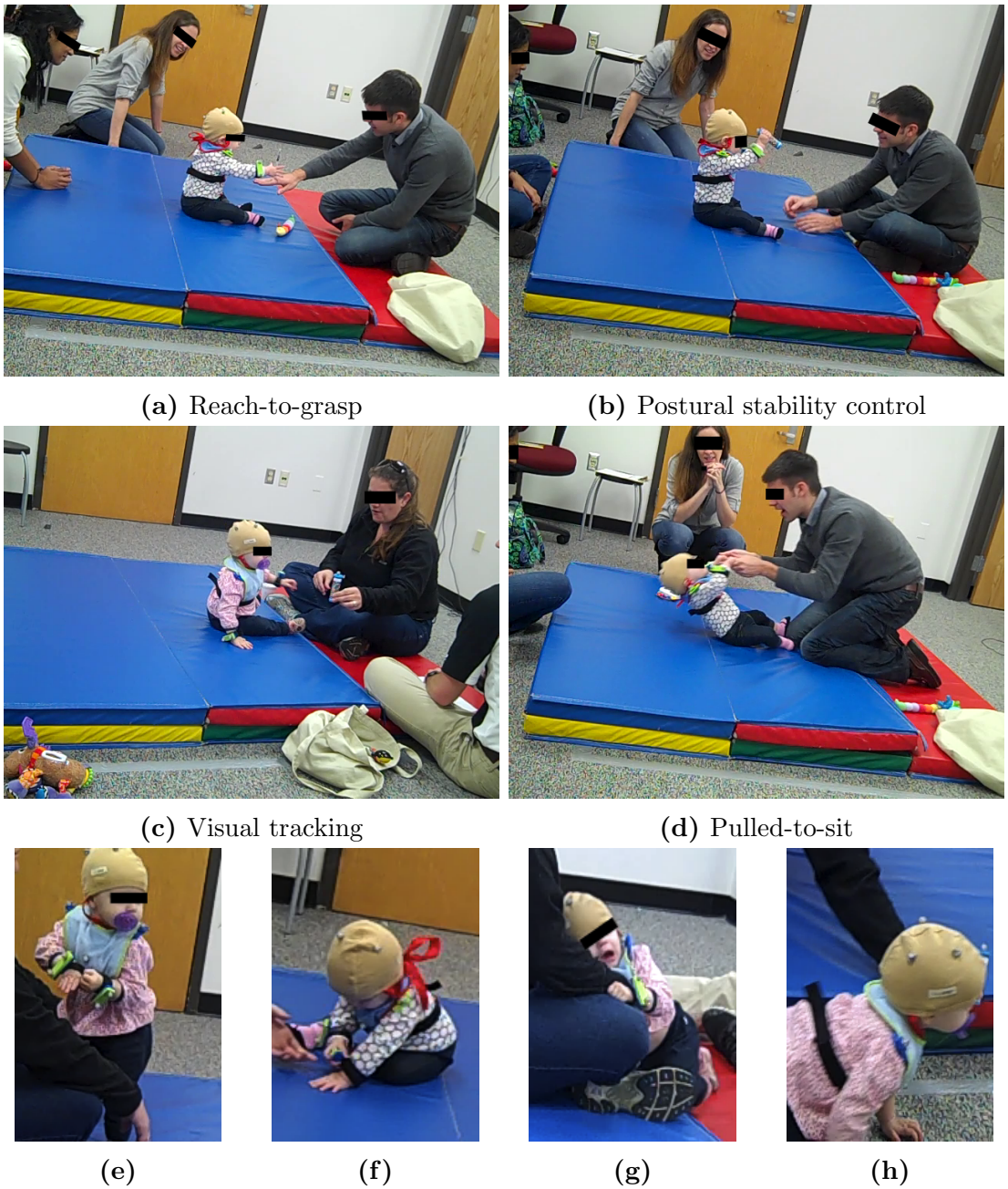
Scores for each task were produced and compared with the assessment of one or more experts, and general agreement was observed. Although this group's vision-based, markerless head tracking system is arguably a form of markerless motion capture, the focus of this account lies on their second vision tool, which tracks body poses through frames and evaluates arm asymmetries. Both this tool, and the one of Dogra et al. [74] will be reviewed in the upcoming section.

## 4.3 Markerless motion capture of infants

Markerless motion capture of infants (or markerless infant mocap) remains a vastly unexplored terrain, despite the very interesting potential applications, as will be summarized next. Previous work can be roughly split into two streams: pressure sensor-based and optical-based, depending on which devices are utilized to read in infants' movement data. Efforts resulted from individual and collaborative work, and spanned a variety of backgrounds, among which engineering, robotics, computer science, psychology and medicine.

### 4.3.1 Pressure sensor images

Perhaps the first approach to infant mocap was to build pads or mats with pressure-driven sensors and to place them on special cribs where babies would be laid on. From the digitized 2-D projection of the baby's pressure against the sensor surface, *pressure sensor images* are produced. An array of such images forms a *pressure video*, which records changes in pressure that almost invariably result from the



**(a)** Reach-to-grasp **(b)** Postural stability control

**(c)** Visual tracking **(d)** Pulled-to-sit

**(e)** **(f)** **(g)** **(h)**

**Figure 4.3:** Selected moments of AMIRA test sessions. (a–d) The behaviors tested. (e–f) The subject removes one of the hand trackers during the test session. (g) The subject gets scared by one of the testers and hides on her mother's lap. (h) Subject gets distracted and crawls out of the capture volume.

baby's movements within the crib.

Along this line, we begin by describing the work of Weinberg et al. [64] which introduced the *BabySense* system and was aimed at children from 0 to 12 months. It used a one of such custom-designed sensor pads to detect a short number of limb motions and behaviors, such as sitting and standing up or playing with toys, based on changes in the capacitance of the built-in fabric electrodes. More than just a tool for psychologists and parents, the creators of *BabySense* wanted the system to help babies develop their sensorimotor capabilities, by allowing them to interact locally and remotely with objects and other humans, including peer babies. For example, the system would react to a particular baby's behavior locally, by showing her lights and making sounds, or remotely, so that when a peer baby played with a toy, the same toy in her crib would wiggle.

Harada and colleagues [1] proposed a similar setup to measure behaviors of six-month olds but, as opposed to *BabySense*, they have provided technical details on sensors and algorithms. The way they turn pressure videos into predictions of baby postures, behaviors and body parts is summarized by the dependency acyclic graph (DAG) of Fig. 4.4. They begin by computing an overall movement measured referred to as *activity score* (AS): a time-series where each data point integrates the intensity of body movement over the measurements of all 384 pressure sensors at once. There were three behaviors (or states) of concern: quiet, crying and what we here call *otherwise*. Each of these corresponded to 3 intervals of the AS scale: the quiet state was marked by light movement of chest and abdomen while breathing, and was fired anytime an 1-minute average AS $< \theta_{AS}^{quiet}$. Crying often resulted

from hard breathing, and was characterized by hard movements of the head, chest, abdomen, plus some arm and leg motion, and was triggered when a 1-minute average of AS $> \theta_{AS}^{crying}$. Lastly, the otherwise state corresponded to average AS values within the interval between the first two or $\theta_{AS}^{quiet} <$ 1-minute average $AS < \theta_{AS}^{crying}$. These range thresholds were determined by observation.

One of the requirements of their system was to accommodate infant growth, so one of the modules was responsible for estimating the physique of the child. Weight and height are computed first, as soon as the baby is observed in the quiet state. Weight estimation is based on a regressed curve that related the digital pressure output of a sensor and its corresponding load in grams. Height, on the other hand, is further estimated as a quadratic function of weight. Finally, the lengths of body segments head-chest, head-abdomen and head-hip are estimated from the computed height, in accordance to a model for 6-month old babies or younger. From the computed physique, the system then attempted to obtain posture information by checking the pressure image for the number of contact areas on the pad. Babies were expected to be on supine or prone position, that is, lying on the back or stomach, respectively. For the supine position, head, back and hip contacts were expected to produce 3 areas of significant pressure, while for prone position, head, abdomen and both legs should give rise to 4 pressure areas. So, the posture is chosen by thresholding the pressure image and counting the number of connected components through all instants. If the 1-minute average number of areas $\leq \theta_{pressure}^{supine}$ the system assumes the baby to be in supine position, otherwise it assumes prone position. The last task is to estimate the positions of head, trunk,

chest and abdomen and respective motions by looking back at the pressure image and using the body segments calculated moments earlier. To find the head and the trunk, the system continuously thresholds the pressure image with decreasing cutoffs, until the number of connected components $< 2$. It then sets imaginary circles of radius equal to the length of head-chest segment, centered around the two remaining connected components. A new thresholding is done to the original pressure map, now at a much lower cutoff (more permissive), and the number of binary elements (areas) of each circle are counted; the one with the largest area is recognized as the trunk, and its center is considered the trunk position. The head label is assigned to the other area, and its center is calculated the same way. Chest, abdomen and hip are localized by imaginary circles around the head center, with radii set to the lengths of head-chest, head-abdomen and head-hip, respectively. The element with highest pressure value at a head-chest distance away from the head is consider to be the chest. Similarly, the abdomen and hip positions will be arise from the elements with highest pressure at radius head-hip and head-abdomen, in turn. The intensity of movement at each location is approximated by its pressure values.

As per results, AS *versus* time plots were used to prove that proposed cutoffs $\theta_{AS}^{quiet} and\ \theta_{AS}^{crying}$ would work to recognize the behaviors quiet, crying and otherwise for a pair of infants 2 and 5 months old. For another pair of babies 1 and 4 months old, graphs with number of contact areas *versus* time were used to show that prone and supine postures were properly determined by thresholding the number of contact areas at the proposed $\theta_{pressure}^{supine}$ cutoff. For the same pair of subjects, exam-

ples of where the body parts were properly located were also provided. Moreover, plots of another infant's chest and abdomen pressure intensities *versus* time served as a final evidence that movement signals can be indeed obtained by the proposed method.

Apparently, pressure pads are still being researched. In 2010, Boughorbel et al. [65] were able to recognize a set of behaviors (breathing, sitting, standing, lying on the back, crawling and lying on the side) similar to [1], using a set of four pressure mats. Features were extracted by placing imaginary dartboard-like polar grids onto the center of gravity of pressure images, accumulating per grid cell pressures, and forming a rotation-invariant feature vector consisting of the mean, standard deviation, kurtosis and skewness across cells. The best recognition results arose from combining a single-frame classifiers with and vote-based classifiers, as follows: for a giving frame, both classifiers are run. If the top-2 voted classes had almost the same number of votes, the system would then pick the classification result of the single-frame classifier, otherwise it would just choose the top voted class. The system was tested on 3 sequences of the same 1-year old child collected on 3 different testing dates.

### 4.3.2 Optical images

Since the mid-2000's, digital video cameras, storage and communication technologies have experienced a dramatic cost drop, and as a result a variety of camera-based surveillance products became accessible to the average consumer. One of the ideas

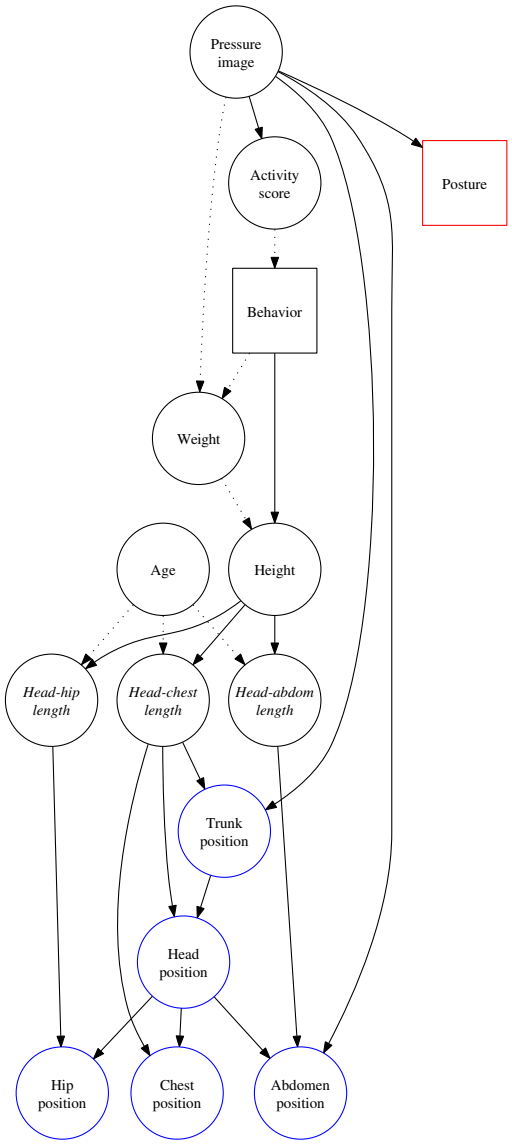



**Figure 4.4:** Dependency DAG derived by interpreting the work of Harada et al. [1]. Discrete and continuous entities appear as squares and circles, respectively. The position of body parts (blue circles) are the target variables. An solid arrow from one variable to the second means that the second depends on the first to be determined, while a dashed arrow indicates the second is a function of the first, and the function is known prior or computed with regressions. Variables that are determined based on estimated height and age are displayed with italic captions (a model for babies of age $\leq$ 6-months was used): segments from head to chest, from head to abdomen and from head to hip. Height itself depends on estimated weight and whether the baby's behavior is recognized to be *quiet*. Note that posture type (red box) is recognized and accessible, but is not being used to help solving other tasks.

back then was to develop video-based system to help looking after unattended babies 6 to 12 months old, with the job of issuing alarms when (1) *the baby's hand moved progressively near the mouth* and when *the hands occluded the mouth alone* (2) or *with an object* (3) [89]. The first situation is detected from tracking the hand to head distances over time, while the second come from measuring the increase or decrease of skin color pixels within the head region. Detection and tracking of body parts is done in a very rudimentary fashion, by relying on simple heuristics and strong posture constraints. For example, the baby is assumed to be facing the camera while sitting upright against a dark, non-skin color background, and is expected to be wearing short-sleeves and short pants so hands and legs are visible to the camera. To detect body parts and track the baby's movement, the system first looks for skin patches by thresholding the image and pulling out regions from connected components. Regions are labeled based on the assumed position of the baby: the topmost detected region is the head, second and third topmost regions are the hands, and the lowest regions are the feet. Eyes are found out of the darkest two points in the head region, and the mouth is estimated from the inter-eyes position. Occlusions with head and legs are computed with templates. The system was tried on 10 sequences, apparently with the same baby, from which a few pictures with the tracked body parts were presented in the paper, both when the system worked and when it failed. Problems with 8 of the sequences were attributed to the system's poor ability to deal with head rotations and fast motions of hands.

#### 4.3.2.1 Diagnosis of epileptic seizures

Members of the health care community have also foreseen digital image and video processing as a useful means to assist providers with more reliable diagnoses. A project that stood out was the computer-based recognition of certain types of seizures undergone by babies. These seizures are known by experts to be characterized by patterns of arms and/or leg motions, so the tracking of body parts becomes a natural first step towards a final system that can discriminate among seizures and rule out irrelevant behavior. Karayiannis and colleagues adopted the block matching paradigm, according to which image regions (blocks) corresponding to body parts (or *anatomical sites* of interest) are tracked by assuming that they preserve appearance throughout the recordings. In particular, the use of *robust motion tracking* framework was proposed, that is, a modular solver for tracking the motion of image blocks that is specified by a *transform model* and a *tracking error function* [69]. The former controls geometry and holds the to-be-optimized parameter set, while the second defines the search space for the optimal parameters of the first, and also controls how outliers are handled during optimization.

In short, tracking of each block between two frames $I^t$ and $I^{t+\tau}$ is done by finding the optimal parameter vector $\mathbf{z}$ that minimizes an error measure $\epsilon(\cdot)$ that depends on the tracking error function $\phi(\cdot)$ of inter-frame appearance differences $\Delta I = I^{t+\tau} - I^t$ within a $W$-pixel neighborhood around the location of the body part block being tracked. By approximating the error measurement by a first-order Taylor expansion, it can be shown that the solution amounts to finding an optimal

step vector $\delta\mathbf{z}$ that minimizes the error function of Eq. 4.1 and make $\mathbf{z}^{t+\tau} = \mathbf{z}^t + \delta\mathbf{z}$:

$$\epsilon = \sum_W \phi(\Delta I + \nabla_{\mathbf{z}}(I^{t+\tau})^\top \delta\mathbf{z}). \tag{4.1}$$

$\nabla_{\mathbf{z}}(I^{t+\tau})$ is the gradient of the error measure $\epsilon$ w.r.t. $\mathbf{z}$. If gradient descent is used, $\delta\mathbf{z}$ will arise from:

$$\delta\mathbf{z}^{i+1} = \delta\mathbf{z}^i - \alpha\nabla_{\delta_{\mathbf{z}}}(\epsilon)\epsilon,$$

where $i$ is the index of the current iteration and $\alpha$ is a usually small scaling constant. The gradient of the error measure $\epsilon$ w.r.t. the step can be shown to be:

$$\nabla_{\delta_{\mathbf{z}}}(\epsilon) = \sum_W \nabla_{\mathbf{z}}(I^{t+\tau}) \underbrace{\phi'(\Delta I + \nabla_{\mathbf{z}}(I^{t+\tau})^\top \delta\mathbf{z})}_{\substack{\text{Contribution} \\ \text{of the tracking} \\ \text{error function}}}, \tag{4.2}$$

and $\nabla_{\mathbf{z}}(I^{t+\tau})$ can be factored as:

$$\nabla_{\mathbf{z}}(I^{t+\tau}) = \underbrace{\nabla_{\mathbf{u}}(\mathbf{v})^{-1}}_{\substack{\text{Contribution} \\ \text{of the transform} \\ \text{model}}} \nabla_{\mathbf{z}}(\mathbf{v})\nabla_{\mathbf{u}}(I^t). \tag{4.3}$$

Vectors $\mathbf{u} = [x^t, y^t]$ and $\mathbf{v} = [x^{t+\tau}, y^{t+\tau}]$ are the coordinates of the block at $I^t$ and $I^{t+\tau}$, respectively. Note that the transform model is plugged into the tracker through the first two factors on the right side of Eq. 4.3, since gradients $\nabla_{\mathbf{u}}(\mathbf{v})$ relates the coordinates of the block before and after the transform, whereas $\nabla_{\mathbf{z}}(\mathbf{v})$ relates the transformed coordinates with the parameters of the model. The (derivative of the) tracking error function affects the gradient descent step of Eq. 4.2. The rest

remains the same for all tracking models, and will depend on the computation of image differences or spatial derivatives. Also, tracking error functions must satisfy the *admissible function criteria*, which are: to be positive everywhere, monotonically increasing and decreasing when $x > 0$ and $x < 0$, respectively and to be piecewise differentiable. To satisfy the *robustness criterion*, it has to increase slower than $\phi(x) = \frac{x^2}{2}$ as $x$ moves away from the $x = 0$ in either direction[2].

Still in [69], different robust motion trackers had their performance tested, by varying both transform models and error functions. Two experiments were carried out, each of which on two distinct sets of 18 sequences containing myoclonic and focal seizures plus random movements, six sequences each. These sequences are part of the CRCNS (Clinical Research Centers for NeoNatal Seizures) database with hundreds of both EEG signals and video recordings of 46 individuals. The type of seizure assigned to each sequence was collectively decided by a team of clinical neurophysiologists and neonatal electroencefalographers who carefully analyzed the data during face-to-face group reviews. The results were reported in terms of how close the motion activity signals produced by the tracker models matched manually annotated counterparts.

In the first experiment, the goal was to find out which transform model would perform best in the first 18 selected sequences, so the tracking error function was fixed to be the baseline function $\phi(x) = \frac{x^2}{2}$ and the following transform models were tested: simple translation, affine, fractional and generalized fractional. The last two models were found out to be the most successful ones, with the generalized fractional

[2]The use of $\phi(x) = \frac{x^2}{2}$ makes $\epsilon(\cdot)$ into a sum-of-squared error criterion.

model being the best overall. In the second experiment, different tracking error functions were tried and the transform model was set to be the generalized fractional model. It came out that the two error functions proposed, $\phi(x) = \frac{\ln(\cosh \beta x)}{\beta^2}$ and $\phi(x) = \frac{\tanh^2 \beta x}{2\beta^2}$ were the best performing ones. With pictures showing the manually labeled motion signals and the system's estimations, the study has presented evidence that these functions were indeed able to handle certain jerky motions typical of seizures, in comparison to the baseline function and a selected competitor, $\phi(x) = \frac{x^2}{x^2+\sigma^2}$.

Prior to developing these trackers, members from the same research team had tried/proposed a number of techniques to acquire motion signals for the same CRCNS dataset, among which optical flow techniques [71], the Kanade-Lucas-Tomasi feature tracker [67], plus adaptive and predictive block matching [68, 72]. They have also tried to estimate image block motions by minimizing a second order Taylor expansion (rather than the first order approximation of Eq. 4.1) with a simple translation model together with $\phi(x) = \frac{x^2}{2}$ [73]. Last but not least, part of their also work focused on a procedure to automatically select anatomical sites on moving body parts and to track multiple individual sites *but at separate sections* of the same video sequence, so seizures could be described by more than one anatomical site. In short, selection is done by first thresholding the optical flow image and applying morphological operations to the resulting blobs [90]. Next, the position of the anatomical site in the current section of the video sequence is set to be the center of the blob with either largest area or with maximum average velocity (equivalent outcomes for both choices were reported). The image block surrounding that position

is then tracked until a new site is automatically detected. We should note that the robust motion trackers reviewed here were shown to top all these preceding techniques, being the most successful tools to extract seizure signals and discriminate among myoclonic seizures, focal seizures and random movements, as was shown by the results in [70], at least when the CRCNS epilepsy dataset is concerned. Last but not least, part of their also work focused on a procedure to automatically select anatomical sites on moving body parts and to track multiple individual sites *but at separate sections* of the same video sequence, so seizures could be described by more than one anatomical site.

Still in the realm of infant seizure detection, Ferrari et al. [91] proposed to detect clonic seizures as a function of whole body periodic motions rather than body parts. First, they turn every 10 s window of the video into a 1-D signal by differentiating neighbor frames, thresholding and eroding the resulting image, and finally making the normalized non-zero pixel counts of each frame into a data point. The resulting signal is called an *average luminosity motion signal*. To estimate fundamental periods, they pass that signal through some hybrid auto-correlation process and look for points of minima, from which they also estimate the fundamental frequency. If the minima count has more than 1 element, the window is considered to present periodicity, and when three of such windows are observed in a row, a clonic seizure event is fired. The algorithm was shown to be consistent with clinical ground-truth on about 1800 three-consecutive half-interlaced (5 s overlap) 10 s windows from various recordings and different lengths.

#### 4.3.2.2 Assessment of neurodevelopment disorders

From infant epilepsy we turn to the study of neurodevelopmental disorders, where very recent vision-based infant motion capture work is being done towards obtaining objective measurements of the developmental markers that are crucial to the accurate diagnosis of these disorders. We begin with the work of Dogra et al. [74] and their attempt to improve the assessment of the Hammersmith Infant Neurological Examination (HINE) pulled-to-sit test via an automated process that estimates performance scores. The test is such that the infant initially lays on his or her back, so the head-neck and neck-torso segments are collinear. Next, the physician or another trained experimenter pulls up the baby by both arms, while the whole motion is observed and/or recorded (as in Fig. 4.3d). According to this study, the HINE protocol prescribes scores 0, 1 and 3 to a pulled-to-sit trial (while a score of 2 is not applicable) based on whether the baby: does not react enough to the pulling of the head, wobbles the head more than once during the pull, or otherwise keeps the head fairly aligned with the torso throughout the examination, respectively. In exchange, a score of 3 is considered to be a predictor of normal development, as opposed to the other two.

A systematic way of extracting body parts and computing the necessary angles and corresponding scores was thus proposed. Each trial is filmed by a lateral view camera that records grayscale frames. These videos are off-line processed in semi-automatic fashion: first, through the use of a touchpad, the system is fed with initial positions of the $p$ landmark body parts (head, shoulder and torso). Then,

tracking is done independently for each body part according to the following block matching method: the system searches frame $t+1$ for the $k$ nearest blocks (in terms of minimizing a pixel difference metric) to the coordinates of the considered body part at time $t$, that is, it produces $k$ possible candidate positions for that part in the next frame. The tracking algorithm is designed to keep only $k$ possible paths per level, leading to a tree with only $k$ leaf nodes/possible full-trajectories. This means it has to examine up to $k^2$ nodes at every level to choose the next $k$ ones to be expanded, but since $k$ is usually a small number, overall this represents small computational effort. This path-pruning process keeps the number of possible trajectories from growing exponentially as a function of the video length $t$.

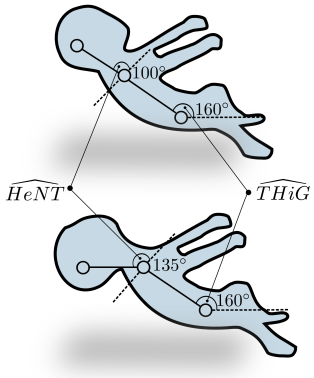
In the following step, with the body parts properly tracked, a simple geometric model and a few rules are used to evaluate temporal variations of the head-neck-torso $\frac{\partial \widehat{HeNT}}{\partial t}$ and torso-hip-ground $\frac{\partial \widehat{THiG}}{\partial t}$ angles at every $t$ and decide for HINE scores (Fig. 4.5a). The infant is considered not to react enough when $\widehat{HeNT} \geq 120°$ *throughout the whole trial*, and in that case a score of 0 was assigned[3]. Else, if $\left|\frac{\partial \widehat{THiG}}{\partial t}\right| > 30°$ is observed *more than once but not always during the examination*, a score of 1 is given. Finally, if $\left|\frac{\partial \widehat{THiG}}{\partial t} - \frac{\partial \widehat{HeNT}}{\partial t}\right| \leq 15°$ *throughout the entire examination*, a score of 3 is assigned. The system was tested on 43 infants and results were reported in terms of sensitivity and specificity out of comparing score assignments with ground-truth labels provided by participating physicians. The proposed markerless tracking worked only on 30 of the subjects, for which 5 false negatives

[3]In Table I of [74], the rule is re-stated as "the head always remain below 30° with respect to torso" which we interpret as $\widehat{HeNT} > 120°(= 90° + 30°)$, since that angle is at least 90°, according to the model diagram in the third figure of the same article, i. e. we just add 90° to the threshold of 30°.

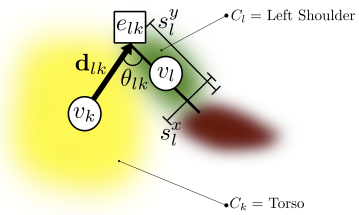
(abnormal pulled-to-sit classified as normal) and 1 false positive (misclassified normal pulled-to-sit) were reported, resulting in overall 80% of sensitivity and 89% of specificity. Micropore markers were placed on the other 13 subjects to help with the tracking, and as a result sensitivity and specificity rates went up to 92% and 96%, respectively (single false positive). Apparently, only one pulled-to-sit trial per subject was taken into account to produce HINE scores.

In the context of early assessment of Autism Spectrum Disorder (ASD), Spina et al. [2] have designed an markerless infant mocap system that measures arm asymmetry of toddlers and infants while they walk unsupported. The goal is to help in the early assessment of the disorder, following a recently discovered connection between asymmetric behavior in early age and the later development of ASD [83]. Their work has appeared previously in [75, 76] but [2] emphasize their markerless infant mocap solution, which was only briefly discussed in the previous manuscripts. A full camera-based system that reads in videos, tracks body parts, and computes 2-D joint positions and angles, plus the asymmetry data was developed in the study. The core of the approach is to estimate the child's pose in between frames, by modeling the child's body as an articulated Cloud System Model, a 4-tuple $\Omega = \{C, A, G, F\}$. We will discuss each component individually and later elaborate on their interaction. We may replace the original notation with our own whenever we find it simpler.

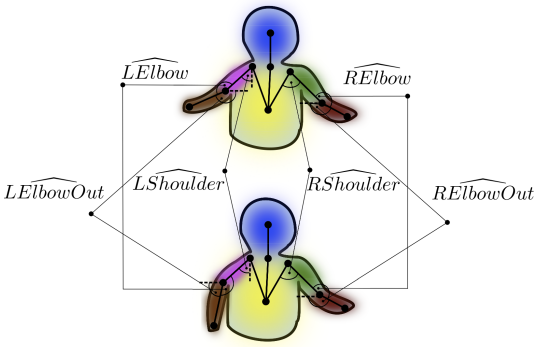
To begin, $C = \{C_1 \cdots C_n\}$ is a set of $n$ point clouds, each one formed by image pixels augmented with membership values within the $[0, 1]$ interval, so a point is defined as $\mathbf{x} = [x_r, x_g, x_b, x_l]$. These clouds are at the heart of the model and will approximate the child's body parts, hence, they may also be referred to as *body part*




**(a)** Head lag assessment of Dogra and colleagues [74]




**(b)** The ACSM of Spina et al. [2]




**(c)** Arm asymmetry assessment in [2, 75, 76]

**Figure 4.5:** Two vision-based markerless infant mocap models aiming at the assessment of neurodevelopmental disorders. (a) Quantizing head lag in pulled-to-sit head angles measured by Dogra et al. [74]. If throughout the whole trial the baby posture looks like the top figure, i. e. if $\left|\frac{\partial \widehat{THiG}}{\partial t} - \frac{\partial \widehat{HeNT}}{\partial t}\right| \leq 15° \ \forall t$, a HINE score of 3 is assigned, configuring a normal head pull. Bottom: for example when $|\widehat{HeNT}| \geq 120° \ \forall t$, a score of 0 is assigned, since the baby fails to pull the head up to keep it aligned with the torso. (b) The ACSM model of Spina et al. [2]: nodes $v_k$ and $v_l$, corresponding parent and child clouds $C_k$ and $C_l$ for the torso and left shoulder in (c), plus edges $e_{lk}$ of the skeleton graph. Note the parameter set $\mathbf{\Gamma}_l^t = \{s_l^y, s_l^x, \mathbf{d}_{lk}, \theta_{lk}\}$. (c) Quantizing symmetric (top) and asymmetric (bottom) arm behavior in walking. Differences of elbow and shoulder angles are used in the asymmetry measures $AS_f = sigm(|\widehat{LElbow} - \widehat{RElbow}|)$, $AS_u = sigm(|\widehat{LShoulder} - \widehat{RShoulder}|)$, $AS^* = max(AS_f, AS_u)$ and $AD_f = |\widehat{LElbowOut} - \widehat{RElbowOut}|$ in [2, 75, 76]. See text for details.

*clouds*. The delineation algorithm $A$ is responsible for establishing crisp boundaries for the clouds so they can be tracked across frames. It functions according to two major steps: first, it outputs a set of superpixels $R$ for the whole current frame being processed. Then, for each superpixel[4] $r \subset R$, the system inspects each of its member points $\mathbf{x}$, and if $\mathbf{x}$ is completely inside a body part cloud $C_l$, i. e. its membership $x_l = 1$, it is marked as belonging to cloud $C_l$. Otherwise, the system (1) populates two sets of points $S_l^f$ and $S_l^b$ that are 8-connected to $\mathbf{x}$ and belong to the interior (foreground) and exterior (background) of $C_l$ and (2) runs a graph segmentation algorithm that determines which of the sets $S_l^f$ or $S_l^b$ will have the member that produces the shortest path to $\mathbf{x}$ (as if these sets were competing for that point) and mark $\mathbf{x}$ to belong *in* or *out* of cloud $l$, accordingly. Everywhere, the weights of that graph are assigned as the average gradient between 8-connected points, so paths that cross image edges are expensive. Finally, the set of points belonging to $C_l$'s interior form a virtual crisp boundary.

Graph G is a tree model that enforces inter-cloud skeletal structure: each body part cloud $C_l$ has a corresponding vertex $v_l$, and a joint between $C_l$ and another adjacent body part $C_k$ (with vertex $v_k$) are represented as edges $e_{lk}$ (Fig. 4.5b). Here, $C_k$ will refer to $C_l$'s parent according to and hierarchical joint model which rooted at the torso. Each $v_l$ holds length and width parameters $s_l^y$ and $s_l^x$ which are the lengths of the first and second major axes of the cloud, in turn. An edge $e_{lk}$ holds displacement vector $\mathbf{d}_{lk}$ from the center of $C_k$ to $e_{lk}$ plus the angle $\theta_{lk}$ between

[4]Although $r$ is termed as superpixel, technically, it is a set of pixels augmented with memberships $\mathbf{x} = [x_r, x_g, x_b, x_l]$.

that displacement vector and the major axis of $C_l$, i. e. the angle between the two neighbor body parts. The displacement parameter is necessary to accommodate posture changes along the depth axis w.r.t. camera.

Functional $F$ is the final component of the model and it does the job of evaluating how well a cloud at time $t > 0$ will match another at time $t = 0$, by averaging $1 - \chi^2$ distances of corresponding histograms, for all clouds. In other words, $F$ imposes an appearance constraint that is enforced while clouds are tracked throughout the frames.

Tracking starts out with the user manually entering one contour for each of the $n$ considered body parts on the initial video frame. These contours are set as the initial boundaries of all clouds in $C$, and pixels $\mathbf{x} = [x_r, x_g, x_b]$ are assigned to their enclosing clouds. To augment $\mathbf{x}$ with membership scores $x_l$, first each cloud's contour is turn into an image mask, which is later converted into a signed distance map, where inside (or foreground), at-the-boundary and outside (or background) pixels result in negative, near-zero and positive values, respectively. That map is further processed by a function that thresholds negative and and positive distances into values 1 and 0, respectively, while near-zero ones are scored according to a logistic function (thresholds and parameters are custom-selected). With the clouds initialized, the next step is to initialize the skeletal graph G's parameters. For the $v_l$ vertices corresponding to the head and the torso, length and width parameters $s_l^y$ and $s_l^x$ are set to be proportional to the major axes of the clouds computed with Principal Component Analysis (PCA). It was noted that this initialization process does not deal well with toddlers' arms and legs, which proportions are more square-

like than rectangular when compared to older children and adults. The problem was circumvented by thinning out limb clouds, and only then applying PCA. For example, to find parameters for the upper arm and the adjacent forearm clouds, the system will first produce a collective arm skeleton using pixels from a binary image obtained from the points of both clouds. Then, for each individual cloud in turn, points that overlap with the arm skeleton are selected as the source data for PCA, so individual $s_l^y$ and $s_l^x$ parameters are computed as done for the head and torso. The last parameters to be initialized are the inter-cloud joint angles and displacements, that is, the set of parameters of edges $e_{lk}$. Overall, joints are properly placed by constraining their coordinates to be at the intersection between the major axis of the child cloud and simultaneously close to both its center and its parent's. The displacement vectors and joint angles arise naturally from knowing the edge position and the centers of the parent-child clouds, as can be seen from Fig. 4.5b. In particular, the displacement vector of the torso cloud, which is the root of $G$, is set to its position within the image frame.

With both the cloud system and the subset of parameters ${}^0\Gamma_l^t = \{s_l^y, s_l^x, \mathbf{d}_{lk}, \theta_{lk}\}$ initialized, the system is ready to estimate body part cloud dynamics from frame $t = 0$ to $t + \Delta t$, where $\Delta t$ is the temporal sampling interval. To optimize for parameters, a multi-scale search algorithm that minimizes functional $F$ given the initial solution is run: first, the search algorithm offers a small number of candidate solutions within some pre-specified parameter intervals $\Delta\Gamma_l = \{\Delta s_l^y, \Delta s_l^x, \Delta\mathbf{d}_{lk}, \Delta\theta_{lk}\}$. Then, for each candidate solution, the delineation algorithm is run, histograms are computed and matched against the corresponding ones at time $t = 0$, and the one

that best satisfies[5] appearance constraints imposed by $F$ is considered the optimal candidate solution, and the corresponding optimal parameter set constitutes the tracked pose at $t + \Delta t$. From $t > 0$ on, the same tracking procedure is utilized, except that the current frame's optimal $\Gamma_l^t$ is not used as the next ${}^0\Gamma_l^{t+\eta\Delta t}$ ($\eta > 0$), as it would be expected. Instead, the optical flow of non-background pixels was measured and used to warp[6] the cloud system $\Omega$ from $t$ to $t + \eta\Delta t$, leading to an initial ${}^0\Gamma_l^{t+\eta\Delta t}$ that is supposed to be closer to the next optimal $\Gamma_l^{t+\eta\Delta t}$.

Like in [74], the motion capture system was evaluated on the basis of how well a particular developmental score would be assigned to a trial, this time by considering inter-arm asymmetries. For such, a number of scores were proposed: $AS_f$, $AS_u$ result from applying a sigmoid function to left-right angle differences of elbows $|\widehat{LElbow} - \widehat{RElbow}|$ and shoulders $|\widehat{LShoulder} - \widehat{RShoulder}|$, respectively, whereas score $AS^*$ is defined as the maximum of the those two. Moreover, $AD_f = |\widehat{LElbowOut} - \widehat{RElbowOut}|$ tries to pick up situations at which the arms point to different directions. Video sequences of six babies were looked at, of which two were from infants (age $\leq$ 12 months old). These children were all previously classified to be at risk for ASD: one of them had an ASD sibling, a couple were premature, two others presented developmental delays, and another presented clear signs of ASD already at the age of 16 months. Each participant was represented by

[5] Even though histograms indirectly depend on parameters $\Gamma_l^t$, the average $\chi^2$ function in $F$ itself is not directly related to the parameters, so gradients and Hessians are apparently not available during optimization. Anyhow, it was mentioned that minimization is done on a gradient descent fashion.

[6] An example of warping would be to choose one of the robust motion trackers in [69], make $\mathbf{z} = \Gamma_l^t$ in Eq. 4.1, set $\nabla_{\mathbf{u}}(I^t)$ in Eq. 4.3 to the non-background optical flow and solve for each individual cloud.

at least 5 seconds of video data (150 frames) from either one or two segments per participant. In total, 10 segments of walking unassisted were processed. Asymmetry events were fired anytime $AS^* \geq 1.0$ and $AD_f \geq 45°$, a criterion that was chosen after manually inspecting their asymmetry scores and corresponding scores produced out of available ground-truth skeletons. Based on that rule, *Static* and *Dynamic Symmetries* (SS and DS, respectively) were computed for each participant, considering all of his/her sequences. The first metric is the percentage of a participant's number of frames where asymmetries were fired. The second is a smoothed version of the first: half-second windows were classified as asymmetric *whenever at least one of its frames was considered asymmetric*, and the percentage of such windows was output as DS.

In the first experiment, the percentage of automatically detected asymmetries compared to the same number of asymmetries computed from the ground-truth skeletons was inspected. Strong agreement was observed, except for a 15-month old participant with developmental delays. For this subject, the system scored $SS = 5\%$ and $DS = 21\%$ asymmetries, while the ground-truth based indicators scored 0% on both. In the second experiment, it was proposed that DS should be thresholded at 30% to classify a segment as being overall asymmetric or not, and compared the results with a clinician's evaluation. The system's outcome matched the expert's assessment in all cases but the first segment of the 16-month old with an ASD sibling.

## 4.4 Principle of dynamical stability and canonical postures

Table 4.1 summarizes the reviewed work on markerless motion capture of infants. Although very sparse – as Bhatt et al. [89] put it: "research concerning child behavior is still not explored in computer vision" – there are a handful of achievements worth noting. First, the results of pressure-based images suggest that coarse, holistic representations of infant motions may be sufficient to allow for the inference of canonical postures or behaviors: we saw success in finding prone and supine positions [1] as well as in differentiating amongst more complex stances such as sitting, lying on the back or crawling on the basis of global features [65]. This leads us to believe that, in computer vision, analogous results could be achieved from the use of depth sensors and global contour features, so this could be exploited in future endeavors. Data from the vision studies have also provided, if not proof, strong evidence that the state-of-the-art camera-based tracking technology allied with very simple pattern recognition such as blob detection, fundamental frequency estimation or block matching can foresee and classify epilepsy seizures and perhaps other events of medical concern, for example, SIDS (Sudden Infant Death Syndrome). Moreover, infant tracking has evolved from pressure-based crib stations to virtually unconstrained camera-based capture volumes, where infants can move freely in space while being recorded. Last but not least, current results of vision-based computations of head lags and arm asymmetries have established the possibilities that can result from solving markerless infant mocap, that is, they have shown that *the use of babies as models in the study of human behavior and its disorders is possible*

*also from a computational perspective*, since numerous other developmental markers could be measured using the same frameworks.

Nevertheless, the problem is far from being completely solved, and there is a lot of room for improvement. The pressure sensor paradigm has obvious limitations that prevent more complex infant tracking to be achieved, so we will concentrate our comments on the vision-based approaches to markerless infant mocap. First, it is currently very hard to discuss progress in terms of the quality of data being obtained, because the analyses presented by most studies is too qualitative or application driven, and there are almost no comparative performances reported. Except for [69], none of the methods has provided direct measurements of how well the estimated movement signals matched a ground-truth, and some have only presented plots for a small number of subjects. The two vision systems motivated by behavioral studies have reported evaluations only on the basis of the developmental markers they propose to measure, by checking their systems' results against expert assessments, but no comparison against previously labeled signals, even when available [2]. It is also very hard to visualize the scalability of these studies: except for two reported experiments, the maximum number of subjects tested was 5 (which is understandable, given that recruiting infants and having them collaborate in test sessions is very resource-demanding). In addition, the evaluation of success based on expert assessments has to be taken very carefully, since experts themselves often disagree upon a diagnosis. Take for instance the results of Hashemi et al. [76] for a visual tracking task: while they did observe strong agreement between the system's outcomes and a collaborating clinician's evaluation, at the same time they noted

disagreement between the diagnoses of that same clinician and the child/adolescent psychiatrist and two psychology students that also provided their inputs. Besides, because no cross-validation of thresholds and model parameters were reported in any of these studies, it is very hard to judge how well the predictions (scores, diagnoses) would generalize unseen data; it could be the case that parameters were merely overfitting the assessment of the participating experts.

Indeed, the problem remains very challenging, which we can tell from the need for manual initialization [2, 74–76] or the eventual resorting to supplemental micropore markers to improve block matching performance [74]. There are also behavior constraints: in [69], the baby must be in supine position with the camera on top, the model of [74] is planar and lateral, and [2] will process unassisted walking but not crawling. Ideally, we will want a markerless infant mocap system to be able to capture data from these children in a variety of postures and orientations with respect to the camera. Another point worth discussing is that the reviewed vision-based studies assume that image block motion is always a result of body part motions, which is not generally the case: data sessions of infants are usually highly-staffed, so one should expect the infant to interact with one or two people (Figs. 4.8d, 4.8f and 4.8j) and to play with objects of various natures (Figs. 4.8b, 4.8e and 4.8f). Explicit modeling of motions of other humans and objects may be necessary.

Conceptually speaking, except for Harada and others [1], *none of the other methods utilized the unique physique properties of infants and/or the occurrence of postures that are more likely to be displayed by infants to significantly bias the tracking process*. We agree with their position that "in order to recognize the infant

behavior, it is necessary to base on the characteristics of infant's unique physique". This explains why state-of-the-art motion capture systems should fail if tried on a very young child, and unless modeling physique is incorporated, current infant-targeted approaches such as [2] could quickly reach an applicability plateau[7]. For example, during infancy, arms and legs are of very similar lengths and the head is at its biggest size with respect to the rest of the body (Fig. 4.7a). As a result, the postures displayed by infants will be very peculiar: hands will often reach for the feet, and there will be a lot of fast, jerky movements around the elbow joints and neck. They will also be expected to crawl, roll or drag themselves. A number of these postures will be re-occurring, as they reflect goals that are common to most infants under similar environmental circumstances, somewhat in accordance with the *principle of dynamic stability* advocated by Esther Thellen and her collaborators [92] p.563:

> "Behavior fluctuates, but within limits. That is, organisms tend to show a delimited number of behavioral patterns, which within certain boundary configurations will act like dynamic attractors. These states will be the preferred configuration from a number of initial conditions, and they will be relatively resistant to perturbation. As a consequence of this dynamic assembly, developing organisms remain flexible in the face of tasks, but only within the constraints of their energetically stable possible states."

[7]Although Spina et al. [2] initialize arms and legs' orientation parameters of their skeletal graph to reflect their observation that "limb proportions are different than those of the adults" (see Sec. 4.3.2.2, p. 75), their system is not explicitly aware that it is tracking an infant subject.

These preferred configurations, attractors, or states of energy minima, as Thellen puts, could be understood as clusters or hidden states that we here term as *canonical postures* (Fig. 4.8). Canonical postures would then condition the parameters of the infant's physical model, that is, some postures or motions should be more likely to be observed than others given that a certain canonical posture has been observed. Spina et al. [2] could have used this concept to narrow down the search space for initial parameters (Sec. 4.3.2.2, p. 76), by finding the solution triple $\{{}^0\Gamma_l^t = {}^0\gamma_l^t, \Delta\Gamma_l^t = \Delta\gamma_l^t, \Lambda^t = \lambda^t\}$ that maximizes the joint probabilities of the current canonical posture $\Lambda^t \in \{\text{crawl, sit, kneel, stand, prone, supine}\}$, initial cloud parameters ${}^0\Gamma_l^t$ and parameter ranges $\Delta\Gamma_l^t$, conditioned to the current cloud parameter set $\Gamma_l^{t-\eta\Delta t}$, the optical flow of non-background pixels $\nabla_{\mathbf{u}}(I_l^t)$, the current shape $S^t$ of cloud $C_l$ (recall that global, pressure-based, shape-like features were successful in discriminating postures) and the previous canonical posture $\Lambda^{t-\eta\Delta t}$, or:

$$\underset{{}^0\Gamma_l^t={}^0\gamma_l^t, \Delta\Gamma_l^t=\Delta\gamma_l^t, \Lambda^t=\lambda^t}{\operatorname{argmax}} P\left({}^0\Gamma_l^t, \Delta\Gamma_l^t, \Lambda^t | \Gamma_l^{t-\eta\Delta t}, \nabla_{\mathbf{u}}(I_l^t), S^t, \Lambda^{t-\eta\Delta t}\right)$$

but because:

$$P\left({}^0\Gamma_l^t, \Delta\Gamma_l^t, \Lambda^t | \Gamma_l^{t-\eta\Delta t}, \nabla_{\mathbf{u}}(I_l^t), S^t, \Lambda^{t-\eta\Delta t}\right) = \frac{P\left({}^0\Gamma_l^t, \Delta\Gamma_l^t, \Lambda^t, \Gamma_l^{t-\eta\Delta t}, \nabla_{\mathbf{u}}(I_l^t), S^t, \Lambda^{t-\eta\Delta t}\right)}{P\left(\Gamma_l^{t-\eta\Delta t}, \nabla_{\mathbf{u}}(I_l^t), S^t, \Lambda^{t-\eta\Delta t}\right)}$$

and also:

$$\underset{{}^0\Gamma_l^t={}^0\gamma_l^t,\Delta\Gamma_l^t=\Delta\gamma_l^t,\Lambda^t=\lambda^t}{\operatorname{argmax}} \frac{P\left({}^0\Gamma_l^t,\Delta\Gamma_l^t,\Lambda^t,\Gamma_l^{t-\eta\Delta t},\nabla_{\mathbf{u}}(I_l^t),S^t,\Lambda^{t-\eta\Delta t}\right)}{P\left(\Gamma_l^{t-\eta\Delta t},\nabla_{\mathbf{u}}(I_l^t),S^t,\Lambda^{t-\eta\Delta t}\right)}$$

$$= \underset{{}^0\Gamma_l^t={}^0\gamma_l^t,\Delta\Gamma_l^t=\Delta\gamma_l^t,\Lambda^t=\lambda^t}{\operatorname{argmax}} P\left({}^0\Gamma_l^t,\Delta\Gamma_l^t,\Lambda^t,\Gamma_l^{t-\eta\Delta t},\nabla_{\mathbf{u}}(I_l^t),S^t,\Lambda^{t-\eta\Delta t}\right) \tag{4.4}$$

the problem can thus be simplified to maximizing the joint probability distribution of Eq. 4.4.

Another important aspect that could have been better explored by vision systems is the role of *age* or *developmental stage* in movement prediction; the human body and mind are perhaps growing at its fastest rate during infancy, as can be noted from Fig. 4.7b. According to the more traditional Piaget's theory of cognitive development, infancy corresponds to the first half of the sensorimotor stage, when an individual's acting abilities range from basic reactions to prehension coordination, or even walking. In terms of cognition, during that period, the child learns important concepts such as object persistence and how to associate basic actions to consequences, intentionality and some language. Again, age could be incorporated to the model of Spina et al. [2] as vector of variables $A$ that would somehow encode information on the infant's developmental stage (for example, by processing the results of standard assessment tests) plus condition canonical postures and parameter ranges, so the problem would now become:

$$\underset{{}^0\gamma_l^t={}^0\Gamma_l^t,\Delta\gamma_l^t=\Delta\Gamma_l^t,\lambda^t=\Lambda^t}{\operatorname{argmax}} P\left({}^0\Gamma_l^t,\Delta\Gamma_l^t,\Lambda^t,\Gamma_l^{t-\eta\Delta t},\nabla_{\mathbf{u}}(I_l^t),S^t,\Lambda^{t-\eta\Delta t},A\right). \tag{4.5}$$

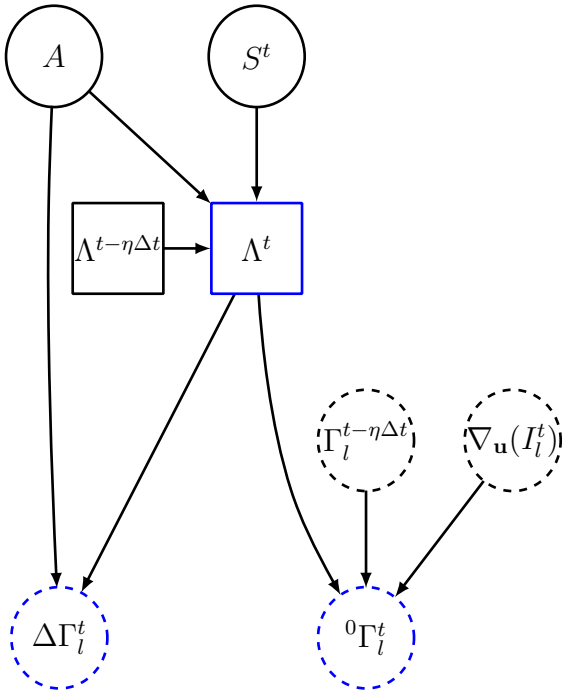


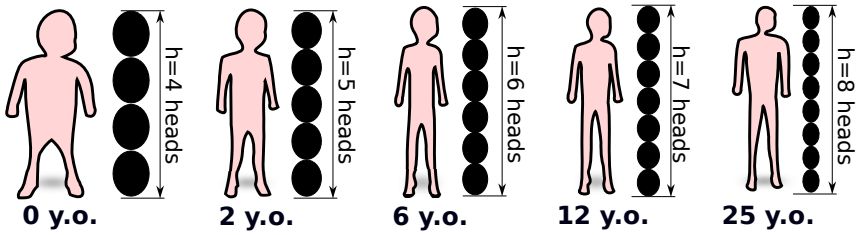

**Figure 4.6:** Bayesian network proposed to extend the initialization of parameters of Spina et al. [2]. Discrete variables are shaped as squares, continuous ones as circles. Blue variables are the ones jointly estimated. Dashed variables correspond to the components currently utilized in their model, which pre-defines $\Delta\Gamma_l^t$ and determines ${}^0\Gamma_l^t$ by means of warping, given the dense optical flow $\nabla_{\mathbf{u}}(I_l^t)$ and the previous cloud parameters $\Gamma_l^{t-\eta\Delta t}$; for more details, see the method's review in the previous section. Added variables incorporate infant's physique information by encoding it as canonical postures $\Lambda^t, \Lambda^{t-\eta\Delta t} \in \{\text{crawl, sit, kneel, stand, prone, supine}\}$ enhanced with age-related information. By exploring the fact that global shape-like features such as blobs were successful in discriminating postures evinced by previous work on pressure-based markerless infant mocap, we can partially condition a canonical posture on contour features $S^t$. Nowadays, contour data can be more easily obtained by the use of depth sensors, which have proven efficient when it concerns human pose estimation [66]. The fact that canonical postures will appear differently given the stage of development suggests that these postures should also be conditioned on age features $A$. Moreover, given appropriate temporal sampling, canonical shapes should be coupled from one instant to another, therefore $\Lambda^t$ should also depend on $\Lambda^{t-\eta\Delta t}(\eta > 0)$. See text for further discussion.

Note that, by exploring the independencies prescribed by the Bayesian network of Fig. 4.6, the joint distribution can be factored as:

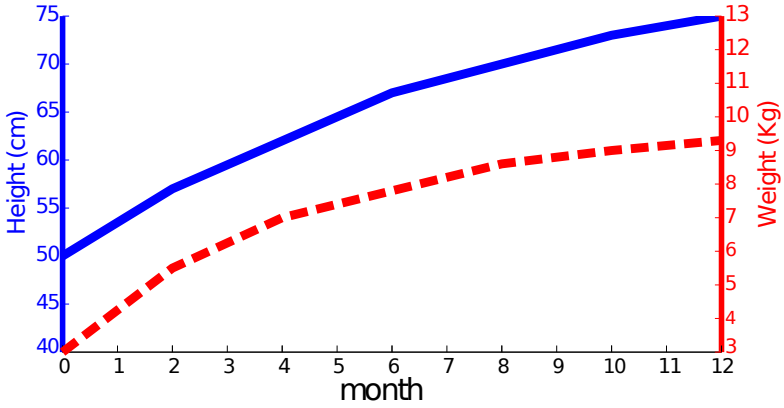
$$P\left({}^{0}\Gamma_l^t, \Delta\Gamma_l^t, \Lambda^t, \Gamma_l^{t-\eta\Delta t}, \nabla_{\mathbf{u}}(I_l^t), S^t, \Lambda^{t-\eta\Delta t}, A\right) =$$

$$\underbrace{P\left({}^{0}\Gamma_l^t | \Lambda^t, \Gamma_l^{t-\eta\Delta t}, \nabla_{\mathbf{u}}(I_l^t)\right)}_{\substack{\text{Next initial solution given} \\ \text{current canonical posture and optical flow} \\ \text{plus previous cloud parameters}}} \cdot \underbrace{P\left(\Lambda^t | A, S^t, \Lambda^{t-\eta\Delta t}\right)}_{\substack{\text{Current canonical posture} \\ \text{given features of age, contour} \\ \text{and the previous canonical posture}}} \cdot \underbrace{P\left(\Delta\Gamma_l^t | A, \Lambda^t\right)}_{\substack{\text{Initial parameter ranges} \\ \text{given age features and} \\ \text{current canonical posture}}} \cdot K, \tag{4.6}$$

where $K = P(A) \cdot P(S^t) \cdot P(\Gamma_l^{t-\eta\Delta t}) \cdot P(\Lambda_l^{t-\eta\Delta t}) \cdot P(\nabla_{\mathbf{u}}(I_l^t))$ is a factor that does not depend on the choice of the to-be-optimized variables and it is in practice disregarded.



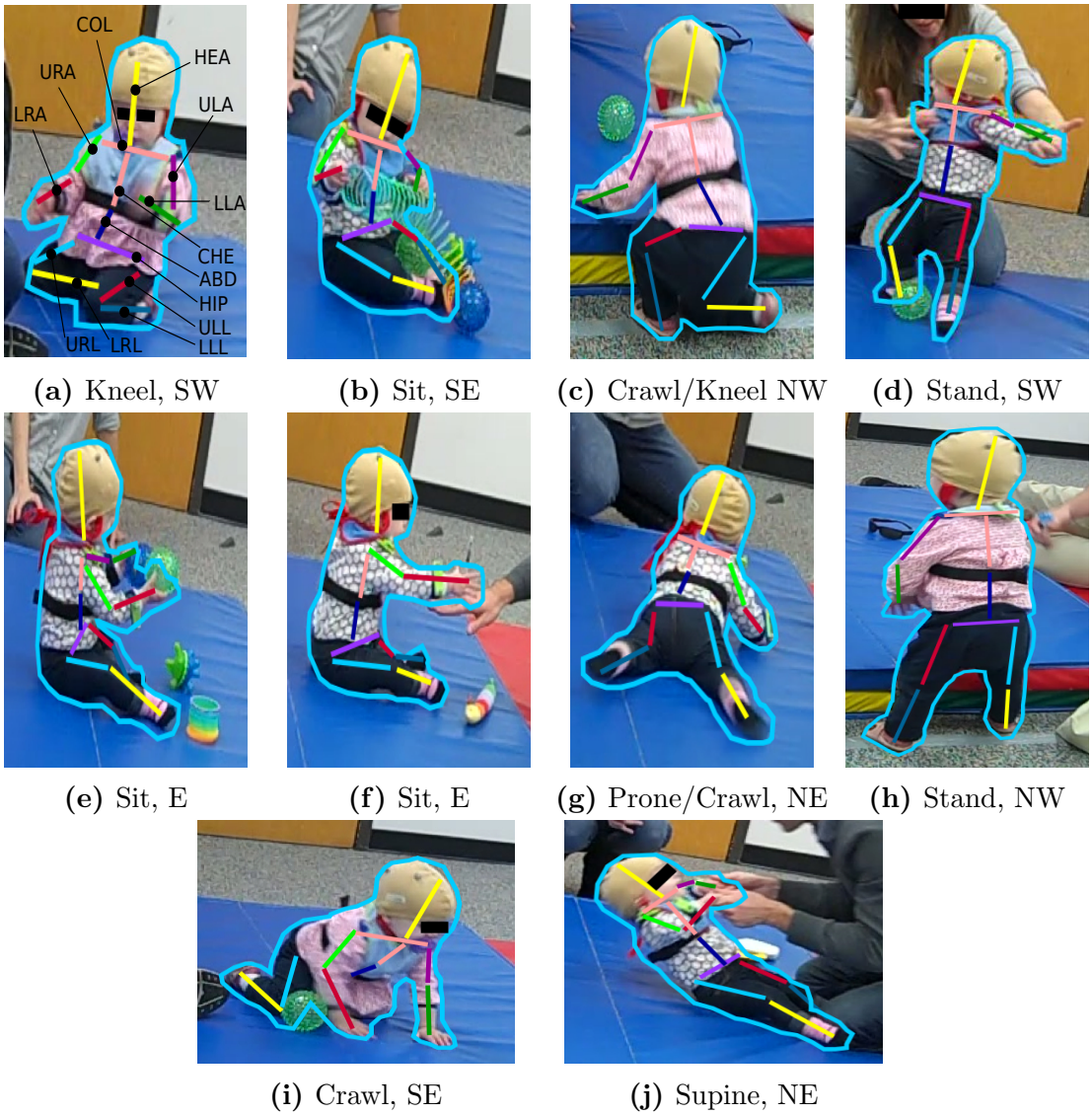

**(a)** Changes in the human physique as a function of age (male). Height $h$ is displayed as a function of the head length. Data from [1].




**(b)** Cross-gender average height (blue) and weight (red) growths during infancy (0-12 months). Data adapted from [1].

In summary, the intuition is the following: contour data, as observed in certain pressure-based studies, and age/developmental stage data should help us estimate probabilities of the next canonical posture assumed by the acting infant. This estimation is smoothed by the previous canonical posture, in a Markovian fashion. Suppose now, that all that information led us to believe that crawling is the most probable current canonical posture (inbound links to $\Lambda^t$ on Fig. 4.6); in that case, parameter search would then bias subspaces that corresponded to values of initial cloud configurations ${}^0\Gamma_l^t$ and ranges $\Delta\Gamma_l^t$ observed to co-occur with crawling in the training data (outbound links from $\Lambda^t$ on the same Figure). In analogy, age/developmental stage would also impact on deciding for the range of motion of the infant, thus the direct link between the two. It should be easy to notice the physical and behavioral constraints herein proposed (top sub-graph nodes of 4.6 represented with solid lines) could enhance virtually any infant mocap model, and not just the articulated cloud system of Spina et al. [2].

Finally, data from [1] show that the infant's body grows a great deal during the first year, when they get about 50% taller and three times heavier (Fig. 4.7b). Meanwhile, the height, as a function of head length does not vary much within the same period (4 heads). In fact, from 1 to 2 years old, the head changes from 25% to 20% of the infant's height and reaches 12.5% by the age of 25 (Fig. 4.7a). This tell us that the head grows slower relatively to the body, and during infancy in particular, it could be assumed to have constant length. As a consequence, both the infant's height and the lengths of body parts could be expressed in head length units. In other words, a system that could somehow find out about the head length




**(a)** Kneel, SW **(b)** Sit, SE **(c)** Crawl/Kneel NW **(d)** Stand, SW

**(e)** Sit, E **(f)** Sit, E **(g)** Prone/Crawl, NE **(h)** Stand, NW

**(i)** Crawl, SE **(j)** Supine, NE

**Figure 4.8:** Example of two subjects and the different canonical postures, cardinal directions and corresponding manually-labeled contours and body segments after a skeletal model compatible with the Eskhol-Wachman system [2, 81]. As in (a), selected body segments are: ULL (upper left leg), LLL (lower left leg), URL (upper right leg), LRL (lower right leg), HIP (segment that transverses the hips), ABD (abdomen), CHE (chest), ULA (upper left arm), LLA (lower left arm), URA (upper right arm), LRA (lower right arm), COL ( collar) and HEA (head length segment).

would be able to use it as an additional constraint in the tracking of limbs. This could again be translated into another improvement to Spina et al. [2]: first, one would try to detect the head and infer its length $s^y_{head}$, something that could be partially solved with current face detection technology. Second, one would learn the

relationship $\ell(s_l^y, s_{head}^y)$ between lengths of each body part (main axis parameters $s_l^y$) and the head, for example by regressing some parametric model $\hat{s}_l^y = f(s_{head}^y)$. Thus, $\hat{s}_l^y$ would work as a ground-truth, so while estimating $s_l^x$ and $s_l^y$ (more details in Sec. 4.3.2.2, p. 75) one would extend the PCA error criterion to include a term that penalized candidate solutions $s_l^y$ based on the disparity $\hat{s}_l^y - s_l^y$.

### 4.4.1 Canonical posture classification

So far, we have accepted the evidence that holistic contour features properly characterize the previously outlined canonical postures; we then tested this hypothesis by running a linear posture classifier on labeled contours (note the blue contours in Fig. 4.8) given our selection of canonical postures and a choice of features that describe contours as a whole. Note that, unless the hypothesis holds, the central $\Lambda^t$ node in Fig. 4.6 will be of limited use.

The first set of contour features we tried was *shape context* [93]. These features are standard in computer vision, and characterize a contour by tessellating a neigborhood around each of its points and counting the number of points that fall within each of the cells. A shape context feature vector is often long and sparse, of the order of the number of cells times the number of points in the contour: in our experiments, we tessellated the contours with 5 distinct radii and 12 orientations around 50 points uniformly sampled through the contour, which led to 3000 features per contour in the dataset. This number is much bigger than the number of examples in our data, so the resulting feature matrix ($48 \times 3000$) became singular and

made linear discrimination inviable. We have then shrunk these 3000 dimensions by projecting the data onto subsets of the first 30 right singular vectors $\mathbf{v}$ of the shape context feature matrix in the following fashion: for $K = \{1, 2 \dots 30\}$, the data was projected on sets of vectors $\{\{\mathbf{v}_1\}, \{\mathbf{v}_1, \mathbf{v}_2\} \dots \{\mathbf{v}_1 \dots \mathbf{v}_{30}\}\}$ respectively. In other words, $K$ is the new dimensionality of the compressed shape context features. We then trained classifiers for the 30 $K$ settings and ran a leave-one-out validation experiment: a hit rate was obtained from averaging over the individual performances of the classifier on each left-out examples. The top hit rate was obtained when $K = \{4, 5, 7, 8, 9\}$, for each of which $hr = 0.75$. The number of occurrences of each posture in the data was *crawl*=11, *sit*=17, *stand*=10, *prone*=2, *supine*= 4 and *kneel*=4. The average hit rate per posture within the reported $K$ range was: $hr = \{0.6136, 0.6765, 0.8750, 0, 1, 0.6875\}$, respectively.

We also tried a set of segment attributes computed from binary masks that result from the manually labeled contours. These attributes are currently being developed by a peer group in the Maryland's Computer Vision Lab [94]. We tried 6 of their attribute features: roundness, straightness of boundaries at 6 different scales, elongatedeness, convexity and segment rotation. When we tried to discriminate postures based on the attribute set alone, we saw poor results: the average hit rate was only $hr = 0.271$. However, when combined with the compressed shape context features, different configurations of these features were shown to improve the previous best hit rate to as high as $hr = 0.771$. These improvements ought to be credited to the simple elongatedness attribute alone (which is the length of the skeleton divided by the average width of the segment) as can be seen from

Table 4.2. These effects can also be seen from the confusion matrices in Table 4.3, where the trace of the matrix increased of 2.8 units, meaning approximately 3 (out of 48) more postures were correctly classified with the mixed setting. These initial numbers *suggest that for the purposes of discrimination, a good canonical posture description appears to profit of a hybrid feature space with both low-level and mid-level attribute-based cues.*

## 4.5 Conclusions and final remarks

Our major goals were to (1) discuss the current demands for infant behavior data (2) provide evidence that infant movement acquisition has to be as least invasive as possible, and defend the position that (3) measuring human movement has to be rethought to deal with infants. We went over the literature and stressed the importance of making use of results in developmental psychology as guidance; in particular, we suggested the use of canonical postures as means to improve existing pose estimation systems, and selected a number of such postures based on observed infant behavior. We also showed that the selected postures can be classified from a hybrid feature set consisting of holistic contour features allied with mid-level segment attributes.

**Table 4.1:** Summary of previous approaches to markerless motion capture of infants.

| Study | Year | Application | Sensor | Tracking | Events | Test data |
|---|---|---|---|---|---|---|
| Weinberg et al. [64] | '98 | Monitoring | Pressure | Limb motions | Behavior | Not informed |
| Harada et al. [1] | '00 | Monitoring & biometrics | Pressure | Head, chest, hip and abdomen | Postures and behavior | 5 sequences of 5 subjects |
| Bhatt et al. [89] | '03 | Monitoring | Optical | Hands, head, eyes, mouth | Danger, possible danger | 10 sequences of 1 subject |
| Karayiannis et al. [67–73] | '01 to '05 | Diagnosis | Optical | Motion signals of arms and legs | Myoclonic and focal seizures | 36 sequences |
| Ferrari et al. [91] | '10 | Diagnosis | Optical | Global body motion signal | Clonic seizures | 1823 frames |
| Boughorbel et al. [65] | '10 | Monitoring | Pressure | Not available | Postures and behavior | 3 sequences of 1 subject |
| Dogra et al. [74] | '12 | Diagnosis | Optical | Head-torso angles | Pulled-to-sit scores | 43 subjects |
| Hashemi et al. [75, 76], Spina et al. [2] | '12 and '13 | Diagnosis | Optical | Body parts and angles | Arm asymmetry scores | 6 sequences (150 frames) of 6 subjects |

| | | Hit rate vs. SC (0.75) | |
|---|---|---|---|
| 0 | *All* | 0.729 | ↓ |
| 1 | *Roundness* | 0.667 | ↓ |
| 2 | *StrBound*1 | 0.667 | ↓ |
| 3 | *StrBound*2 | 0.708 | ↓ |
| 4 | *StrBound*4 | 0.708 | ↓ |
| 5 | *StrBound*8 | 0.708 | ↓ |
| 6 | *StrBound*16 | 0.729 | ↓ |
| 7 | *StrBound*32 | 0.708 | ↓ |
| *8 | *Elongatedeness* | 0.771 | ↑ |
| 9 | *Convexity* | 0.729 | ↓ |
| 10 | *Rotation* | 0.729 | ↓ |

| | Hit rate vs. SC (0.75) | |
|---|---|---|
| $2-7$ | 0.667 | ↓ |
| $1, 8, 9$ | 0.75 | – |
| $1, 8$ | 0.75 | – |
| $1, 9$ | 0.708 | ↓ |
| $*8, 9$ | 0.771 | ↑ |

**Table 4.2:** Performance of the compressed shape context (SC) features combined with segment attributes. First (top table) we tried augmenting SC with each attribute individually, and noted that the elongatedness attribute was the only one to improve the SC-only performance ($0.75 \rightarrow 0.771$). Next (bottom table) we tried augmenting SC only with straightness of boundaries attributes ($2-7$) and with combinations of roundness (1), elongatedness (8) and convexity (9). The best performances arose from sets of attributes that had elongatedness as the only commonality, thus leading us to conclude that it was the cause of SC's improvement.

| | Cra | Sit | Sta | Pro | Sup | Kne | |
|---|---|---|---|---|---|---|---|
| **Cra** | 6.8 | 1.0 | 0.0 | 1.4 | 0.8 | 1.0 | / 11 |
| **Sit** | 2.8 | 11.8 | 0.0 | 2.0 | 0.0 | 0.4 | / 17 |
| **Sta** | 0.0 | 0.0 | 8.8 | 0.0 | 0.0 | 1.2 | / 10 |
| **Pro** | 0.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | / 2 |
| **Sup** | 0.0 | 0.0 | 0.0 | 0.0 | 4.0 | 0.0 | / 4 |
| **Kne** | 1.0 | 0.0 | 0.2 | 0.0 | 0.0 | 2.8 | / 4 |

| | Cra | Sit | Sta | Pro | Sup | Kne | |
|---|---|---|---|---|---|---|---|
| **Cra** | 7 | 1 | 0 | 1 | 1 | 1 | / 11 |
| **Sit** | 3 | 13 | 0 | 1 | 0 | 0 | / 17 |
| **Sta** | 0 | 0 | 9 | 0 | 0 | 1 | / 10 |
| **Pro** | 0 | 1 | 0 | 1 | 0 | 0 | / 2 |
| **Sup** | 0 | 0 | 0 | 0 | 4 | 0 | / 4 |
| **Kne** | 1 | 0 | 0 | 0 | 0 | 3 | / 4 |

**Table 4.3:** Confusion matrices summarizing the classification results of canonical postures (cra=crawl, sit, sta=stand, pro=prone, sup=supine, kne=kneel) from compressed shape-context features-only (top) and augmented with the elongatedeness segment atribute (bottom). The left matrix was computed based on the average per-posture hit rates of $K = \{4, 5, 7, 8, 9\}$, for which the same best overall hit rate was observed ($hr = 0.75$). Blue numbers in the diagonal present the average number of correct classifications per canonical posture. In red, we note that the two prone samples were incorrectly assigned to sits. The right matrix was computed based on the average per-posture hit rates of $K = 7$, for which the best per-posture hit rate improvement was seen for the combined features ($hr = 0.771$). The green values along the main diagonal indicate more correct classifications of crawls, sits, stands, prones and kneels.

# Chapter 5: Clinical descriptions of infant behavior can help predict risk for neurodevelopmental disorders

## 5.1 Introduction

Here we continue the discussion of recording movement data from human infants; in the last chapter we saw that relatively recent results have linked the presence of sensorimotor impairments in infancy to the manifestation of neurodevelopmental disorders such as Autism Spectrum Disorder (ASD) and Cerebral Palsy [58] a few years later in the child's life (see [95] and more recently [37]). This exciting new understanding has opened an opportunity for the administration of early therapies that can prevent typical traits from advancing and help including these individuals in society. This can improve the quality of life of several families and decrease healthcare costs, especially when we consider that disorders like ASD are become more and more prevalent [78].

To assess the risk for ASD and related disorders, the clinician will carefully observe how the child behave in their natural environment or when performing batteries of tests. In many cases, the diagnosis is not fully conclusive, and hard to quantify, and sometimes more than one evaluation is needed. Current computer

vision has been shown to help diagnosing ASD behavior in children [75], but the acquisition of movement data from infants is generally very challenging and demands time and resources that are often unavailable (see Chapter 4).

However, while inspecting an infant performing a task, the health professional will often create descriptions of how they perceive the way that child reacts to the behavioral tasks and how they conform to developmental milestones, for example a sentences like "This position doesn't require him to work as hard against gravity so he is comfortable here and looks much more symmetrical." (Fig. 5.1), which are nothing but freely available linguistic counterparts to the actual, low-level movement signals.

Here we begin to study how these descriptions could be used as a proxy to the movement signals observed in infant behavioral trials, in the hope that it will trade low-level description for an easier-to-obtain, interpretable and multi-centered representation of tasks. Our current results show that, at the level of sentences, traditionally used text features such as term frequencies and TF-IDF computed from unigrams and bigrams can be potentially helpful.

## 5.2 Predicting risk for atypical development

When assessing risk for atypical development, the clinician will typically subject the child to a behavioral battery of tasks and make a judgement based on his or her impression and expertise. Computationally speaking, this configures a binary classification problem where one would learn a mapping between task-related movement

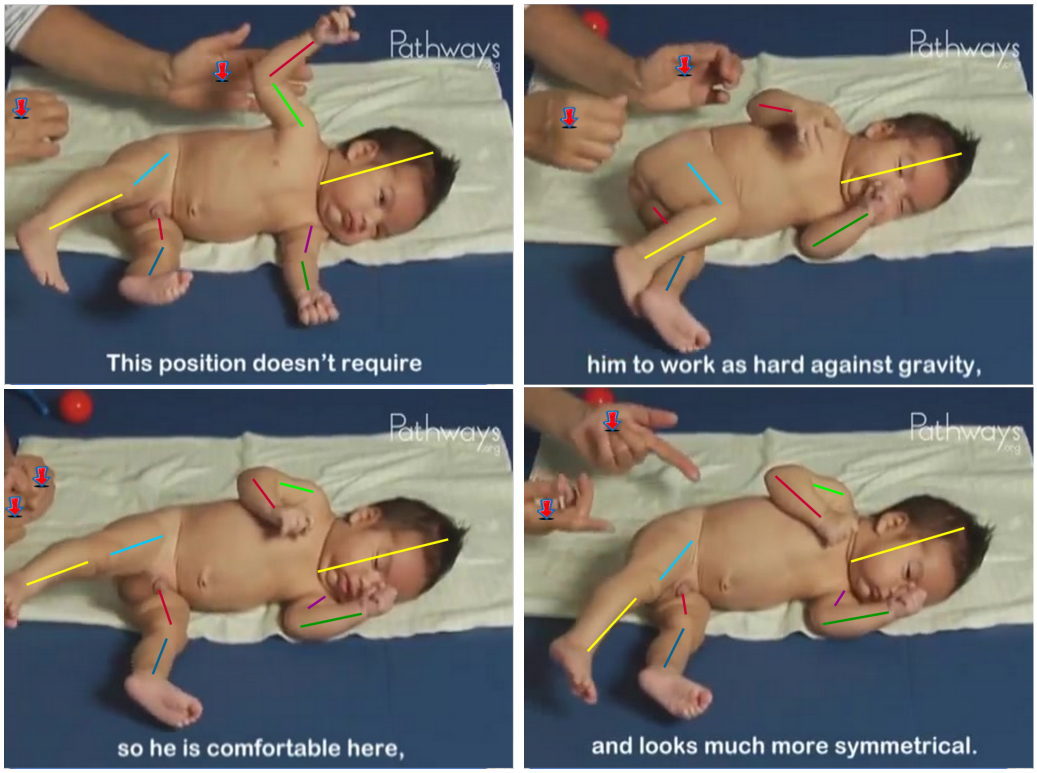



**Figure 5.1:** Top: selected frames for one of the sentences from the Pathways.org dataset describing Owen (atypical) at 2 months of age performing part of the Sidelying task. Overlaid colored sticks and small arrows are annotations of the baby's body parts and the tester hands, respectively.

features to labels *typical* or *atypical* that are known for a number of individuals, and use this learned model to assign labels to sets of features for which labels are unknown. Commonly, these *movement features* will be derived directly from the low-level movement data like the kinematics of body parts [75], but here we propose instead to use language as a proxy to movement, which we refer to as *language features*, *text features* or just *text*.

Using the methodology of [97] tailored to text classification rather than regression, we studied the discriminative power of text features, as we describe next.

| | | Train | | | Test | | |
|---|---|---|---|---|---|---|---|
| Task | Vocab | Typical | Atypical | Total | Typical | Atypical | Total |
| Supine | 474 | 97 | 89 | 186 | 22 | 24 | 46 |
| Sidelying | 508 | 109 | 100 | 209 | 10 | 13 | 23 |
| Prone | 495 | 94 | 95 | 189 | 25 | 18 | 43 |
| Pull-to-sit | 531 | 105 | 100 | 205 | 14 | 13 | 27 |
| Sit | 517 | 103 | 99 | 202 | 16 | 14 | 30 |
| Horizontal suspension | 534 | 110 | 104 | 214 | 9 | 9 | 18 |
| Protective extension | 531 | 110 | 103 | 213 | 9 | 10 | 19 |
| Stand | 530 | 105 | 101 | 206 | 14 | 12 | 26 |

**Table 5.1:** Leave-one-task-out data splits. Tasks for which we already have movement data appear in blue.

## 5.3 Experiments

### 5.3.1 Data, features and setup

We begin by introducing the Pathways.org dataset, the first public dataset with text descriptions of typical and atypical infants engaged in behavioral tasks in a longitudinal fashion.[1] Besides text, the data include annotated body parts of children, tester and objects central to the tasks. The dataset was produced by our team by manually processing three of the educational videos in [98]. These videos feature two subjects, Marty (typical) and Owen (atypical) performing 8 different tasks when they were 2, 4, and 6 months old respectively. These tasks are typical of infant behavioral battery tests and assess sensori-motor and social development: Supine, Sidelying, Prone, Pull-to-sit, Sit, Horizontal suspension, Protective extension and Standing. To create the actual data, we manually annotated each individual sen-

[1]These videos were originally intended to help parents to learn how to interact with their babies and watch out for developmental delays, but the quality of filming is so high, that we realized it could be used as scientific data.

tences and sampled 20 image frames per sentence. These sentences were grouped by task and collapsed over ages, so there were 8 sets of sentences (rows of Table 5.1) both for Marty and Owen ("Typical" and "Atypical" columns in the same table).

We experimented with term frequencies, TF-IDF and log1p measurements of lemmatized unigrams plus n-grams. After computing features for each sentence, we set up a typical vs. atypical binary classification experiment evaluated using what we call *leave-one-task-out* cross-validation, that is, we trained a linear SVM on a set of sentences coming from 7 out of the 8 considered behavioral tasks and tested on the remainder (one versus all). This let us create individual models for each task and discuss the results in terms of what we know about these tasks. The sentence/behavior distribution for the second task and its breakdown for typical and atypical is shown on Table 5.1.

## 5.4 Results and Analysis

Quantitative F-measure results for text features can be seen from Table 5.4: these numbers are all beyond chance, and mostly within 0.7–0.8, with the exception of Horizontal suspension and Protective extension whose scores were 0.9 or greater. The best results came from features based on term frequencies or log1p, and the linear SVMs were by far the best performing model overall, with the exception of the Sidelying task.

Table 5.3 shows selected tasks (columns) and top-20 words more associated with atypical and typical sentences (first and bottom rows, respectively) based on

the SVM models. The more positive the weight assigned by task's SVM to a word, the more it relates to atypical sentences, and the opposite for typical sentences.

The top-scoring bags-of-words shown on that table may help explain the performance numbers. First, text features seem to going beyond being a mere proxy to movement, as we first thought would be the case; we rather see that *text is enriching movement description* (whose words appeared labeled as *mov*, *qual* and *body* on Table 5.3) *by incorporating other information also related to the physics of the movement that would be very hard or near impossible to grasp directly from movement signals*, because they are very abstract. Examples of these words are **freedom** (to perform some movement) and **abl_sustain** (able to sustain, an indication of strength). *Text is also conveying information on the state of mind and cognition that concurs with/is part of the task*, for example through words like **calm** (revealing how comfortable the infant is while engaged in the task) or even **investigate**. Text is also incorporating a third-party's perspective on the movement that is virtually impossible to obtain from the movement low-level data. *This expert's sentiment towards the child's performance* is also evident from top-scored words: **overshoot**, **import(ant)**, **poor**, **hard**, **lower** (than), **greater** (than), **productive** and so on.

A second explanation to why language would help discriminate typicals and atypicals come from how *top words seems to be very well-locked to tasks they characterize*; for example, Supine, Prone and Sidelying (Table 5.3) are tasks that demand a postural control, while at tummy up/tummy down positions or rolling on the surface, respectively, so there are usually differences in the symmetry of behavior, balance and the ability to sustain weight between typical and atypical individu-

als, and these are reflected by task-related textual counterparts like **antigravity**, **weight**, **posture**, **thirty (degrees)**. Differences in attention are reflected by task-related words like **toy** (utilized to check visual engagement with objects) **vision** and **looking**. Some of these *top words refer to the exact qualities that help distinguish typical from atypical behavior*. Tasks like Pull-to-sit and Sit involves controlling the upper-trunk. In normal behavior, the head is supposed not to fall back or to the side, so we may see the atypically developing child to overshoot the 90 degrees head position and display a curved silhouette. Top-weighed, quality-related words like **greater** (than 90°), **overshoots** combined with body-related words **upper_trunk** and **upper_thoracic_spine** express this difference.

## 5.5 Related work

Our work belongs in an emerging field within NLP that is the application of computational linguistics to problems in clinical psychology, more notably the works of [99], [100], [101] and [102] who have shown that it is possible to discriminate between normal subjects and those affected with depression, post-traumatic stress disorder (PTSD) and other mental health signals. Different from these studies, we predict typicality/atypicality based on sentences and not individuals. However, as discussed earlier, the "true" movement described in these sentences carry the signal that can predict the disorder, so we are *indirectly* assessing subjects.

These studies rely on processing large volumes of social data "in the wild" using, among other things, features based on topic models or a LIWC dictionary.

| **Supine** | **Sidelying** | **Prone** | **Pull-to-sit** |
|---|---|---|---|
| TFIDF, uni + big, stem, F=0.68 | log1p uni, lemma, F=0.72 | log1p, uni+big, lemma, F=0.72 | TF, uni + big, F=0.77 |
| extend 1.06 (mov) | away 0.20 | down 0.13 | narrower 0.47 (qual) |
| briefli 0.98 (qual) | whole 0.14 | finger 0.13 (body) | individual 0.41 |
| asymmetri 0.97 (mov) | variety 0.13 | how 0.12 | greater 0.36 (qual, sent) |
| poor 0.94 (sent) | due 0.13 | upper_thoracic_spine 0.11 (body) | attempts 0.34 (state) |
| overshoot 0.94 (qual, sent) | mobilize 0.12 (state) | variety 0.10 | overshoots 0.31 (qual, sent) |
| come 0.89 | calm 0.12 (state) | atypical 0.10 (qual, sent) | vision 0.30 (task) |
| top 0.86 | get 0.12 | entire 0.10 | looking 0.28 (task) |
| abl_sustain 0.85 (state) | work 0.11 | round 0.09 (state) | course 0.28 |
| keep 0.83 | posture 0.11 | extend 0.09 (mov) | readily 0.27 (qual) |
| horizont 0.82 (task) | keep 0.10 | brushing 0.09 (mov) | challenge 0.27 (state) |
| ten 0.81 | briefly 0.10 (qual) | turn 0.09 (mov) | and 0.27 |
| appear 0.80 | create 0.10 | when 0.08 | rattle 0.27 (task) |
| carri 0.80 | sustain 0.09 (state) | strength 0.08 (state) | reciprocal 0.26 (mov) |
| upright_posit 0.80 (qual) | horizontal 0.09 (task) | quickly 0.08 (cal) | two 0.26 |
| saw 0.80 | readily 0.09 (state) | lifting 0.08 (mov) | spinal 0.25 (body) |
| freedom 0.73 (state) | unlikely 0.09 (sent) | core 0.07 (body) | handling 0.24 |
| hip 0.72 (body) | strategy 0.09 | more 0.07 (sent) | saw 0.24 |
| unbalanc 0.71 (qual) | lot 0.09 | now 0.07 | brushing 0.24 (mov) |
| augment 0.71 (sent) | also 0.09 | femoral 0.07 (qual) | presented 0.24 |
| immedi 0.67 (qual) | hold 0.09 (state) | balance 0.07 (qual) | sustains 0.23 (state) |

**Table 5.2:** Bags-of words with the top-20 more important words along with weights assigned by SVM classifiers for selected tasks (one per column, along with the pre-processing strategy used). Terms that describe movement = *mov*, qualify movement = *qua* refer to body parts = *body*, qualify the physical and or mental state of the actor = *sta*, qualify the task itself = *task* and terms that somehow reflect the sentiment of the analyst towards the performance = *sent*. Terms considered uninformative or too general were grayed out.

However, despite the good topics and beyond chance-level prediction scores obtained on the CLPsych 2015 shared tweets (e. g. the system proposed in [103]) these results were not translated into more concrete insights in the understanding of distinctions between depression, PTSD and normal subjects. Because of the limited size of our data, we focused less on prediction scores, and more on examining the weights of linear SVMs learned for each task.

| **Supine** | **Sidelying** | **Prone** | **Pull-to-sit** |
|---|---|---|---|
| TFIDF, uni + big, stem, F=0.68 | log1p uni, lemma, F=0.72 | log1p, uni+big, lemma, F=0.72 | TF, uni + big, F=0.77 |
| movement -1.21 | infant -0.18 | over -0.16 | upper_trunk -0.40 (body) |
| follow -1.12 | femoral -0.14 (qual) | lumbar_spine -0.14 (body) | most -0.38 |
| over -0.95 | flex -0.13 (mov) | bang -0.13 (mov) | counter -0.31 (qual) |
| result -0.94 | newborn -0.10 | also -0.11 | many -0.30 |
| reaction -0.92 (state) | immediately -0.10 (cal) | take -0.09 | hard -0.30 |
| roll -0.92 (mov) | drive -0.09 (mov) | investigate -0.08 (state) | somewhat -0.28 |
| possess -0.89 | typically -0.09 (qual) | delay -0.08 (qual) | degrees -0.27 |
| minim -0.84 (sent) | always -0.08 | area -0.08 | productive -0.27 (sent) |
| import -0.81 (sent) | area -0.08 | body -0.08 | initially -0.25 |
| begin -0.81 | movement -0.08 | rolling -0.08 (mov) | holding -0.24 |
| handl -0.75 | symmetrically -0.08 (mov) | vision -0.08 (task) | remain -0.24 |
| core -0.74 (body) | utilize -0.07 | hand -0.07 (body) | olds -0.23 |
| fulli -0.73 (qual) | week -0.07 | low_extremity -0.07 (body) | left -0.23 |
| lower -0.72 (qual, sent) | pseudo -0.07 (qual) | then -0.07 | month -0.22 |
| howev -0.68 | age -0.07 | typical -0.07 (qual) | let -0.21 |
| antigrav -0.68 (task) | rather -0.07 (sent) | thoracic_spine -0.06 (body) | increasing -0.21 (sent) |
| toy -0.65 (task) | instead -0.07 (sent) | table -0.06 | keep -0.21 |
| weight -0.61 (task) | attain -0.07 (state) | sustain_posture -0.06 (state) | versa -0.21 |
| postur -0.60 (task) | choose -0.06 (state) | two -0.06 | bouts -0.21 (task) |
| thirti -0.60 (task) | humeral -0.06 (qual) | kick -0.06 (mov) | turned -0.21 (task) |

**Table 5.3:** Bags-of words with the top-20 more important words along with weights assigned by SVM classifiers for selected tasks (one per column, along with the pre-processing strategy used). Terms that describe movement = *mov*, qualify movement = *qua* refer to body parts = *body*, qualify the physical and or mental state of the actor = *sta*, qualify the task itself = *task* and terms that somehow reflect the sentiment of the analyst towards the performance = *sent*. Terms considered uninformative or too general were grayed out.

| Feature | Supine | Sidelying | Prone | Pull-to-sit | Sit | Hor. susp. | Prot. ext. | Standing |
|---|---|---|---|---|---|---|---|---|
| TF | 0.5581 | 0.4545 | 0.5882 | **0.6000** | **0.6667** * | **0.7500** * | **0.7500** * | 0.6897 * |
| TFIDF | **0.6047** ** | 0.4348 | 0.5882 | **0.6000** | **0.6667** * | **0.7500** * | **0.7500** * | 0.6897 * |
| log1p | 0.5366 | **0.5833** | **0.6667** * | **0.6000** | 0.6400 ** | 0.7272 ** | 0.7059 * | **0.7200** * |

**Table 5.4:** F-measures of each method per task using linear SVM. We ran permutation tests with N=500 so that (*) $p < 0.05$ and (**) $p < 0.1$.

## 5.6 Conclusions and next steps

The absolute quantitative performance of task-based SVMs together with the observed top-scoring typical and atypical bags-of-words suggest that language descrip-

tions of these tasks should not be taken as a mere poor man's representation, but the opposite: we have reason to believe that text features can provide extra discriminative power by incorporating information distributed over a number of latent variables that qualify task behavior in dimensions that are at best only indirectly related to the original low-level movement signal. Our next discrimination model will thus account for these variables explicitly. Precision-recall numbers support a significant statistical relationship between this "movement language" and the typical/atypical labels, since they that discarded independence in 6 out of the 8 tasks, despite the small size of the data (N=500 permutations, $p < 0.05, 0.1$). Future work would involve comparing and combining/comparing language with movement features deriving from inter-segmental angles like the ones shown on Fig. 5.1.

# Chapter 6: Conclusions and future directions

## 6.1 A computational sketch of action generation based on SB-ST

In Chapter 2 we saw that SB-ST and three other methods suggested that, considering intersegmental joint angles of the legs, a single spatial basis should be compressing the postural space. We saw that, despite the large age range of our jumpers, that both the coefficients and amount of variance explained by that SB was consistent between subjects. We could then hypothesize that, in the absence of other dofs being considered, that the process of generating and parsing a jump would take place like what is shown on Fig. 6.1: the supplemental motor area (SMA) will issue a motor plan for the jump that specifies SB-1 as the dominant primitive, the family of basis functions $\mathbf{\Phi}_{\boldsymbol{\tau_i}}$ that define the pattern of spatio-temporal activation and and control and coordination parameters that will tune those functions, representing when in time that primitive will be recruited (coordination parameters, $\tau_{1,1}$ and $\tau_{1,2}$ in Fig. 6.1) and how strong will be the activation at that instant (control parameters, $c_{1,1}$ and $c_{1,2}$). When the jump is *generated* these primitives, functions and parameters (compressed motor information) would be communicated from the primary motor cortex to the spinal cord, whereas when the jump is *perceived*, under a simulation hypothesis, execution would be suppressed, so that the role of the mo-

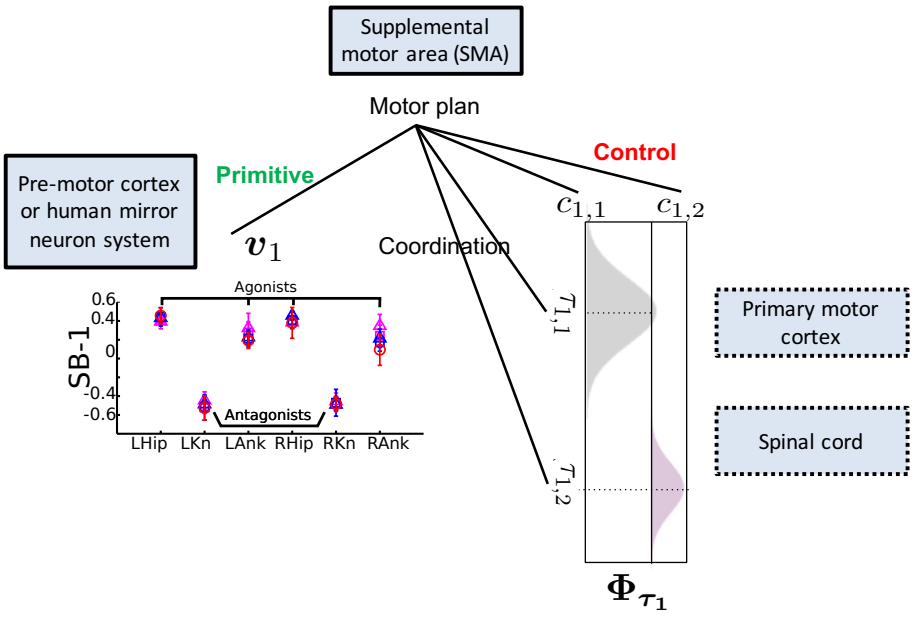



**Figure 6.1:** Generating and parsing jumps under the SB-ST model.

tor cortex would resume to providing higher-level areas with the compressed motor representations they need to retrieve information from the action and actor. According to what we have observed from the spatio-temporal parameters, differences between subjects would be observed in control parameters, which could suggest that humans would be equipped with very similar primitives (or motor programs) and families of basis functions, and problems would take place somewhere during planning. But the previous was an exploratory exercise, with a single action and a small number of dofs, and no statistical inference was carried out on the parameters, so these outcomes have to be taken more as insights than scientific results. Moreover, without reconciling behavioral data with brain data, it is very hard to make new assumptions about the neural basis of jumping or another sensorimotor task for that matter.

## 6.2 The encoding of groups and tasks in the spatial bases

We then turn to a different but related question: how are the spatial bases encoding information needed to recognize populations (here we will use the term groups) and tasks? To address this question, we looked at a a subset of the same motion capture data we collected in the sessions described in Chapter 2 except that this time we used tasks involving bimanual coordination and object manipulation rather than jumps. We chose this particular set both because it has large task variability and because it included most of the participants, maximizing statistical power. The tasks were (1) clapping, (2, 3) bouncing a ball with each arm, (4) catching and throwing a ball to a person and (5) pretending to scoop beans from one jar to another. Our participants were a member of one of the following: TD (typically developing) DCD (Developmental Coordination Disorder) YAD (young adults) SAD (senior adults) and PD (Parkinson's disease seniors). In total, we had data for 53 distinct participants in this study, and all subjects performed all 5 tasks, so subjects and tasks were equally represented. The number of subjects per group, however varied: 16 TD, 6 DCD, 11 YAD, 14 SAD and 6 PD. The only normalization applied to the raw data was the swapping of left and right arms for subjects that used their left arms to scoop, because they were fewer, the rest was set just like in the jump experiment. The dofs used were: left and right shoulders, elbows and wrists at $x$, $y$ and $z$ rotations, that is a total of 18 dofs.

We tested each individual dof of all spatial bases (more specifically, the absolute value of the coefficients at each dof within the corresponding singular vectors)

with a 2-way ANOVA withx factors *group* and *task*. For each SB-$i$ (singular vector $\boldsymbol{v}_i$) we split dofs into those with significant interactions ($\boldsymbol{v}_i^{int}$) and those without it. The latter were further subdivided into dofs with significant main effects for groups ($\boldsymbol{v}_i^{groups}$) tasks ($\boldsymbol{v}_i^{tasks}$) and no significant main effects ($\boldsymbol{v}_i^{nme}$). For the sake of simplicity, we may call the former two *group discriminating* and *task discriminating* dofs, respectively, or simply group and task dofs.

In practice, the decomposition is ran by zeroing out all dofs within SB-$i$ that do not belong to each subdivision. This creates a hierarchical subdecomposition of each SB-$i$ where each leaf node consists of an exclusive subset of dofs, so all leaves are orthogonal. The amount of discriminative information per leaf can be assessed by the ratio of the number of dofs at which that leaf is statistically significant over the maximum number of dofs ($p < 0.05$, Bonferroni corrected). For example, the hierarchy for for SB-1 will look like Fig. 6.2a, and the ratio of task discriminant dofs would be $(12 + 3)/18 = 83.33\%$. Because all leaves are orthogonal, this implies that the explained variance within SB-$i$ is also partitioned per subdecomposition, with the fractions defined by replacing the singular vector corresponding to SB-$i$ with the equivalent sum of orthogonal vectors in the right factor of the rank-1 expansion of

SVD, that is:

$$
\begin{aligned}
s_i \boldsymbol{u}_i &= \mathbf{Y} \boldsymbol{v}_i && \text{level 0 (root)} \\
s_i \boldsymbol{u}_i &= \mathbf{Y}[\boldsymbol{v}_i^{int} + \boldsymbol{v}_i^{nint}] && \text{level 1} \\
s_i \boldsymbol{u}_i &= \mathbf{Y}[\boldsymbol{v}_i^{int} + \boldsymbol{v}_i^{me} + \boldsymbol{v}_i^{nme}] && \text{level 2} \\
s_i \boldsymbol{u}_i &= \mathbf{Y}[\boldsymbol{v}_i^{int} + \boldsymbol{v}_i^{tasks} + \boldsymbol{v}_i^{groups} + \boldsymbol{v}_i^{nme}] \quad \text{or} && \text{level 3} \\
s_i &= \frac{\boldsymbol{u}_i^\top \mathbf{Y}}{s_i}[\boldsymbol{v}_i^{int} + \boldsymbol{v}_i^{tasks} + \boldsymbol{v}_i^{groups} + \boldsymbol{v}_i^{nme}],
\end{aligned}
\tag{6.1}
$$

and from Equation 6.1: $\boldsymbol{u}_i = \frac{\mathbf{Y}\boldsymbol{v}_i}{s_i}$, so:

$$
s_i = \frac{(\mathbf{Y}\boldsymbol{v}_i)^\top \mathbf{Y}}{s_i}[\boldsymbol{v}_i^{int} + \boldsymbol{v}_i^{tasks} + \boldsymbol{v}_i^{groups} + \boldsymbol{v}_i^{nme}],
$$

with the fractions per term being obtained in terms of the data matrix, spatial bases and explained variances by dividing the both sides by $s_i$:

$$
1 = \frac{(\mathbf{Y}\boldsymbol{v}_i)^\top \mathbf{Y}}{s_i^2}[\boldsymbol{v}_i^{int} + \boldsymbol{v}_i^{tasks} + \boldsymbol{v}_i^{groups} + \boldsymbol{v}_i^{nme}].
$$

In other words, the fraction of SB-$i$ variance explained by one of its leaf subdecompositions $\boldsymbol{v}_i^{leaf}$ is:

$$
s^{\%}_{i,leaf} = \frac{\boldsymbol{v}_i^\top \mathbf{Y}^\top \mathbf{Y} \boldsymbol{v}_i^{leaf}}{s_i^2} \tag{6.2}
$$

with $i = 1 \ldots k$ where $k \leq$ maximum number of spatial degrees of freedom (here 18) and $leaf \in \{int,\ nme,\ tasks,\ groups\}$. Because each SB-$i$ basis covers exclusive parts of the variance, and so does each leaf within that basis, integrating Equa-

tion 6.2 over all leaves and bases will result in the entire variance in the decomposed data:

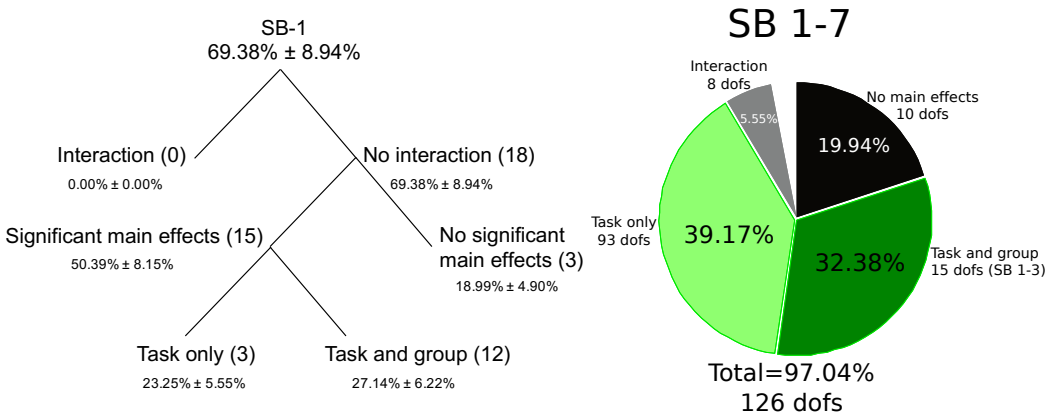
$$\sum_{i}\sum_{leaf} s^{\%}_{i,leaf} = 1. \tag{6.3}$$

The amount of explained variance per leaf expresses the how much data reconstruction power is accounted for by the different leaves. Table 6.1 shows the mean $\pm$ standard deviation explained variances of each SB-$i$ from 1 to 7 (we stopped at 7 because this is the last spatial basis with at least 1% of the average explained variance). Fig. 6.2b shows the total number of significant dofs and accumulated mean variances accumulated for the same spatial bases. These data tells us the following: first, although in comparison with jumps, the tasks involved in this experiment utilize more degrees of freedom and include more tasks and populations, the first spatial basis took care of an even higher average amount of variance, that is, 69.38% (Fig. 6.2a) suggesting that 3D biological motion can be very efficiently compressed into key postures. Second, and more importantly, all dofs with significant main effects for groups are also significant for tasks, so group discriminating dofs are actually group *and* task discriminating, or alternatively, task-discriminating dofs are *task-exclusive* discriminating dofs. This does not mean that group and task are necessarily interacting, and in fact, we found only a few dofs with significant interactions (8 out of 126). In practice, this would mean that engaging in cognitive tasks such as perceiving, imagining and performing both a certain action **and** that action but with the addition of the traits that characterizes a certain group of interest would involve the same cost in terms of representation, that is the cost relative

to the task alone.

Computationally, the fact that all group variance is embedded within the same discrimination manifold as the task variance means that as we recreate an action based on the spatial bases, we also recreate the actor's way of performing it. From what we see here, *recreating the actor would happen faster than the action* because the largest chunk of group discriminating dofs and explained variances are within SB-1 and SB-3 (see third and fourth columns of Table 6.1). Note as well that the number of group discriminative dofs actually goes down with SB-$i$ as $i$ increases, as opposed to what happens with task-exclusive ones (see seventh and eighth columns of Table 6.1). A compression scheme that allows processing actors faster than actions could be indicate an underlying bias to social information processing from motion signals over action recognition, but this would have to be investigated further. Anyhow, it is interesting to see that group discrimination information can afford such compression, especially considering the wide age range of individuals tested (from 6 to 80 years old, with and without neuromotor disorders).

Although the goal here is not to look at the particular dofs that discriminate these groups, we can get an idea of what type of differences these spatial bases seem to be expressing by looking at the results of pairwise post-hoc analyses; for example, the analysis in Table 6.2 reveals various subsets of numerous dofs that tell children from senior adults (row DCD and column TD) and adults from seniors (column YAD) but none that discriminate between the two children groups. Plus it shows a single set with two dofs that discriminate the two senior groups (PD, SAD). The boundary between young adults and children seems to be a bit blurry, since there are




**(a)** Hierarchical subdecomposition of SB-1 resulting from 2-way ANOVA. **(b)** Average explained variance of subdivisions in (a) accumulated over SB 1 to 7.

**Figure 6.2**

only two single-dof subsets separating this group from the children (column YAD). We can hence deduce that SB-1 is mostly representing *normal differences* between the different developmental stages, with TD and DCD forming a single cluster, YAD forming another cluster with large overlap with the former, and a third cluster with SAD and PD far from the rest. The same analysis but at SB-2 and SB-3 reveals a dof that discriminates TD from DCD and SAD and PD, respectively, so these two bases are more likely to be connected to *abnormal differences* between TD and DCD as well as SAD and PD, respectively. However, one has to be careful when looking at SB-2 and SB-3 group differences since they amount to a subset of small average explained variances (see third column of Table 6.2, second row: SB-2 = $4.77 \pm 5.82\%$ and third row: SB-3 = $0.47 \pm 0.61\%$).

| | $s^{\%}_{i,int}$ | $s^{\%}_{i,nme}$ | $s^{\%}_{i,groups}$ | $s^{\%}_{i,tasks}$ | $\mathrm{df}_{int}$ | $\mathrm{df}_{nme}$ | $\mathrm{df}_{groups}$ | $\mathrm{df}_{tasks}$ | $s_i(\%)$ |
|---|---|---|---|---|---|---|---|---|---|
| **SB-1** | | $18.99 \pm 4.90$ | $27.14 \pm 6.22$ | $23.25 \pm 5.55$ | 0 | 3 | 12 | 3 | $69.38 \pm 8.94$ |
| **SB-2** | $4.48 \pm 4.87$ | | $4.77 \pm 5.82$ | $3.96 \pm 2.66$ | 5 | 0 | 2 | 11 | $13.21 \pm 4.20$ |
| **SB-3** | $1.07 \pm 1.04$ | $0.46 \pm 0.56$ | $0.47 \pm 0.61$ | $4.39 \pm 1.95$ | 3 | 1 | 1 | 13 | $6.40 \pm 2.39$ |
| **SB-4** | | | | $3.46 \pm 1.51$ | 0 | 0 | 0 | 18 | $3.46 \pm 1.51$ |
| **SB-5** | | $0.16 \pm 0.23$ | | $2.01 \pm 0.93$ | 0 | 1 | 0 | 17 | $2.17 \pm 0.98$ |
| **SB-6** | | $0.17 \pm 0.19$ | | $1.25 \pm 0.59$ | 0 | 3 | 0 | 15 | $1.42 \pm 0.67$ |
| **SB-7** | | $0.15 \pm 0.14$ | | $0.85 \pm 0.47$ | 0 | 2 | 0 | 16 | $1.00 \pm 0.52$ |
| **Sum** | 5.55 | 19.94 | 32.38 | 39.17 | 8 | 10 | 15 | 93 | 97.04 |

**Table 6.1:** Number of dofs with statistically significant main effects within $\boldsymbol{v}_i^{\{int,groups,tasks\}}$ and dofs without significant main effects for groups or tasks $\boldsymbol{v}_i^{nme}$ and corresponding variances from SB-1 to SB-7 ($p < 0.05$ Bonferroni corrected). The last row accumulates the dofs or means on the previous ones, depending on the column. Empty cells mean no variance was explained by the corresponding subdecompositions.

| | **PD** | **SAD** | **TD** | **YAD** |
|---|---|---|---|---|
| **DCD** | 4 | 7 | | 1 |
| **PD** | | 2 | 5 | 6 |
| **SAD** | | | 7 | 6 |
| **TD** | | | | 1 |

**Table 6.2:** SB-1 post-hoc group analysis: number of group discriminating dofs per group pair (out of 12) ($p < 0.05$ Bonferroni corrected). Empty cells or missing group pairs mean no dofs were found to discriminate between the corresponding members.

| | **TD** | **YAD** |
|---|---|---|
| **DCD** | 1 | |
| **PD** | 2 | 1 |
| **SAD** | 1 | |

**Table 6.3:** SB-2 post-hoc group analysis: number of group discriminating dofs per group pair (out of 2) ($p < 0.05$ Bonferroni corrected). Empty cells or missing group pairs mean no dofs were found to discriminate between the corresponding members.

| | SAD | TD |
|---|---|---|
| **PD** | 1 | 1 |

**Table 6.4:** SB-3 post-hoc group analysis: number of group discriminating dofs per group pair (out of 1) ($p < 0.05$ Bonferroni corrected). Empty cells or missing group pairs mean no dofs were found to discriminate between the corresponding members.

## 6.3 Next steps

Assuming the focus is on the analysis of groups and not tasks, the natural next step to this research is to look at the statistics of SB-1, SB-2 and SB-3 temporal behavior to understand how the recruitment of postures vary between groups. We can do it combining the procedures in Chapter 2, Section 2.4.4 with what we did in the last section. In Section 2.4.4 we analyzed differences in the spatial-temporal profile of the single spatial basis (SB-1) which we judged would be the same for all groups, based on the statistics of Fig. 2.6 (top, left). In other words, we believed that groups (TD, YAD and DCD) were "registered" with respect to that spatial basis. As a consequence, the inputs to the spatial-temporal analysis in the step that followed were the full spatio-temporal profiles of SB-1 (i. e. ST-1) calculated as the projection $\boldsymbol{z}_1 = \boldsymbol{Y}\boldsymbol{v}_1$.

Here, the procedure would be slightly modified. Let us only consider SB-1 to simplify the description: from Table 6.1, we saw that only 6 out of 18 dofs with no statistically significant group differences (dofs for which the two-way ANOVA could not reject the null hypothesis) so these are the only dofs we are interested to consider further when analyzing spatio-temporal profiles. That said, we would then (1) zero out all 12 others thus creating a modified $\boldsymbol{v}_1^{null}$ whose non-zero elements

would register all groups with respect to SB-1 in the sense described earlier. Next (2) we would calculate ${}^{null}ST$-1 profiles by projecting $\boldsymbol{Y}$ onto the modified $\boldsymbol{v}_1^{null}$. To calculate the amount of explained variance corresponding to the spatial-temporal profile, we would use Equation 6.3 just like shown in the previous section.

With the spatio-temporal profiles properly calculated, and assuming we would stick with univariate analysis, two different steps could be carried out next: (1) testing every single time instant within the maximum length $T$ of ${}^{null}ST$-1 under the same two-way ANOVA paradigm as before. The problem with this approach is that it might end up underpowered, as a result of correcting for multiple comparisons. Alternatively, (2) we would do it as in the SB-ST algorithm/jump experiment, where we ran VARPRO and fit a small number $N << T$ of Gaussians to spatio-temporal profiles, and inference would take place only at the $2 \cdot N$ resulting control and coordination parameters, thus avoiding the large multiple comparisons problem. To choose a suitable $N$ we could look for an "elbow" in the plot $R^2$ versus $N$ like in Fig. 2.4 (top).

This type of temporal analysis can help spotting disparities in how groups recruit subsets of common factors in time (candidates for action primitives) and how much these common factors contribute to task performance. These disparities can be related of abnormalities and provide a better understanding of conditions such as DCD or PD. However, this is just the behavioral side of the question; it would be interesting to see if there are neural correlates to these differences. Based on the principle of direct matching (Chapter 1) it could be the case that the action observation network in a normal person's human mirror neural system [104]

would react differently to experiencing typical and atypical performances of a trial. A richer experiment would be to subject the typical and atypical populations to normal and abnormal trials and run a similar study. In any case, using action observation networks to engage the motor system during observation may be a clever approach to study behavior that would not otherwise be possible because of technology limitations (e. g. it is impossible for participants to bounce a ball in an fMRI scanner) or to help in rehabilitation of subjects with limited mobility, like certain stroke patients [105]. There is also an ongoing effort to figure out how to use state-of-the-art knowledge on neural mirror systems to understand the relationship between the integrity of action primitives and its relation with movement disorders, sensorimotor injuries [106] and control impairments [107].

# Bibliography


[1] Tatsuya Harada, Akihiko Saito, Tomomasa Sato, and Taketoshi Mori. Infant behavior recognition system based on pressure distribution image. *International Conference on Robotics and Automation*, pages 4082–4088, 2000.

[2] Thiago Vallin Spina, Mariano Tepper, Amy Esler, Vassilios Morellas, Nikolaos Papanikolopoulos, Alexandre Xavier Falcao, and Guillermo Sapiro. Video human segmentation using fuzzy object models and its applications to body pose estimation of toddlers for behavior studies. Unpublished manuscript. http://arxiv.org/abs/1305.6918v1, May 2013.

[3] Michael T. Turvey. Coordination. *Am. Psychol.*, 45(8):938–953, 1990.

[4] Mark L. Latash, John P. Scholz, and Gregor Schöner. Toward a new theory of motor synergies. *Motor Control*, 11(3):276–308, 2007.

[5] Mark L. Latash, Mindy F. Levin, John P. Scholz, and Gregor Schöner. Motor control theories and their applications. *Medicina (Kaunas)*, 46(6):382–392, 2010.

[6] Tamar Flash and Binyamin Hochner. Motor primitives in vertebrates and invertebrates. *Curr. Opin. Neurobiol.*, 15(6):660–666, 2005.

[7] Ferdinando A. Mussa-Ivaldi, Simon F. Giszter, and Emilio Bizzi. Linear combination of primitives in vertebrate motor control. *PNAS*, 91:7534–7538, 1994.

[8] Ferdinando A. Mussa-Ivaldi. Nonlinear force fields: a distributed system of control primitives for representing and learning movements. In *Proc. of CIRA*, pages 84–90, 1997.

[9] Ferdinando A. Mussa-Ivaldi and Emilio Bizzi. Motor learning through the combination of primitives. *Phil. Trans. R. Soc. B*, 355:1755–1769, 2000.

[10] Andrea d'Avella and Matthew C. Tresch. Modularity in the motor system: decomposition of muscle patterns as combination of time-varying synergies. *Proc. of NIPS*, 14:141–148, 2001.

[11] Andrea d'Avella, Philippe Saltiel, and Emilio Bizzi. Combinations of muscle synergies in the construction of a natural motor behavior. *Nat. Neurosci.*, 6(3):300–308, 2003.

[12] Andrea d'Avella, Philippe Saltiel, and Emilio Bizzi. Shared and specific muscle synergies in natural motor behaviors. *PNAS*, 102(8):3076–3081, 2005.

[13] Marco Santello, Martha Flanders, and John F. Soechting. Postural hand synergies for tool use. *J. Neurosci.*, 18(23):10105–10115, 1998.

[14] Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.*, 14:201–211, 1973.

[15] Gunnar Johansson. Spatio-temporal differentiation and integration in visual motion perception. an experimental and theoretical analysis of calculus-like functions in visual data processing. *Psychol. Res.*, 38(4):379–393, 1976.

[16] James E. Cutting. Coding theory adapted to gait perception. *Journal of Experimental Psychology: Human Perception & Performance*, 7:71–87, 1981.

[17] James E. Cutting and Dennis R. Proffitt. Gait perception as an example of how we may perceive events. *Intersensory Perception and Sensory Integration*, pages 249–273, 1981.

[18] Winand H. Dittrich. Action categories and the perception of biological motion. *Perception*, 22(1):15–22, 1993.

[19] Bennett I. Berthenthal and Jeannine Pinto. Global processing of biological motions. *Psychological Science*, 5(4):221–225, 1994.

[20] Nikolaus F. Troje. Decomposing biological motion: a framework for analysis and synthesis of human gait patterns. *J Vis*, 2:371–387, 2002.

[21] Marc Jeannerod. *Motor cognition: what actions tell to the self*, chapter 6: The simulation hypothesis of motor cognition. Oxford University Press, USA, 2006.

[22] Marc Jeannerod. *Motor cognition: what actions tell to the self*, chapter 2: Imagined actions as a prototypical form of action representation. Oxford University Press, USA, 2006.

[23] Marc Jeannerod, Michael A. Arbib, Giacomo G. Rizzolatti, and H. Sakata. Grasping objects: the cortical mechanisms of visuomotor transformation. *Trends in Neuroscience*, 18(7):314–320, 1995.

[24] Melvyn A. Goodale and A. David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

[25] Melvyn A. Goodale and David A. Westwood. An evolving view of duplex vision: separate but interacting cortical pathways. *Current Opinion in Neurobiology*, pages 203–211, 2004.

[26] Andrew H. Fagg and Michael A. Arbib. Modeling parietal-premotor interactions in primate control of grasping. *Neural Networks*, 11(1277-1303), 1998.

[27] Jeannette Bohg and Danica Kragic. Grasping familiar objects using shape context. *International Conference on Advanced Robotics*, pages 1–6, 2009.

[28] Ashutosh Saxena, Justin Driemeyer, Justin Kearns, and Andrew Y. Ng. Robotic grasping of novel objects. *Advancements in Neural Information Processing Systems*, 19, 2006.

[29] Abhinav Gupta and Larry S. Davis. Objects in action: an approach for combining action understanding and object perception. *International Conference on Computer Vision*, pages 1–8, 2007.

[30] Hedvig Kjellström, Javier Romero, David Martínez, and Danica Kragić. Simultaneous visual recognition of manipulation actions and manipulated objects. *European Conference on Computer Vision*, 2:336–349, 2008.

[31] Roman Filipovych and Eraldo Ribeiro. Recognizing primitive interactions by exploring actor-object states. *Computer Vision and Pattern Recognition*, pages 1–7, 2008.

[32] J. Randall Flanagan and Roland S. Johansson. Action plans used in action observation. *Nature*, 424:769–771, 2003.

[33] Marc Jeannerod. *Motor cognition: what actions tell to the self*, chapter 5: How do we perceive and understand the action of others. Oxford University Press, USA, 2006.

[34] Yuri P. Ivanenko, Richard E. Poppele, and Francesco Lacquaniti. Motor control programs and walking. *Neuroscientist*, 12(4):339–348, 2006.

[35] Neil D. Lawrence. Learning for larger datasets with the gaussian process latent variable model. *Proc. of AISTATS*, pages 243–250, 2007.

[36] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. Gaussian process dynamical models. *Proc. of NIPS*, pages 1141–1148, 2005.

[37] A. Bhat, R. J. Landa, and J. C. Galloway. Current perspectives on motor functioning in infants, children, and adults with autism spectrum disorders. *Physical Therapy*, 91(7):1116–1129, 2011.

[38] Gene Golub and Victor Pereyra. The differentiation of pseudo-inverses and non-linear least squares problems whose variables separate. *SIAM J. Numer. Anal.*, 10(2):413–432, 1973.

[39] Dianne P. O'Leary and Bert W. Rust. Variable projection for nonlinear least squares problems. *Comput. Optim. Appl.*, 54(3):579, 593 2013.

[40] Matthew C. Tresch, Vincent C. Cheung, and Andrea d'Avella. Matrix factorization algorithms for the identification of muscle synergies: evaluation on simulated and experimental data sets. *J. Neurosci.*, 95(4):2199–2112, 2006.

[41] Yuri P. Ivanenko, Germana Capellini, Nadia Dominici, Richard E. Poppele, and Francesco Lacquaniti. Coordination of locomotion with voluntary movements. *J. Neurosci.*, 25(31):7238–7253, 2005.

[42] John P. Scholz and Gregor Schöner. The uncontrolled manifold concept: identifying control variables for a functional task. *Exp. Brain Res.*, 126(3):289–306, 1999.

[43] Auke J. Ijspeert, Jun Nakanishi, Heiko Hoffmann, Peter Pastor, and Stefan Schaal. Dynamical movement primitives: learning attractor models for motor behaviors. *Neural Computation*, (25):328–373, 2013.

[44] Neil D. Lawrence. Gaussian process models for the visualization of high dimensional data. *Proc. of NIPS.*, 16:329–336. Source code: GPLVM toolbox available at http://ml.sheffield.ac.uk/~neil/gplvm/, 2004.

[45] Raquel Urtasun, David Fleet, and Pascal Fua. 3D people tracking with Gaussian process dynamical models. *Proc. of CVPR*, 2006.

[46] Angela Yao, Juergen Gall, Luc V. Gool, and Raquel Urtasun. Learning probabilistic non-linear latent variable models for tracking complex activities. *Proc. of NIPS*, 2011.

[47] Guoliang Fan, Xin Zing, and Meng Ding. Gaussian process for human motion modeling: a comparative study. *Proc. of MLSP*, pages 1–6, 2011.

[48] Jody L. Jensen, Sally J. Phillips, and Jane E. Clark. For young jumpers, differences are in the movement's control, not its coordination. *Res. Q. Exerc. Sport*, 65:258–268, 1994.

[49] Sheila E. Henderson and David A. Sugden. *Movement assessment battery for children*. The Psychological Corporation, London, England, 1992.

[50] Marc Jeannerod. The timing of natural prehension movements. *Journal of Motor Behavior*, 16:235–254, 1984.

[51] Alex Pentland. *Honest Signals: How They Shape Our World*. MIT Press, 2008.

[52] Tetsunari Inamura, Iwaki Toshima, and Hiroaki Tanie. Embodied symbol emergence based on mimesis theory. *International Journal of Robotics Research*, 23(4-5):363–377, 2004.

[53] Justine Cassel. A framework for gesture generation and interpretation. *Computer vision for human-machine interaction*, pages 191–216, 1998.

[54] Justine Cassell, Stefan Kopp, Paul A. Tepper, Kim Ferriman, and Kristina Striegnitz. Trading spaces: how humans and humanoids use speech and gesture to give directions. *Engineering approaches to conversational informatics*, pages 133–160, 2007.

[55] David McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, 1992.

[56] Eren Erdal Aksoy, Alexey Abramov, Johannes Dörr, Kejun Ning, Babette Dellen, and Florentin Wörgötter. Learning the semantics of object-action relations by observation. *International Journal of Robotics Research*, 30:1229–1249, 2011.

[57] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders (4th ed.)*. Washington, DC, 2000.

[58] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders (5th ed.)*. Washington, DC, 2013.

[59] Peter Hobson. *The cradle of thought: exploring the origins of thinking*. Macmillan, Oxford, 2002.

[60] M. Mari, U. Castiello, D. Marks, C. Marraffa, and M. Prior. The reach-to-grasp movement in children with autism spectrum disorder. *Phil. Trans. Royal Society of London*, (358):393–403, 2003.

[61] F. Shic, K. Chawarska, J. Bradshaw, and B. Scassellati. Autism, eye-tracking, entropy. *International Conference on Development and Learning*, pages 73–78, 2008.

[62] Victoria L. Chester and Matthew Calhoun. Gait symmetry in children with autism. *Autism Research and Treatment*, 2012, 2012.

[63] Eileen Chai, Jillian Chavis, Kevin Chodnicki, Tim Crisci, Nathan Destler, Duncan Graham, Kesshi Jordan, Richard Landa, Conrad Merkle, Soh Park, Christopher Paxton, Rachita Sood, James Tanner, and Brendan Wray. Assessing the viability of studying motion indicators of autism spectrum disorders in infants at high and low risk for asd using a passive motion capture system. Technical report, University of Maryland, http://hdl.handle.net/1903/12484, 2012.

[64] Gili Weinberg, Rich Fletcher, and Seum-Lim Gam. The BabySense environment: enriching and monitoring infants' experiences and communication. In *Conference on Human Factors in Computing Systems*, pages 18–23, 1998.

[65] Sabri Boughorbel, Fons Bruekers, and Jeroen Breebaart. Baby-posture classification from pressure-sensor data. *International Conference on Pattern Recognition*, pages 556–559, 2010.

[66] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. *Computer Vision and Pattern Recognition*, pages 1297–1304, 2011.

[67] Nicolaos B. Karayiannis, Seshadri Srinivasan, Rishi Bhattacharya, Merrill S. Wise, Jr. James D. Frost, and Eli M. Mizrahi. Extraction of motion strength and motor activity signals from video recordings of neonatal seizures. *Transactions on Medical Imaging*, 20:965–980, 2001.

[68] Nicolaos B. Karayiannis, Abdul Sami, James D. Frost Jr., Merrill S. Wise, and Eli M. Mizrahi. Quantifying motion in video recordings of neonatal seizures by feature trackers based on predictive block matching. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, pages 1447–1450, 2004.

[69] Nicolaos Karayiannis, Yaohua Xiang, James D. Frost Jr., Merrill S. Wise, and Eli M. Mizrahi. Quantifying motion in video recordings of neonatal seizures by robust motion trackers based on block motion models. *Transactions on Biomedical Engineering*, 52(6):1065–1077, 2005.

[70] Nicolaos Karayiannis, Guozhi Tao, Yaohua Xiang, Abdul Sami, Bindu Varughese, James D. Frost Jr., Merrill S. Wise, and Eli M. Mizrahi. Computerized motion analysis of videotaped neonatal seizures of epileptic origin. *Epilepsia*, 6(901-917), 46.

[71] Nicolaos B. Karayiannis, Bindu Varughese, Jr. James D. Frost, Merrill S. Wise, and Eli M. Mizrahi. Quantifying motion in video recordings of neonatal seizures by regularized optical flow methods. *Transactions on Image Processing*, 2005.

[72] Nicolaos B. Karayiannis, Abdul Sami, James D. Frost, Jr., Merrill S. Wise, and Eli M. Mizrahi. Automated extraction of temporal motor activity signals from video recordings of neonatal seizures based on adaptive block matching. *Transactions on Biomedical Engineering*, 52(4):676–686, 2005.

[73] Nicolaos B. Karayiannis, Yaohua Xiong, Jr. James D. Frost, Merrill S. Wise, and Eli M. Mizrahi. Improving the accuracy and reliability of motion tracking methods used for extracting temporal motor activity signals from video recordings of neonatal seizures. *Transactions on Biomedical Engineering*, 52(4):747–749, 2005.

[74] Debi P. Dogra, Arun K. Majumdar, Shamik Sural, Jayanta Mukherjee, Suchandra Mukherjee, and Arun Singh. Toward automating Hammersmith

pulled-to-sit examination of infants using feature point based video object tracking. *Transactions On Neural Systems and Rehabilitation Engineering*, 20(1):38–47, 2012.

[75] Jordan Hashemi, Thiago Vallin Spina, Mariano Tepper, Amy Esler, Vassilios Morellas, Nikolaos Papanikolopoulos, and Guillermo Sapiro. A computer vision approach for the assessment of autism-related behavioral markers. *International Conference on Development and Learning and Epigenetic Robotics*, pages 1–7, 2012.

[76] Jordan Hashemi, Thiago Vallin Spina, Mariano Tepper, Amy Esler, Vassilios Morellas, Nikolaos Papanikolopoulos, and Guillermo Sapiro. Computer vision tools for the non-invasive assessment of autism-related behavior. http://arxiv.org/abs/1210.7014v2, November 2012.

[77] University of Miami. New knowledge: Of machines and men.

[78] Centers for Disease Control and Prevention. Prevalence of Autism Spectrum Disorders Autism and Developmental Disabilities Monitoring Network, 14 sites, United States, 2008. *Morbidity and Mortality Weekly Report*, 61(3), 2012.

[79] S. A. Naimer, H. B. Alonim, D. Tayar, and B. Schipper. A model of autism treatment: Using early intensive and sequential multidisciplinary intervention. *The Israel Journal of Family Practice*, 16(132):56–61, 2006.

[80] J. W. Jacobson, J. A. Mulick, and G. Green. Cost-benefit estimates for early intensive behavioral intervention for young children with autism - general model and single state case. *Behavioral Interventions*, 13(4):201–226, 1998.

[81] Osnat Teitelbaum, Tom Benton, Prithvi K. Shah, Andrea Prince, Joseph L. Kelly, and Philip Teitelbaum. Eshkol-Wachman movement notation in diagnosis: the early detection of asperger's syndrome. *Proceedings of the National Academy of Sciences*, 101(32):11909–11914, 2004.

[82] D. Alie, M. H. Mahoor, W. I. Mattson, and D. R. Anderson. Analysis of eye gaze pattern of infants at risk of autism spectrum disorder using markov models. *Workshop on Applications of Computer Vision*, pages 282–287, 2011.

[83] Gianluca Esposito, Paola Venuti, Fabio Apicella, and Filippo Muratori. Analysis of unsupported gait in toddlers with autism. *Brain and Development*, 33:367–373, 2011.

[84] Marcel Zentner and Tuomas Eerola. Rhythmic engagement with music in infancy. *Proceedings of the National Academy of Sciences*, 107(13):5768–5773, 2010.

[85] Renee L. Carrico Neil E. Berthier. Visual information and object size in infant reaching. *Infant Behavior and Development*, 33(4):555–566, 2010.

[86] J. Artigas, W. Mattson, D. Messinger, P. Ruvolo, T. Wu, and J. Movellan. Rethinking motor development and learning, 2011.

[87] The Karolinksa Institutet and the Uppsala University. EASE - Early Autism Sweden. 2013.

[88] Lonnie Zwaigenbaum, Susan Bryson, Tracey Rogers, Wendy Roberts, Jessica Brian, and Peter Szatmari. Behavioral manifestations of autism in the first year of life. *International Journal of Developmental Neuroscience*, 23(2-3):143–152, 2005.

[89] Jigna Bhatt, Niels da Vitoria Lobo, Mubarak Shah, and George Bebis. Automatic recognition of a baby gesture. *International Conference on Tools with Artificial Intelligence*, pages 610–615, 2003.

[90] Abdul Sami, Nicolaos B. Karayiannis, Jr. James D. Frost, Merrill S. Wise, and Eli M. Mizrahi. Automated tracking of multiple body parts in video recordings of neonatal seizures. *International Symposium on Biomedical Imaging: Nano to Macro*, pages 312–315, 2004.

[91] G. Ferrari, G. M. Kouamou, C. Copioli, R. Raheli, and F. Pisani. Low-complexity image processing for real-time detection of neonatal clonic seizures. In *International Symposium on Applied Sciences in Biomedical and Communication Technologies*, pages 1–5, 2010.

[92] Esther Tellen. Self-organization in developmental processes: can systems approaches work? *Systems and Development: The Minnesota Symposium in Child Psychology*, 22:77–117, 1989.

[93] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(24):509–521, 2002.

[94] João V. B. Soares and Andrea Baraldi. Operational estimation of a comprehensive set of complementary shape, size, and photometric attributes of image-objects. Unpublished manuscript. Available at http://www.umiacs.umd.edu/~joao/shapeIndexesPaper/08Aug2014/paper.pdf.

[95] Philip Teitelbaum, Osnat Teitelbaum, Jennifer Nye, Joshua Fryman, and Ralph G. Maurer. Movement analysis in infancy may be useful for early diagnosis of autism. *PNAS*, 95(23):13982–13987, 1998.

[96] Leonardo Claudino and Yiannis Aloimonos. Studying human behavior from infancy: on the acquisition of infant postural data. In *International Conference on Development and Learning and on Epigenetic Robotics*, pages 256–261, 2014.

[97] Shimon Kogan, Dimitry Levin, R. Bryan Routledge, S. Jacob Sagi, and A. Noah Smith. Predicting risk from financial reports with regression. In *ACL HLT*, pages 272–280, 2009.

[98] Pathways.org. Educational videos http://pathways.org/watch/motor/. 2015.

[99] Philip Resnik, Anderson Garron, and Rebecca Resnik. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1348–1353, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[100] Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. Beyond lda: Exploring supervised topic modeling for depression-related language in twitter. *2nd Workshop on Computational Linguistics and Clinical Psychology (CLPsych) in conjunction NAACL HLT, 2015*, 2015.

[101] Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[102] H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. Towards assessing changes in degree of depression through facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[103] Philip Resnik, William Armstrong, Leonardo Claudino, and Thang Nguyen. The university of maryland clpsych 2015 shared task system. *2nd Workshop on Computational Linguistics and Clinical Psychology (CLPsych) in conjunction NAACL HLT, 2015*, 2015.

[104] Marco Iacoboni and Mirella Dapretto. The mirror neuron system and the consequences of its dysfunction. *Nature Reviews Neuroscience*, 7:942–951, 2007.

[105] Kathleen A. Garrison, Lisa Aziz-Zadeh, Savio W. Wong, Sook-Lei Liew, and Carolee J. Winstein. Modulating the motor system by action observation after stroke. *Stroke*, 44:2247–2253, 2013.

[106] Marco Santello and Catherine E. Lang. Are movement disorders and sensimotor injuries pathologic synergies? when normal multi-joint movement synergies become pathologic. *Frontiers in Human Neuroscience*, 8(1050):1–13, 2015.

[107] Yuri P. Ivanenko, Germana Capellini, Marco Molinari, and Francesco Lacquaniti. *Introduction to Neural Engineering for Motor Rehabilitation*, chapter Motor control modules of human movement in health and disease, pages 39–60. John Wiley and Sons, Inc., 2013.