Neurocomputational Principles of Action Understanding: Perceptual Inference, Predictive Coding, and Embodied Simulation

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Tobias Fabian Schrodt
aus Kassel

Tübingen
2018

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	11.05.2018
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Martin Butz
2. Berichterstatter:	Prof. Dr. Heiko Neumann

Abstract

The social alignment of the human mind is omnipresent in our everyday life and culture. Yet, what mechanisms of the brain allow humans to be social, and how do they work and interact? Despite the apparent importance of this question, the nexus of cognitive processes underlying social intelligence is still largely unknown. A system of mirror neurons has been under deep, interdisciplinary consideration over recent years, and farreaching contributions to social cognition have been suggested, including understanding others' actions, intentions, and emotions. Theories of embodied cognition emphasize that our minds develop by processing and inferring structures given the encountered bodily experiences. It has been suggested that also action understanding is possible by simulating others' actions by means of the own embodied representations. Nonetheless, it remains largely unknown how the brain manages to map visually perceived biological motion of others onto principally embodied states like intentions and motor representations, and which processes foster suitable simulations thereof. Seeing that our minds are generative and predictive in nature, and that cognition is elementally anticipatory, also principles of predictive coding have been suggested to be involved in action understanding. This thesis puts forward a unifying hypothesis of embodied simulation, predictive coding, and perceptual inferences, and supports it with a neural network model. The model (i) learns encodings of embodied, self-centered visual and proprioceptive, modal and submodal perceptions as well as kinematic intentions in separate modules, (ii) learns temporal, recurrent predictions inside and across these modules to foster distributed and consistent simulations of unobservable embodied states, (iii) and applies top-down expectations to drive perceptual inferences and imagery processes that establish the correspondence between action observations and the unfolding, simulated self-representations. All components of the network are evaluated separately and in complete scenarios on motion capture data of human subjects. In the results, I show that the model becomes capable of simulating and reenacting observed actions based on its embodied experience, leading to action understanding in terms of motor preparations and inference of kinematic intentions. Furthermore, I show that perceptual inferences by means of perspective-taking and feature binding can establish the correspondence between self and other and might thus be deeply anchored in action understanding and other abilities attributed to the mirror neuron system. In conclusion, the model shows that it is indeed possible to develop embodied, neurocomputational models of the alleged principles of social cognition, providing support for the above hypotheses and opportunities for further investigations.

Abstract

Die soziale Orientierung des menschlichen Geistes ist in unserem Alltag sowie unserer Kultur allgegenwärtig. Welche Vorgänge im Gehirn führen jedoch dazu, und wie funktionieren und interagieren sie? Trotz des offensichtlichen Gewichts dieser Fragestellung sind die der sozialen Intelligenz zugrundeliegenden Zusammenhänge und kognitiven Prozesse weitestgehend ungeklärt. Seit einigen Jahren wird ein als Spiegelneuronensystem benannter neuronaler Komplex umfangreich und interdisziplinär betrachtet. Ihm werden weitreichende Implikationen für die soziale Kognition zugeschrieben, so etwa das Verstehen der Aktionen, Intentionen und Emotionen anderer. Die Theorie der *'Embodied Cognition'* betont, dass die verarbeiteten und hergeleiteten Strukturen in unserem Geist erst durch unser Handeln und unsere körperlichen Erfahrungen hervorgebracht werden. So soll auch unser Verständnis anderer dadurch zustande kommen, dass wir ihre Handlungen mittels der durch unseren eigenen Körper erworbenen Erfahrungen simulieren. Es bleibt jedoch zunächst offen, wie etwa visuell wahrgenommene Bewegungen anderer Personen auf grundsätzlich sensomotorisch koordinierte Zustände abgebildet werden, und welche mentalen Prozesse entsprechende Simulationen anstoßen. In Anbetracht der antizipatorischen Natur unseres Geistes wurden auch Prinzipien der prädiktiven Codierung (*'Predictive Coding'*) mit Handlungsverständnis in Zusammenhang gebracht. In dieser Arbeit schlage ich eine kombinierende Hypothese aus *'Embodied Simulation'*, prädiktiven Codierungen, und perzeptuellen Inferenzen vor, und untermauere diese mithilfe eines neuronalen Modells. Das Modell lernt (i) Codierungen von körperlich kontextualisierten, selbst-bezogenen, visuellen und propriozeptiven, modalen und submodalen Reizen sowohl als auch kinematische Intentionen in separaten Modulen, lernt (ii) zeitliche, rekurrente Vorhersagen innerhalb der Module und modulübergreifend um konsistente Simulation teilweise nicht beobachtbarer, verteilter Zustandssequenzen zu ermöglichen, und wendet (iii) top-down Erwartungen an um perzeptuelle Inferenzen und perspektivische Vorstellungsprozesse anzustoßen, so dass die Korrespondenz von Beobachtungen zu den gelernten Selbstrepräsentationen hergestellt wird. Die Komponenten des Netzwerks werden sowohl einzeln als auch in vollständigen Szenarien anhand von Bewegungsaufzeichnungen menschlicher Versuchspersonen ausgewertet. Die Ergebnisse zeigen, dass das Modell bestimmte Handlungtypen simulieren und unter Zuhilfenahme der eigenen körperlichen Erfahrungen beobachtete Handlungen nachvollziehen kann, indem motorische Resonanzen und intentionale Inferenzen resultieren. Desweiteren zeigen die Auswertungen, das perzeptuelle Inferencen im Sinne von Perspektivübernahme und Merkmalsintegration die Korrespondenz zwischen dem Selbst und Anderen herstellen können, und dass diese Prozesse daher tief in unserem Handlungsverständnis und anderen den Spiegelneuronen zugeschriebenen Fähigkeiten verankert sein können. Schlussfolgernd zeigt das neuronale Netz, dass es in der Tat möglich ist, die vermeintlichen Prinzipien der sozialen Kognition mit einem körperlich grundierten Ansatz zu modellieren, so dass die oben genannten Theorien unterstützt werden und sich neue Gelegenheiten für weitere Untersuchungen ergeben.

To my parents, without whose faith,
encouragement, and (action) understanding
throughout my life this thesis
would never have been written.

Acknowledgements

Particular appreciation goes to my doctoral supervisor Prof. Dr. Martin Butz who opened a great many of prospects, helped me to surpass myself, and prosper both as a scientist and person. Thank you for the years of trust and support during my doctorate and beyond. I would like to also thank Prof. Dr. Heiko Neumann for taking the time to review this thesis.

I kindly thank my colleagues, first and foremost Dr. Johannes Lohmann, Dr. Anna Belardinelli, and Dr. Sebastian Otte, for help, support and the right balance of inspiring discussions and constructive distraction.

Credits to my good old friends from Kassel and all the wonderful people that became friends during and because of my study and doctorate in Tübingen. Without you, many a rough time would have been unbearable, and you gave me so many valuable views from so many different perspectives.

This dissertation was kindly supported by postgraduate funding of the state of Baden-Württemberg, Germany. The motion tracking data used in this project was obtained from Carnegie Mellon University, created with funding from NSF EIA-0196217.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Humans are exceptionally socially capable creatures. We interact with and learn from others our entire life, beginning at a very young age when we start imitating others and recognize the meaning of their movements, gestures, gaze directions, facial expressions, et cetera (Carpenter et al., 1998; Want and Harris, 2002; Elsner, 2007; Klinnert, 1984). Not so much later in life, we are able to empathize with them and to slip into their shoes, we can project our experiences onto their situations, we can figure out the reasons for, and intentions behind their deeds and words even in very complex scenarios. Apparently, a form of *emotional intelligence* (Goleman and Griese, 1996; Murphy, 2014) develops from very basic perceptual skills which is steadily enriched by the experiences we make in our lifetime. Human cognition is built on social cognitive skills like these so much that a culture has co-evolved being specifically centered around social norms, communication, interaction, and coexistence in general (Helman, 2007; Tomasello, 1999). Accepting the computational theory of mind (Rescorla, 2017) – the notion that the brain is an information processing system – how it manages to be social on a functional level is a matter of much speculation, and not even the involved basic perceptual mechanisms are identified and understood with full certainty to this day. With good reason, the distinctive social orientation of the human mind is a highly disputed topic among cognitive scientists.

Many forms of social behavior require an understanding of others' bodily actions. As introductory example for facets of this quality, let us assume that we observe a man crossing a street. While this situation may appear trivial at first glance, in fact, it can be stated with some humor that if we completely understand (our perception of) the street-crossing man, we are able to completely understand major aspects of what makes us human, and finally shed light on the *"dark matter of social neuroscience"*, as social cognition was recently named (Przyrembel et al., 2012). Let me at first elaborate on well-chosen cognitive processes involved in this example that will help to classify this work in the huge context of neuroscientific, developmental and behavioral studies centered around action understanding. The mentioned aspects will resurface in more technical terms in the course of this thesis and will be detailed and underpinned in the following chapters.

First, when we observe the man crossing the street, we were able to recognize the shape of the man within fractions of a second. As usually happens, we did not have trouble to segregate them from the street, or distracting or irrelevant elements. We rather selectively focused the man and directed attention to visual features of the scene that are salient, where saliences are vigorously governed by the information that our brain seeks, or in other words, the expectations that it tries to confirm or contradict (Oliva et al., 2003). Amongst others, this information comes in terms of grouped features – or forms – and their motion (see e.g. Grossberg, 2007). Thus, several of visual features were aggregated to form a stable, descriptive percept of the man, his body, and his current movement.

From this percept of bodily, biological motion, we were able to identify a type of action – the man is presumably walking – a perception of which we are arguably used to. We learned the appearance of such a movement from a rich amount of similar experiences in our lives. While some of the first impressions of this kind may certainly have been obtained by observing our parents and others (as for example accentuated by Heyes, 2010; Nagai, Kawai, and Asada, 2011; Saby, Marshall, and Meltzoff, 2012; Froese, Lenay, and Ikegami, 2012), much of our actual *understanding* of walking comes from the perception of our own body (cf. Gallese and Goldman, 1998; Meltzoff, 2007; Gallese, 2007; Catmur, Walsh, and Heyes, 2007; Gallese et al., 2009). Most obviously, we cannot directly observe or feel the muscular activations of another person. We nonetheless have a rather precise notion of how we would have to move our muscles if we were in their situation. Interestingly, we might even find ourselves reflexively moving in response to the action that we see – a phenomenon which was termed *ideomotor* *effect* by Carpenter already in 1852 (Carpenter, 1852), emphasizing that *ideas* (by means of more abstract mental representations) and *motor* representations are involved in the process of observing others. Similarly, observation-triggered motor preparations in human and nonhuman primate brains, also referred to as *motor intentions* (cf. Jeannerod, 1994), have been continuously reported in more recent neuroimaging literature (e.g. Di Pellegrino et al., 1992; Fadiga et al., 1995; Toni, Thoenissen, and Zilles, 2001), highlighting the fine line between observation and action.

On a rather latent level, action understanding thus comes in the form of preparing motor responses that correspond to the actions that we see. On higher and predominantly conscious levels of mental abstraction, understanding certainly goes way beyond motor reenactment and recognition of action classes. By the example of the walking man, we may probably ask ourselves questions like: Where is the man going to, what is his destination? Is he aware of the car approaching from the right while he is looking to the left? Why didn't he take the crosswalk nearby? To answer questions like these, we evidently need to be able to integrate the whole situational context, and to attribute intentionality and a whole mindset to the observed person. In doing so, we need to *simulate* possible subsequent and alternative situations, mindsets, and intentions. Simulating means activating memories of perceptions we do not currently experience, while nonetheless being somewhat triggered by and predicted from the recent observations. These simulations are fetched for imagination more or less consciously, and may range from concrete sensory perceptions like touch to high-level mental concepts like words. However, high-level social reasoning and the inclusion of context are not in the focus of this thesis (for own work on this, see Schrodt et al., 2017; Schrodt, Röhm, and Butz, 2017; Schrodt, Lohmann, and Butz, 2016; Ehrenfeld, Schrodt, and Butz, 2015), and respective modeling approaches usually rely on discrete, symbol-like representations. Following the idea of universal, subsymbolic principles of information processing in the brain (e.g. Friston, Kilner, and Harrison, 2006; Barsalou, 2008; Clark, 2013; Butz, 2016), I argue that the same principles that allow the basic simulation of actions in visuomotor domains can be applied to more complex simulations of context and theory of mind, which is why we have to understand the former in the first place. As a first step towards this understanding, this thesis considers the inference of abstract, intention-like action types from observed actions and visuomotor simulations, and it completely relies on neurocomputational cognitive models.

In a nutshell, our understanding of observed actions resides in modally different but strongly entangled cognitive domains that – at the least – represent visual, motor and intention-like encodings, which are governed by our own embodied experience, and which also respond to the observation of others. Our expectations influence where we direct attention to, how it is integrated, and how we interpret and understand our observations. Still, after making these likely assumptions, candidates for functional links between them are yet to be exposed.

First, how does the brain *identify and recognize* action-related visual features that it learned from self-observation, when someone else is observed from a completely different, allocentric vantage point? I suggest that during action observation (i) top-down saliences are primarily driven by encodings known from self-perception, such that features are selected and grouped that match embodied expectations, and that (ii) visuo-spatial perspective-taking transforms the selective perception onto an egocentric frame of reference to establish the necessary correspondence between self and other. Both of these processes are driven by minimizing the differences between predicted sensory inputs – governed by embodied encodings – and observed sensory inputs. Thus, embodied encodings are generative and predictive in nature to allow for spatial perceptual inference of (visual) correspondences. They are synergistically activated through perceptual adaptation and attention, and in turn drive the perceptual adaptation and attention as an active mental ability.

Second, how does the brain infer or prepare *motor representations* from visually recognized features? I suggest that the brain continuously simulates visual as well as motor and intention states at distributed neural sites and strives for overall consistency in the activated representations. Once the perspective of another actor is taken, the visual domains are focused on this observation, which primes motor and intention simulations (cf. Castiello et al., 2002; Edwards, Humphreys, and Castiello, 2003). The three components converge to maximally consistent, distributed state trajectories to reenact the encountered action by means of embodied codes. Given that motor execution is being suppressed, such a distributed attractor can be defined as a fundamental understanding of action. Similarly, this theory can explain how the brain is able to imagine actions without sensory stimulation, and bridge the gap to imitative behavior: Observed actions can be imitated by not suppressing motor execution, or imitated later by activating respective episodic memories of the perception.

The theoretical approach pursued in this thesis agrees with the notion of the direct-matching hypothesis (Gallese et al., 1996; Iacoboni et al., 1999), stating that *"an action is understood when its observation causes the motor system of the observer to 'resonate'"* (Rizzolatti, Fogassi, and Gallese, 2001). In other words, motor cognition (Sommerville and Decety, 2006) is a (self-sustaining) mental simulation based on the experienced, bodily spatial and temporal contingencies, while being primed by visual observations when the observer is willing to take the perspective of another actor. This formulation fits into the theory of embodied simulation, which assumes that we simulate observed actions by means of our own embodied codes to understand others (Gallese, Keysers, and Rizzolatti, 2004). Equally, this theoretic model agrees with the framework of predictive coding, which argues that the brain is a *prediction machine* that continuously matches bottom-up sensory stimuli to top-down expectations, and that one of its primary functions is to minimize emerging deviations on hierarchical levels (Clark, 2013).

In sum, action understanding is seen here as an ability that is strongly related to *embodied simulation*, *predictive encodings*, and *perceptual inference*. This thesis considers the functional origins, organization and principles of the ability to recognize, understand, and reenact others' bodily actions in these terms. It specifically attends to the perceptual foundations of understanding others' bodily actions, and with it a foundation of social cognition. To do so, artificial neural network methods are applied to identify computational candidate mechanisms of action understanding. The neural network model learns visual, motor, and abstracted intention representations from self-observation, and predictively correlates the respective encodings. It thus obtains the ability to simulate forward in time different types of actions in distinct, distributed modalities that mutually predict each other and thus synchronize over time towards consistent overall activations. The simulation can be pushed onto specific attractor state sequences when observing similar actions from arbitrary perspectives, and in potentially arbitrary scenes: Using its embodied, top-down expectations, the network will identify the relevant features, group them correctly to whole-body Gestalt perceptions, and take the perspective of the observed actor gradually. Thus, the network *imagines* the spatial perspective of an observed actor to reenact their actions in terms of the own, predictive action model. In doing so, it is able to identify the type of action, which can be seen as a preparation step for inferring higher mental representations, in which certainly other aspects like the integration of context are involved as well.

In Chapter 2, I will clarify the term *action understanding* in the light of psychological and neuroscientific evidence as matters stand, as well as in a brief historical context. This will reveal the emergence of big open questions over recent years, resulting in the contemporary disputation of the topic. Chapter 3 will go into detail on my hypothesis about action understanding from which I believe that it can unify some of the contradicting positions on the topic, support specific assumptions, and contribute to answering some of the open questions.

I will relate to own modeling work and the work of others in Chapter 4. The neurocomputational model presented in this thesis as well as its assumptions and presuppositions are detailed in Chapter 5, where I will first explain how visual and motor-related perceptions are processed, and how the model is able to selectively group visual features and imagine different perspectives onto them. Then, I will explain how visual, motor-related and intention codes are encoded and mutual predictions are learned that form overall consistent attractor state sequences.

Results of the model are presented in Chapter 6. The model will be trained on three short motion capture trials, and validated on a variety of similar and also dissimilar trials. After explaining the format of these stimuli, I will proceed with explaining the network parameters. I will then show the performance of the network in developing generative and predictive embodied encodings. In the following experiments, I will show that the network is able (i) to identify and bind visual features also of completely novel actions and imagine them from an egocentric perspectives, (ii) to simulate and "understand" the corresponding action intentions and proprioceptions based on the embodied experience, and (iii) to consistently imagine whole-body actions in distributed domains also without sensory stimulation. The perceptual characteristics of the model are compared to studies with human subject groups.

A discussion and conclusion of the insights of the model and results follows in Chapter 7, where prospects for future work and open questions are detailed.

Chapter 2

A Brief Summary on Action Understanding

In 1992, neuroscientist di Pellegrino and colleagues published an article about cells in the macaque inferior premotor cortex that discharge during the execution of an action – and discharge as well while passively observing an actor performing a similar action (Di Pellegrino et al., 1992). At the time, the role of these cells was already speculated to be to *understand* observed motor events. Later, they were termed *mirror neurons*, pointing to the property they seem to mirror the behavior of others (Rizzolatti et al., 1996). A mirror neuron was defined by the property that it matches observation and execution of a particular action by encoding the action visually as well as in terms of corresponding motor responses (Gallese et al., 1996). Rizzolatti et al. proceeded to characterize the role of mirror neurons as to enable primates

> *"... to recognize the presence of another individual performing an action, to differentiate the observed action from other actions, and to use this information in order to act appropriately."*
>
> — Rizzolatti et al., 1996

Although the functional mechanisms behind this mirror neuron property and the implications of the findings remained speculatively, these assumptions established the term *action understanding* as the primary role of mirror neurons.

In 2004, Rizzolatti and Craighero concluded from other neurophysiological findings the existence of a *mirror neuron circuit* (see Figure 2.1), involving two further brain regions besides area F5 in inferior premotor cortex (Rizzolatti and

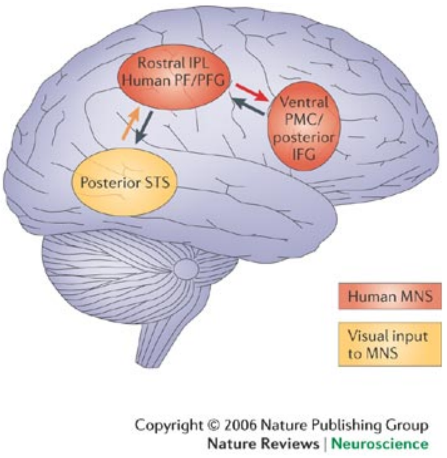


FIGURE 2.1: The mirror neuron system. In macaques, mirror neurons have been found in the premotor cortex (PMC) / inferior frontal gyrus (IFG) area F5 and area PF in inferior parietal lobule (IPL). Together with the superior temporal sulcus (STS), the regions form the mirror neuron circuit. Frontal regions are assumed to be involved with goal states of actions, parietal regions are associated to sensorimotor integration, spatial cognition, and imitation, while temporal regions are assumed to recognize views of actions as input to the mirror neuron system. In humans, the existence of a mirror neuron system has been suggested. Figure from Iacoboni and Dapretto, 2006.

Craighero, 2004). Further mirror neurons were found in area PF of the macaque brain, which forms the rostral part of the inferior parietal lobule. About two thirds of PF cells show the mirror neuron property described beforehand. PF bidirectionally communicates with area F5. Moreover, the superior temporal sulcus (STS) was designated, which was already well known for encoding views of action types visually (Bruce, Desimone, and Gross, 1981; Oram and Perrett, 1994; Perrett et al., 1985). Like mirror neurons, STS cells both respond to observed and executed actions, however do not seem to encode motor responses and thus cannot be termed actual mirror neurons. Nonetheless, STS intercommunicates with PF and was suggested to provide the main visual input for action understanding, thus being a crucial part of the mirror neuron system (Rizzolatti and Craighero, 2004; Ulloa and Pineda, 2007; Pavlova, 2012; Cook et al., 2014).

According to the literature, the functional principle of mirror neurons in F5, PF and STS can be briefly summarized as follows (cf. Rizzolatti and Craighero, 2004; Iacoboni and Dapretto, 2006)[1]. STS provides higher-order *visual descriptions* of observed actions to the parietal region PF. PF mirror neurons extract the corresponding *motor descriptions* and communicate them to the frontal area F5 (Iacoboni et al., 1999; Chaminade, Meltzoff, and Decety, 2005). F5 mirror neurons encode actions in a *goal-directed* manner. They seem to primarily respond to effector actions (such as grasping) and require an interaction with an object (such as food) to do so (see also Rizzolatti and Luppino, 2001). Including *efference copies* sent top-down from F5 to PF, and from PF to STS, the three macaque brain areas are assumed to form an action observation network that involves sensorimotor integration.

Regarding the tremendous extent of other parieto-frontal neural processing pathways (see e.g. Caminiti et al., 2017) that are involved in sensorimotor integration, the mirror neuron system was considered in a broader context than grasping actions. Rizzolatti and Craighero, 2004; Rizzolatti and Craighero, 2005 postulate that the mirror neuron system is at the basis of complex abilities like imitation learning, gestural communication, understanding others' intentions and emotions, and might as well have been involved in speech evolution (Rizzolatti and Arbib, 1998; Arbib, 2010). Rizzolatti and Sinigaglia, 2007 argue that understanding intentionality is firmly grounded in a fundamental interweaving of motor and intention components of action. Taken together, insights into a human action understanding network would have farreaching implications. The implications and theories mentioned beforehand, however, are anything but generally accepted or well established, which I explain in the following.

[1] Note that not all of the neurons in the respective areas are mirror neurons, and that they respond to single or multiple modal stimuli. Other neurons in the respective areas may have a supportive function.

2.1 The Dilemmas of Action Understanding

The existence of mirror neurons and a mirror neuron system also in humans – which was assumed by (Rizzolatti and Craighero, 2004) – has been substantiated further by other neuroscientists (e.g. Dinstein et al., 2007; Chong et al., 2008). However, both the existence of a mirror neuron system in humans, and theories of its potential role for action understanding are still subject to intense controversies. Single cell recordings in humans are practically infeasible, and thus there is no direct evidence for mirror neurons in the human brain. Several studies have found no indirect evidence for the inclusion of motor simulation in action understanding and suggest that, instead, the inclusion of context-sensitive inferential processes plays an essential role (Brass et al., 2007; Kilner and Frith, 2008; Lingnau, Gesierich, and Caramazza, 2009). Moreover, studies with apraxia patients – who can be impaired in producing particular actions despite being unimpaired in understanding these actions when observed (Mahon and Caramazza, 2005) – can be interpreted as evidence that motor simulation is not involved in action understanding. On the other hand, the findings might also indicate that motor simulation is not a *mandatory* component of action understanding, highlighting the flexibility of the human mind.

With respect to learning and mental development, the involvement of mirror neurons in mediating *imitation* was already suspected by Jeannerod, 1994. Iacoboni et al., 1999; Heiser et al., 2003; Molenberghs, Cunnington, and Mattingley, 2009 verified that mirror neuron areas take part in imitation. On the other hand, Rizzolatti and Craighero, 2005 point out that mirror neurons do only map an observed action to the motor system of the observer, if the action *already belongs* to their motor repertoire. This raises the question how and if mirror neurons may help infants to imitate an action that they have never performed before, the question what facilitates *learning* by imitation in the first place. While Rizzolatti and Craighero, 2004; Rizzolatti and Craighero, 2005 argue that imitation cannot be the primary function of mirror neurons from an evolutionary perspective, it might well be that other neural circuits underlie learning by imitation and might thus bootstrap the development of mirror neurons early in life (Buccino et al., 2004; Iacoboni and Dapretto, 2006). Meltzoff and Moore, 1977; Meltzoff and Moore, 1983 report imitation of facial and manual gestures in infants as early as a few days of age and suggest an innate, genetic mechanism for basic imitation abilities. In contrast, Heyes, 2016 states that newborns do not imitate, and imitative behavior is a result of cultural rather than genetic evolution. The contradiction of these theories and positions leads to what can be termed an *imitation dilemma*, as it is still highly debated today.

Another dilemma is known as the *correspondence problem* (Nehaniv and Dautenhahn, 2002; Heyes, 2001). The core assumption of mirror neurons as cognitive mechanism for action understanding, as stated above, is the existence of a matching mechanism from sensory to motor representations. Observed and executed actions are co-encoded in STS (cf. Molenberghs et al., 2010), which provides visual input to the mirror neuron circuit. Yet how does a visual representation of an observed action, which is inevitably viewed from a perspective that does not correspond to a self-perceptual perspective, activate the same network of areas? It is conceivable that both executed and observed actions share a *common code*, such that visual representations of the very same are invariant to differences in perspective (Prinz, 1984), and that mirror neurons implement the physiological mechanism for the common coding of perception and action (Keysers, 2011). However, neuroimaging studies strongly suggest that visual actions are represented view-dependently in STS (see Subsection 2.2.1). Along these lines, the mirror neuron system was proposed to provide an *"automatic transformation"* as *"functional bridge between first-and third-person perspectives"* (Decety and Meltzoff, 2011). Another view on the matching from observed to executed actions – or own behavior to the behavior of others – was formulated by Heyes and Ray, 2000; Heyes, 2001; Heyes, 2010; Cook et al., 2014. The authors assert that *"mirror neurons may be a byproduct of associative learning"*, and that they *"do not play a dominant, specialized role in action understanding"* (Heyes, 2010). Thus, mirror neurons primarily develop from sensorimotor experience when interacting with others, without having a specific evolutionary purpose. As solution to the correspondence problem, the authors suggest that different views of observed actions are directly associated to own motor representations to explain the mirror neuron property (Ray and Heyes, 2011).

In preliminary conclusion, the evolutionary and developmental origins of mirror neurons as well as their functional principles and cognitive roles are still far from being clarified to this day. In particular, there is a need for an investigation of the following questions:

- Are mirror neurons a byproduct of mental development, or is there a genetic predisposition? What is their purpose?
- Do mirror neurons facilitate imitation learning? Or is there an innate mechanism?
- What comes first, learning by imitation, or learning how to imitate?
- Which mental processes equate action and perception? What establishes the correspondence between observations of others, and self-representations?
- Is there even a mirror neuron system in humans?
- What is possible from a computational perspective? What functionality can explain the observations and verify the made assumptions?

To contribute possible answers to these questions, I will first elaborate on how actions are believed to be encoded in the substrates of the mirror neuron system in terms of biological motion, motor encodings and action goals. The main contribution of this work will be to investigate the computational feasibility of specific assumptions and to suggest candidate mechanisms for perceptual and learning processes related to the mirror neuron system.

2.2 Action Encodings in the Brain

A voluntary action can be defined as a goal-directed movement that pursuits a particular, typically reward-directed outcome, or action effect. As such, it follows a specific *intention* by executing *motor commands* that result in *visible, bodily motion*. The respective neural codes are believed to be represented at distributed brain areas, which are linked via bottom-up, as well as top-down, generative processes (see e.g. Clark, 2013). It is typically assumed that action goals and intentions are encoded inferior frontally, motor codes and plans posterior parietally, and biological, mainly visually-driven motion patterns in the superior temporal sulcus (cf. Iacoboni, 2005; Kilner, 2011; Turella et al., 2013). In the following chapters, I will discuss how these three main components of action and – in the context of the mirror neuron system – action understanding are believed to be encoded in the brain. Other aspects of action, like decision making, planning, inclusion of context, and anticipation of environmental effects are meaningful, but not in the focus of this thesis.

2.2.1 Biological Motion

According to the two-streams hypothesis of visual perception (Goodale and Milner, 1992), the ventral stream is primarily involved with form perception and object recognition. The stream extracts particular information, such as orientation, size, color, and shape features via the occipital regions V2 and V4 and forwards them to the temporal lobule. In this context, theories of Gestalt perception consider the question in which way, and by which characteristics (the so-called *Gestalt laws*) the brain integrates visual features and recognizes them as *one* object, or as dynamically connected (Hartmann, 1935; Jäkel et al., 2016). In terms of action observation and understanding, the human body can be considered an important Gestalt percept, and fMRI studies suggest that neurons already in visual cortex are specifically tuned to the perception of the human body (Downing et al., 2001).

Psychological experiments that investigate visual perception related to action observation often make use of *biological motion* stimuli. Biological motion primarily refers to the visualization of human or animal bodily motion. In a great number of studies, these stimuli are abstracted and reduced to contain the minimal amount of information that is sufficient to perceive the underlying movements. As Johansson, 1973 showed, moving light points can trigger the perception of a walking person in the observer. Such biological motion stimuli became common-use in experimental psychology and are mostly termed *point-light displays* (see Figure 2.2). Johansson's results show that observed motion features are integrated by the brain to detect and understand whole bodily motion.

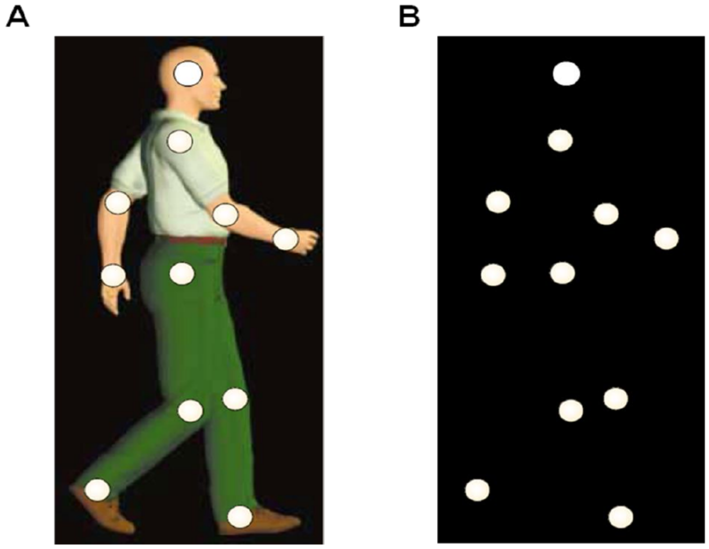


FIGURE 2.2: A point light stimulus example. The point lights (**B**) typically correspond to joint coordinates (**A**). Without motion, actions are hard to recognize from point light stimuli. Figure from Giese, 2013.

Giese and Poggio, 2003 summarize properties critical for the recognition of biological motion. Among others, biological motion recognition is highly invariant to differences in position, scale, speed, body morphology, and exact posture control, as well as incomplete representations or variances in illumination. Thus, the recognition of biological motion is highly robust and general. Nonetheless, studies demonstrate that humans are able to identify the gender (Runeson and Frykholm, 1983) and the identity (Cutting and Kozlowski, 1977) of an observed person solely from point-light displays. Humans perform better in action recognition when observing recordings of their own actions in comparison to the actions of others (Beardsworth and Buckner, 1981). This indicates the involvement of the own motor experience in action understanding. Also, while being most general, the human perceptual system is able to extract very specific and distinct information from few biological motion cues. At that, the relative motion of visual key features may contain the most crucial information (cf. Garcia and Grossman, 2008; Thurman and Grossman, 2008).

Biological motion stimuli are believed to be encoded in STS (Bruce, Desimone, and Gross, 1981; Oram and Perrett, 1994; Perrett et al., 1985). As described before, action representations in STS are considered to provide visual input to the mirror neuron system, and as such they are most important for the development of attributes linked with it (see e.g. Cook et al., 2014; Grossman et al., 2000; Gallese, 2001; Pavlova, 2012; Puce and Perrett, 2003; Ulloa and Pineda, 2007). Amongst others, STS seems to integrate *form and motion* information into *whole body* perceptions of the observed movements (Oram and Perrett, 1996). STS partially encodes biological motion in a *retinotopic* organization (Gattass and Gross, 1981; Huk, Dougherty, and Heeger, 2002), and the major portion of neurons in the posterior STS seems to encode viewer-centered representations of specific movements to the effect that their activation depends on the *type* of movement observed, as well as on the observer's current *vantage point* (Oram and Perrett, 1994; Perrett et al., 1985; Perrett et al., 1989; Perrett et al., 1991). As well, the recognition performance decreases with the amount of rotation an action is perceived from with respect to common or canonical perspectives (Pavlova and Sokolov, 2000).

The findings mentioned beforehand indicate that visual encodings which are assumed to provide the main input to the mirror neuron circuit exist in rather viewer- or eye-centered coordinate frames, which strongly underlines the existence of functional mechanisms that map observations of others to self-observations. They also evidence that encodings of biological motion support the inference of action-related information based on type-specific, whole-body form and motion cues.

2.2.2 Visuomotor Control and Spatial Perception

Motor skills are the brains coupling from intentions to goal-directed behavior. In the context of social cognition, however, they are assumed also to be the coupling from observing others' behavior to understanding their intentions. The respective areas are essentially located in parieto-frontal regions of the brain, which share complex neural circuits for bidirectional communication (cf. Caminiti et al., 2017). In particular, the connectivity between parietal and frontal mirror neuron areas motivates the idea that intentionality can be understood from motivity.

The frontal and parietal cortices are anatomically delimited by the central sulcus. Adjacently in the parietal lobe, neurons in the primary somatosensory cortex integrate and process *tactile and proprioceptive* sensations. Anatomically close, the inferior parietal lobule – which blends into the mirror neuron system – features both somatosensory properties as well as motor properties and seems to be involved in motor planning (Andersen, 2011). Neurons adjacent to the central sulcus in the frontal lobe are located in the primary motor cortex, which combines motor primitives and drives the muscles via self-stabilizing mechanisms in the peripheral nervous system (cf. Butz and Kutter, 2016, p. 278 ff.). Primary motor areas encode a variety of different information, from joint motion and force to spatial goals. The variety of information processed in motor cortical areas reflects their important role in mediating between high-level, goal-directed behavior and concrete motor neuron control (Scott, 2003).

A substantial type of information processed by neurons in motor cortical areas of the frontal lobe has been characterized after recording individual, specific cell tunings to broad *directions* of voluntary limb motion (Georgopoulos et al., 1982). Similarly, cells in area 5 of superior parietal lobule show selectivity for directional limb movements (Kalaska, Caminiti, and Georgopoulos, 1983). In consequence of the findings of Georgopoulos et al., the *population vector hypothesis* was established, suggesting that ensembles of neurons cooperate to generate directional limb movements. Likewise, sensory input to somatosensory and related regions (just as their visual counterparts) is processed by populations of locally receptive cells with tunings to specific stimulus characteristics (Pouget, Dayan, and Zemel, 2000).

Mappings between such population-encoded, afferent and efferent modalities in parietal cortex have been suggested to be established by means of neural *gain-field* structures that modulate the tuning of neural populations in a multiplicative manner (Andersen, Essick, and Siegel, 1985; Salinas and Abbott, 2001; Schrodt and Butz, 2015). This accompanies the finding that the directional tunings of motor cortical cells are not invariant, but rather modulated depending on the current region of the work space (Caminiti et al., 1991). As well, the tuning of cells in primary motor cortex does not represent precisely the direction of effector movement, but is rather skewed along body-relative reference frames

(Scott et al., 2001). When eye movements shift the retinal response to a stimulus, the mappings also have to be adapted. It has been shown that posterior parietal areas compensate for the consequences of eye movements to maintain stable visuo-spatial perceptions (Duhamel, Colby, and Goldberg, 1992). Eye-movements can be considered a form of attention-driven motor control, which highlights the linkage between action, spatial perception, and attention. Taken together, it can be concluded that motor-relevant encodings are represented in *multiple, interacting frames of reference*. Spatial transformations that stem from visual, eye-centered representations *modulate* proprioception and motor perception as well as motor control for goal-directed actions (see also Cohen and Andersen, 2002).

Interestingly, superior and inferior parietal regions – which are involved in the foregoing transformational processes and the mirror neuron circuit – are moreover involved in *imitation* and *perspective-taking* (Jackson, Meltzoff, and Decety, 2006), as well as *mental rotation* (Alivisatos and Petrides, 1997). This substantiates the possibility of a link between spatial abilities and action understanding. The superior parietal cortex indirectly receives input from the dorsal stream of visual processing via the ventral intra-parietal sulcus (Caminiti, Ferraina, and Johnson, 1996) which integrates visual and somatosensory information (Duhamel, Colby, and Goldberg, 1998). Since the dorsal stream processes spatial relationships of perceived objects (Goodale and Milner, 1992), it possibly contributes information about spatial relations that facilitate perspective-taking.

The overall representation of the spatial configuration and location of ones body is termed *body schema* (Gallagher, 2006). The body schema is believed to be represented at different regions of the brain – including the ones mentioned beforehand – and integrates proprioceptive and visual stimuli. It is believed to also represent the body of others for action understanding (Chaminade, Meltzoff, and Decety, 2005). The body schema can be compared to the body image, but differs in the sense that it is closer related to action and space, while the body image is closer related to perception and appearance (Gallagher, 2006).

In conclusion, parietal and frontal regions integrate visual and somatosensory information to control visual attention and motor activity in a goal-directed manner. Multiplicative transformations are applied to translate between reference frames in population-encoded modalities, which include multiple, local reference frames in motor-related encodings, as well as global reference frames in visual encodings. In a similar manner, visual correspondences of the own body scheme to observed persons are possibly established via spatial inferential processes like perspective-taking to facilitate action understanding and imitation.

2.2.3 Goal-directed Action

Behavior is fundamentally reward-oriented and goal-directed. Rewards and goals are partially generated based on bodily states. For example, the hypothalamus regulates body temperature, hunger, and sleep cycles by releasing hormones and neurotransmitters (i.e. leptin in case of hunger) when the respective states deviate from their normal, desired, healthy states, and by releasing dopamine when the bodily states return to the desired level. The release of dopamine has been linked to reward and learning (Wise, 2004). As such, the hypothalamus is part of a homeostatic, motivational system that generates urges based on bodily states and supports learning of rewarding actions and motor skills. The urges result in drives to act in order to bring back the motivational system towards homeostasis. These drives result in goal-directed, intentional behavior that is believed to yield the desired rewards. Thus, the brain is a self-regulatory system to the effect that behavior can partially be seen as active inference of bodily states.

Motivational and reward-related areas such as the hypothalamus are phylogenetically rather old and primitive systems. Clearly, not all of our behavior (at least directly) aims at eating, sleeping, or reproduction. Beside the basic, physiological urges, more complex motivations evolved, such as a need for safety, love, esteem, and self-actualization (Maslow, 1943). These urges primarily correlate with the complex social alignment of the human mind. Thus, more generally, resulting behavior aims at actively producing situations associated with rewards.

The frontal lobe of the human brain is particularly involved in high-level cognition. It has often been mentioned in connection with reasoning, consciousness, theory of mind, personality, social behavior, but also concrete motor control. More importantly in the context of the mirror neuron system, it is believed to produce all types of intentional states that result from the motivational urges mentioned above. For instance, amongst others, the mirror neuron area F5 in macaque inferior premotor cortex represents the execution as well as observation of grasping food with the intention of eating (Rizzolatti and Luppino, 2001). Generally, an intentional action can be decomposed into a goal or action outcome, as well as kinematic parameters, and thus, an *action intention* can be defined in both ways (Butterfill and Sinigaglia, 2014). For example, condensed movement complexes like walking or running can be considered a simple form of *kinematic representations of intentions*, or *motor intentions* (Jeannerod, 1994), which is assumed to be processed in the mirror neuron system (Jacob and Jeannerod, 2005), and which also typically follow a specific action goal (i.e. locomotion to a certain location). Some authors assume that outcome representations of actions and thus *finite goal state intentions* are not processed in the mirror neuron system itself, but inferred form its encodings (Jacob and Jeannerod, 2005), while others argue that mirror neuron regions also co-encode goal states of actions (C. Hamilton and Grafton, 2007).

In the light of the scope of this thesis, I will neglect all other aspects of intentionality in the following chapters, and focus on a first understanding of how compact, kinematic intention encodings may be inferred from more complex motor simulations, and how motor simulations may be inferred from encodings of biological motion.

Chapter 3

A Hypothesis on Action Understanding

As a consequence of the debates and dilemmas mentioned in the last chapter, none of the theories on action understanding has widely been accepted (see also Dinstein et al., 2008; Hickok, 2009; Heyes, 2010). In the following, I will clarify the theoretical approach for modeling action understanding in this thesis. It lies in the nature of the current state of scientific knowledge on this topic that many of the following statements are hypothetical, matter of opinion, and disputable. At the same time, it is necessary to side with specific assumptions to provide a theoretical framework. Although a plenty of side steps, abstractions, and simplifications are adopted following these ideas, the model generally pursuits the approach briefly outlined in the introduction: To model and explain action understanding in terms of embodied simulation, predictive coding, and perceptual inferences.

3.1 Embodied Simulation

Although humans are particularly good at learning from others, many experiences can only be formed by self-perception and introspection, and clearly all of our experiences are gathered through, and shaped by our own body. Cognitive science has therefore recently undergone a pragmatic turn, focusing on the enactive roots of cognition. Today, cognition is seen as being action-oriented and grounded in sensorimotor couplings – as generating meaning of the world through action. Hence, *"the functioning of cognitive systems is thought to be inseparable from embodiment"* (Engel et al., 2013).

Embodied cognitive states, according to Barsalou's simulation hypothesis (Barsalou, 1999; Barsalou, 2008), are situated simulations that temporarily activate – or reenact – particular events by means of a set of embodied modal codes. This means that, when we observe an object, our embodied experiences associated with that object are triggered. These range from sensory experiences like touch or smell to episodic, situated memories and introspective states that occurred during interaction with that object. According to the theory, the brain simulates these experiences to produce perceptual inferences and form entity categories and concepts of the world. In this process, simulation is thought to be a *"core form of computation in the brain"* (Barsalou, 2008), ranging from conscious mental imagery, to unconscious processes of motor simulation (see also Stevens et al., 2000). Barsalou, 2008 concludes that *"simulation is the reenactment of perceptual, motor, and introspective states acquired during experience with the world, body, and mind"*.

Agreeing with this notion, action understanding is possible by simulating others' mental states by means of the own, embodied states. Gallese et al., 2009 also hold that *"action understanding may be primarily based on the motor cognition that underpins one's own capacity to act"*. During the observation of others, however, neither information about their proprioceptions nor their intentions is directly accessible. Although there is a bunch of empirical evidence for simulation processes involved in different cognitive processes, the embodied simulation theory does not answer how exactly simulation might work functionally, and how multimodal, embodied encodings are represented and 'triggered' upon observation. Generally agreeing with the notion of embodied simulation for action understanding, I suggest that predictive coding presents a framework that can answer these questions, which I explain in the following.

3.2 Predictive Coding

Chapter 2 stressed that the particular encodings that appear to be involved in the mirror neuron system and action understanding are encoded at different neural sites, and implement different cognitive functions. Although they rely on similar neural characteristics, the types of information they represent are distinct. To put it simply, parietal regions primarily represent visuo-spatial relations that blend into proprioceptive (and other somatosensory) perceptions and facilitate motor control, temporal regions encode visual, view-dependent form templates of actions, and frontal regions process intention signals that stem from self-regulatory motivational systems.

The brain processes these types of information in a hierarchical manner, locally extracting task-relevant information. This cascading involves bottom-up processing, as well as top-down processing. Following the idea of predictive coding (Clark, 2013), bottom-up processing activates encodings that correspond to particular features of lower cortical levels, while top-down processing in turn *predicts* – or generates expectations about – the activated encodings in lower cortical levels. Emerging prediction errors are minimized by particular cognitive adaptation processes. These adaptations refer to learning, filtering, but also to perceptual inference and action.

Related to the framework of predictive coding, Butz, 2016 suggests that additional learning biases and structural priors are needed to explain cognition on a sub-symbolic level. The author proposes the existence of three different types of top-down and bottom-up predictive encodings, which I interpret freely as follows: *Top-down predictive encodings* generate expectations in the form of perceptual templates or momentary expectations. They are matched locally to submodal perceptions of lower cortical levels. Learning of template encodings occurs via minimizing prediction errors that emerge with respect to these perceptions. The perceptions are computed given structural neural biases and extract particular stimulus characteristics, thus being centered in particular frames of reference, representing particular submodal features. A predictive template might for example represent a specific object-centered bodily Gestalt image or a postural arm configuration perceived proprioceptively. *Spatial predictive encodings* transform different reference frames onto each other. Their role is to activate template encodings by applying spatial transformations that project onto the respective reference frame in which they exist. These spatial transformations include dynamic perceptual inferences that minimize top-down prediction errors systematically. Finally, *temporal predictive encodings* encode the changes of the activated template encodings over time, and thus represent the respective temporal contingencies.

I assume that temporal predictive encodings learn the state progress both locally in a modal or submodal context, as well as selectively in a broader, cross-modal context. They form bidirectional neural circuits that predict the typical progress of their own neural activity, as well as the activity progress in other, related neural encodings. Akin to the ideomotor theory (see Stock and Stock, 2004, for a historic review), this includes predictions given bodily actions (motor commands) about the perceptual effects (the resulting changes in visual perception), and vice versa. For example, given the proprioception of an initial *posture* of the right arm, a movement to a specific *direction*, with a specific *magnitude* ultimately leads to the visual perception of a *new position* of the arm. Conversely, perceiving the movement of the arm in visual space will result in a specific, corresponding new proprioception.

To encode the *continuation* of movements, I further assume that these bidirectional predictions also cover possible subsequent movement directions and magnitudes. Then, the manifold of sensorimotor contingencies can be restricted by modulation to specific sub-sets by applying top-down priors. Such priors may originate from different mental states, for example caused by visual obstacles, physical constraints, tool use, and most importantly, intention-related codes. As a result, different movements can be achieved via top-down modulation, and because of the bidirectional nature of the encodings, intention priors can be predicted given a sequence of observed movements.

Here, the term *temporal predictive encoding* is used somewhat synonymously with the idea of *direct matching* in Rizzolatti and Sinigaglia's statement about mirror neurons:

> *"The functional properties of these neurons indicate that intentional understanding is based primarily on a mechanism that directly matches the sensory representation of the observed actions with one's own motor representation of those same actions. These findings reveal how deeply motor and intentional components of action are intertwined, suggesting that both*

> *can be fully comprehended only starting from a motor approach to intentionality."*
>
> — Rizzolatti and Sinigaglia, 2007

I propose that the matching from observations to motor representations and intentions is less 'direct', but rather predictive, generative, multidirectional, distributed, modular, and hierarchical in nature. Complex neural circuits as well as spatial transformations are necessary to establish the matching.

This notion agrees with the idea of embodied simulation as stated in Section 3.1. Separate modular neural structures can partially predict their own submodal activity progress, and partially be predicted by other neural structures. Thus, neural activity can be seen as a self-sustaining dynamical system that *simulates* gathered experiences, which are mostly embodied in nature, and converges to overall consistent, distributed attractors. These simulations can be *primed* by bottom-up sensory information via spatial transformations which are driven by top-down expectations. This approach can explain mental imagery, where visual encodings are driven predominantly by internal representations, as well as motor resonance, where motor encodings are driven by visual observations.

As a result, the embodied simulation of a movement is congruent in multiple modalities, if the sequence of postural and motion perceptions does not conflict neither with the learned sensorimotor contingencies, nor the intentional priors. Establishing congruent multimodal simulations from action observations relies on spatial inference processes as described in the following chapter.

3.3 Perceptual Inference

As indicated in Chapter 2, there is presumably a cognitive mechanism that matches self-representations and the representations of others. There is also evidence that such a mechanism relies on spatial transformations. In the last chapter, I have made spatial predictive encodings responsible for such a mechanism, and stated that these include dynamic perceptual inferences that minimize prediction errors. In the following, I suggest how perceptual inferences for self-other correspondence can be related to attention and visuo-spatial abilities.

First, to recognize an object such as the human body of another person, the visual system has to integrate separate visual features into a Gestalt percept (see e.g. Koffka, 2013). This problem was described in 1998 by Treisman as follows:

> *"The 'binding problem' concerns the way in which we select and integrate the separate features of objects in the correct combinations. Experiments suggest that attention plays a central role in solving this problem."*
>
> — Treisman, 1998

Thus, the binding problem is related to attention, which is known to be driven by bottom-up salience cues as well as top-down, task-dependent activity (see e.g. Buschman and Miller, 2007). The visual system directs attention to specific salient features and binds them in a way such that objects, object-relations, and important scene information are recognized.

In a naive computational approach, the binding problem has a very high complexity. Assume that out of $N = 400$ salient visual features, we attend to just $n = 40$ which carry task-relevant information that is to be selected and combined. The number of combinations in which we can select the 40 out of 400 features in a singular order is $\frac{N!}{(N-n)!} \approx 1.6 \cdot 10^{103}$, which exceeds the number of atoms in the universe by orders of magnitude. Clearly, what the brain does is not comparable to a sequential or random search in a database of learned feature constellations, thus there have to be intelligent metrics and perceptual biases involved. Also, it cannot be assumed that these underlying perceptual mechanisms are directly comparable to more elaborate image search algorithms that often directly use transformation invariant features (e.g. SIFT feature based algorithms, Lowe, 1999), but rather rely on continuous patterns of neural computations, topological mappings, and expectation-driven adaptations.

Second, thinking along similar lines, how does the brain activate motor encodings from observed biological motion to understand actions? Following the idea of embodied simulation and predictive couplings of visuomotor self-representations as described before, the observer first has to match the observation to a respective self-representation in *visual* modalities to actively drive an embodied simulation. Then, and possibly only then, they can infer or predict corresponding motor perceptions, and intentions respectively. So how does observed biological motion, which is inevitably viewed from a perspective that does not correspond to a self-perceptual perspective, activate the same areas in the mirror neuron system? Brass and Heyes describe this *correspondence problem* (see also Nehaniv and Dautenhahn, 2002; Heyes, 2001) as follows:

> *"When we observe another person moving we do not see the muscle activation underlying their movement but rather the external consequences of that activation. So how does the observer's motor system 'know' which muscle activations will lead to the observed movement?"*
>
> — Brass and Heyes, 2005

Just as in object recognition, where rich perceptions are matched onto learned, canonical representations, the correspondence problem ultimately leads to the question how the brain can match observed biological motion onto visual self-representations.

The *associative sequence learning* hypothesis proposes that – instead of establishing visual correspondence – allocentric views of actions are *directly* associated to embodied motor representations. These associations are formed e.g. while being imitated, perceiving a mirrored self, as well as synchronous activities with others – and thus mirror neurons develop by sensorimotor experience and social interaction (Catmur, Walsh, and Heyes, 2009; Heyes, 2010; Heyes, 2016). The authors argue that the correspondence problem cannot be solved by innate mechanisms, e.g. as proposed by Meltzoff and Moore, 1977; Meltzoff and Moore, 1983.

However, considering human spatial abilities, several mechanisms have been identified in psychometric studies (Lohman, 1979; McGee, 1979; Eliot and Smith, 1983; Carroll, 1993; Hegarty and Waller, 2004). Amongst them, *visuo-spatial perspective-taking* has been described as a progressive ability to adopt the spatial point of view of another person (Newcombe, 1989; Jackson, Meltzoff, and Decety, 2006, cf.). Perspective-taking is also a slightly vague term used for a variety of cognitive phenomena: While visuo-spatial perspective-taking refers to the ability to merely imagine different visual perspectives in the environment – akin to mental rotation (Shepard and Metzler, 1971; Shepard and Metzler, 1988; Hegarty and Waller, 2004) – empathetic, social, or affective perspective-taking describes the adoption of psychological experiences and mental states of another actor (Ashton and Fuehrer, 1993; Farrant et al., 2012).

It has been shown that sensorimotor resonance to the observation of pain was correlated to perspective-taking, and decreased in observers with racial bias towards the model (Lamm, Batson, and Decety, 2007; Avenanti, Sirigu, and Aglioti, 2010), indicating that intergroup commonality is an important factor for empathy (Galinsky and Moskowitz, 2000). Interestingly, but not too surprising, studies have shown a functional coherence between these two forms of perspective-taking: Racial biases could be decreased in observers who took the spatial perspective of an outgroup avatar in virtual reality (Peck et al., 2013). Again, these findings indicate that visuo-spatial abilities as well as abilities attributed to the mirror neuron system – such as empathy – are highly intertwined.

Perspective-taking can be thought of as a non-discrete mental transformation process that aligns allocentric perspectives with egocentric perspectives to make inferences based on embodied codes and self-experience. According self-centered mental states are believed to be important for self-consciousness, including agency and ownership (cf. Vogeley and Fink, 2003). Distinct neural representations for first-person and third-person perspectives, as well as common processes for their transformation into each other have been suggested (Vogeley et al., 2004), and self-localization experiments show that such egocentric reference frames might be located in upper face or upper torso regions (Alsmith and Longo, 2014).

Furthermore, Kessler and Thomson, 2010 found evidence that spatial perspective-taking is an embodied transformation in which large parts of the body schema are mentally rotated. I support this view, linking it to action understanding, and adding that both feature binding and perspective-taking are two structurally biased forms of spatial perceptual inference. They are adaptive predictive encodings that are driven by embodied top-down expectations and minimize prediction errors. Attention is responsible for the selection and binding of visual features, whereas visuo-spatial perspective-taking is responsible for transformations of visual observations onto self-centered views. These adaptive mental processes are continuous in time, and support the visual recognition of actions, such that the correspondence to other action-related modalities can be inferred by principles of embodied simulation and predictive coding.

Chapter 4

Modeling Action Understanding

Cognitive models are computational approximations of mental processes. The purpose of cognitive models is complementary to neuroscientific and psychological studies – which basically *measure* neural and behavioral systematics and perceptual phenomenology – to *emulate* the same. Cognitive models often constrain themselves to specific functional paradigms that are believed to represent building blocks of cognition – such as artificial neurons or related information coding theorems – to investigate if these paradigms allow the approximation of more complex cognitive phenomena. This approach allows a substantiation and verification of hypotheses that often come from neuroscience and psychology while conclusions about the functional concepts of cognition can be drawn.

Computational models thus offer the potential to investigate the feasibility of specific assumptions and theoretical approaches. Unfortunately, neural network models of action understanding are still rather hard to find. In the following, I catalogue own preliminary studies on this subject that implement parts of the theoretical model proposed in Chapter 3, before I compare this approach to related and relevant work of others.

4.1 Preliminary Studies

Aspects of the hypothetical model of action understanding described in Chapter 3 were implemented in a number of previous own studies. First, we implemented a neural model that learns motion patterns from visuo-proprioceptive perceptions of a simple, artificial 2D arm (Schrodt et al., 2014a). We showed that the model is able to transform the observed biological motion gradually to one of multiple canonical reference frames seen during training by minimizing top-down expectation errors – basically taking the perspective seen during training. Minimal information such as the constellation of motion directions of joint positions was sufficient in this process.

Our follow-up study showed that the model also scales on a simulated full body in 3D (Schrodt et al., 2014b). By including top-down modulations in the learning algorithm, clearly distinct, separately tuned cells developed (cf. Layher et al., 2014) for specific canonical perspectives, including the perspective that represents an egocentric frame of reference. The study also revealed the model's dependency of the recognition performance on display orientation comparable to psychometric studies (Pavlova and Sokolov, 2000).

Further investigations showed that the type of information the biological motion patterns are based on provides the model version with complete invariance to differences in body morphology, movement speed, and scale of the observation, as well as robustness to differences in posture control. Also, the developing encodings were found suitable for learning untimed forecasts of the sequence of the recognized motion patterns, which further contributed to the robustness of perspective-taking (Schrodt and Butz, 2014).

The evaluations were confirmed and expanded in a further study (Schrodt et al., 2015), where we applied the model for learning multiple perspectives on different types of actions which came from full body motion tracking recordings of human subjects. The patterns generated by the unsupervised learning algorithm were found to be clearly classifiable with respect to their action type based on the pre-structured neural extraction of directional motion information. While the foregoing model variants did not show how proprioceptions could be inferred from visual observations – although they could establish the visual correspondence – the present study showed how this inference is possible, and that it is strongly entangled with the ability to adopt the visual perspective of the observed person. Furthermore, the model replicated bistable percepts when depth information was suppressed.

While all of the above model variants were based on winner-takes-all cells (cf. Grossberg, 1973) with direct multimodal tunings, we extended the model to encode visual, proprioceptive and also abstract intention signals in *separate modules* (Schrodt and Butz, 2016). Although the introduced modules cannot *directly* be compared to parietal, temporal and frontal regions of the mirror neuron system, the model proposed a step towards a distributed representation of similar encodings. In this work, we showed how separate modal modules can simulate sequences of action patterns while mutually predicting and synchronizing each other. By linking the developing motion patterns to visual snapshots of the perceived biological motion, the model obtained the capability to *imagine* specific action types without sensory input. In a similar manner, the distributed simulation allowed to infer action classes and proprioceptions from observed stimuli.

In Chapter 5, these model variants are unified and their capabilities are expanded. In particular, I add several submodal encodings to the visual and proprioceptive modules to represent different types of information, which then feature different types of perceptual invariance, and thus different types of predictive information. This step is crucial for solving the binding problem – which has been side-stepped in earlier variants of the model – and helpful for solving the correspondence problem. As in earlier model variants, the types of information are encoded by specifically tuned populations of cells. This population coding is extended to support learning and to enable feature binding. As well, the encoding capacity of submodal neural modules is enhanced by learning locally distributed codes by means of cooperating cells, instead of winner-takes-all cells. Previously, predictions of the progress of simulated or imagined motion were based on approximations of Bayesian statistics. The variant of the network introduced in this thesis is expanded to identify more complex, multi-conditional, spatio-temporal dependencies from distributed submodal encodings for its predictions.

4.2 Related Modeling Approaches

To the best of my knowledge, there is no model of action understanding that integrates and tests the proposed concepts. There is also a particular lack of learning, iterative and adaptive models of spatial perspective-taking. Hence, in the following, I present relevant studies that relate to the approach of this work in partial aspects.

Intention Inference Models

Many models that refer to action understanding attend to the inference of future goal states of an observed individual – such as an effector position. As an example, Baker, Saxe, and Tenenbaum, 2006; Baker, Saxe, and Tenenbaum, 2009 describe action understanding and intentional reasoning in a Bayesian framework as inverting a probabilistic generative model from hidden intention states to actions under the assumption of rational behavior. The framework is validated in a simple scenario, in which the model infers the likelihood of goal positions of an observed agent based on movement trajectories. The results are compared to inferences made by humans. Closer related to this thesis, Bütepage, Kjellström, and Kragic, 2017 use a conditional variational autoencoder to simulate forward in time multiple possible upper-body biological motion trajectories that end up in multiple possible goal states.

However, as explained earlier, this thesis considers the inference of simplified kinematic intention states (such as walking) instead of finite modal goal states. The main focus here is on the embodiment, visuomotor interactions and perceptual inferences that could also yield inferences and simulations of more complex hidden variables (like emotional states). Similarly, Friston, Mattout, and Kilner, 2011 propose a model of action understanding that infers compact action codes (hidden states) and proprioceptions (joint angles) from visual trajectories (coordinates) via active inference. Again, the theoretical framework was validated in a simple case (handwriting of a simulated arm with two degrees of freedom), in which however no modal representations of the considered kinematic trajectories were learned. For a short but general overview of models that predict intention-like states in robotics applications and multi-agent systems, refer to Demiris, 2007.

None of the studies mentioned here considers the perspectival differences between executed and observed actions, which is investigated by models of perspective-taking.

Perspective-Taking Models

Johnson and Demiris, 2005 propose a framework that demonstrates the positive effects of perspective-taking on action recognition, and that involves feed-forward and multiple generative visual models to simulate a robot actor's visual perception. Still, despite the availability of prediction errors in this model, the concrete realization of perspective-taking directly applies a transformation to the visual inputs that stems from the gaze direction of the observed actor, which was basically made available to the model.

Similarly, Breazeal et al., 2006 propose a robot architecture that simulates the perspective and belief system of a teacher to resolve ambiguities in the teacher's action demonstrations on a symbolic level. Again, perspective-taking is based purely on pre-defined perceptual cues (a body tracking algorithm) and does not apply learned, embodied expectations. Another cognitive architecture for human-robot interaction that involves perspective-taking by means of simulating the field of an actor's vision via bottom-up information was implemented by Trafton et al., 2005.

Ehrenfeld, Herbort, and Butz, 2013 propose a model that employs a kinematic, modular body scheme for transformations between global and local frames of reference and fusion of sensory information. The model was successfully tested in a visuo-spatial perspective-taking task (Ehrenfeld and Butz, 2014). It does not learn and predict actions and body motion, but uses rather explicitly defined forward and inverse kinematics and sensor fusions to this end. It is also provided with observed feature identities, and their positions and orientations in a number of local frames of reference, such that the perspective estimate is again a result of direct transformations and sensor fusions.

In fact, besides own work mentioned in the last chapter, the functional principles of perspective-taking are strongly simplified, pre-defined, or completely bypassed also in most models of imitation (see e.g. Billard and Matarić, 2001; Cabido-Lopes and Santos-Victor, 2003).

Taken together, although some algorithmic approaches on perspective-taking and imitation exist, neither of these models considers the learning and application of driving signals by which a mental perspective may be adopted, the fundamental neurocomputational mechanisms, and the seemingly continuous, gradual nature of perspective-taking and mental rotation. Computational models that reproduce perceptual characteristics of psychometric studies are hard to find.

Pose Estimation Models

The approach of selecting and sorting visual features to estimate bodily configurations from visual features suggested in this thesis can be compared to algorithms of articulated pose estimation. Most approaches are based on tree-structured graphical models that enforce spatial consistency of the estimated pose (e.g. Felzenszwalb and Huttenlocher, 2005; Andriluka, Roth, and Schiele, 2009; Andriluka, Roth, and Schiele, 2010; Yang and Ramanan, 2011). For example, Andriluka, Roth, and Schiele, 2010 propose a three-stage graphical model that estimates 2D limb positions (based on Andriluka, Roth, and Schiele, 2009), integrates them over time, and reconstructs 3D pose estimates from image sequences. For robustness to differences in perspective, multiple, individual view-point specific pose estimators were trained, whose output was then combined via a support-vector machine.

Tree-structured graphical models for pose estimation are typically confronted with problems that result from symmetries in the appearance, or self-occlusions of body parts. These problems can be avoided by learning strong postural priors from usual activities such as walking (Lan and Huttenlocher, 2005), or by applying occlusion-sensitive local likelihoods (Sigal and Black, 2006). In contrast to graphical models, where dependencies between the variables of the bodily configuration are typically not learned, newer approaches rely on hierarchical, multi-scale filters or classifiers which are trained via backpropagation.

For example, Ramakrishna et al., 2014 propose an inference machine for pose estimation, where each of multiple classifiers on a hierarchical level generates a confidence map for the location of each anatomical landmark. These estimates are combined and gradually refined on the hierarchy to estimate 2D body postures from still images. Similarly, Wei et al., 2016 estimate poses using a deep convolutional neural network. Both approaches were trained on rather large amounts of annotated data to develop the inferential capabilities and robustness to different perspectives, and they do not apply spatial transformations for perspective invariance.

The approach presented in this thesis combines several methods also used for pose estimation as mentioned above, although some of the problems that often arise from raw image processing are basically skipped, such as the extraction of visual candidate features for anatomical landmarks (although it can distinguish between biological and non-biological features), tracking their location over time, or the reconstruction of depth information. The model learns strong spatial and structural priors from bodily, self-centered actions. It uses these embodied priors to gradually transform the coordinate space of observed visual landmarks to infer a corresponding view-point of the respective action. Thus, the model does not have to learn different views of the same actions, although it is potentially capable of encoding multiple, canonical views to speed up convergence (Schrodt et al., 2015). Furthermore, the model uses multiple types of information for pose inferences, which improves the inference abilities, the ability to preserve the inferred feature assignments and perspective over time, as well as robustness and generality in action recognition. Determining visual feature identity and binding, the perspectives, and the resulting posture are separate processes (or stages) in the model proposed here, providing further robustness.

Generative Action Models

A number of generative models capable of crossmodal inferences and simulations have been proposed. For example, Lallee and Dominey, 2013 implemented a model that integrates low-level sensory data of an iCub robot, encoding multimodal contingencies in a single, 3D, self-organizing competitive map. When driven by a single modal stimulus, the multimodal integration enabled mental imagery of corresponding perceptions in other modalities. The modeled self-organizing map is topographic with respect to its discrete multimodal cell tunings. However, temporal dependencies and action patterns are not encoded in this approach.

Taylor, Hinton, and Roweis, 2006 implemented a stochastic, generative neural network model based on conditional restricted Boltzmann machines. When trained on motion capture data, the model is able to reproduce walking and running movements as well as transitions between them in terms of sequences of joint angular postures. However, the model does not learn and generate sequences of other modal perceptions, and does not implement modal or perceptual inferences.

In contrast to the mentioned models, the model proposed in this thesis is generative in several modalities and submodalities, which are able to produce consistent simulations based on predictive crossmodal encodings. Furthermore, the model is able to learn compressed kinematic types from the submodal encodings, which can be used for inference from observations or selective action simulations.

Biological Motion Recognition Models

Although complex models of action understanding are rare, there are several noteworthy models of biological motion and Gestalt perception, which constitutes an important aspect of it. For example, Fleischer et al., 2013 modeled the properties of STS cells during object interaction. The authors' approach includes the encoding of multiple viewer-centered representations of simple, schematic actions to establish a certain degree of orientation invariance. recognition was based on a hierarchy of feature detectors in several neurobiologically inspired domains, like local shape detectors and motion neurons, leading to plausible model predictions about human recognition performance. However, the model does not infer perspectives, but uses separate networks for each encoded viewpoint. Furthermore, the model uses a hard-coded wiring and parameterization that is not trained on data.

Lange and Lappe, 2006; Lange, Georg, and Lappe, 2006 modeled biological motion recognition using viewer-centered, image-based posture templates, where the best matching template responses were integrated over time and decided on the recognized movement. The authors' model is timescale-independent to a certain degree and can distinguish pre-defined walking directions. The approach also produced plausible results with respect to the artificial cell firing rates. Even so, the motion information is only considered indirectly by recognizing whole movements by means of adjacent posture images, and the model was validated only on a single movement in two manually distinguished orientations. Again, no learning was applied to the model's parameters. The approaches mentioned above can be compared to models of Gestalt perception. For a general overview of Gestalt perception models, see (Jäkel et al., 2016).

Taken together, none of the models introduced in this chapter solves the correspondence and binding problems via expectation-driven, iterative, inferential processes, and only a few assess the issue of embodiment in action understanding. The model introduced in this theses combines and unifies several of the above aspects for a neurocomputational model of the principles of action understanding, which is detailed in the following.

Chapter 5

A Neural Network Model of Action Understanding

This chapter proposes a model of action understanding that implements the addressed interpretation of embodied simulation, predictive coding, perceptual inferences, and blends into neuroscientific and psychological findings about the mirror neuron system. The model consists of three main modules: An intention module, a visual module, and a proprioceptive module. All modules communicate via bidirectional couplings that represent complex neural circuits with predictive characteristics.

Note that the three modules can not directly and not exclusively be referred to frontal, temporal and parietal regions of the mirror neuron system. Rather, the visual module infers both spatial relations of visual features (as in parietal cortex), as well as Gestalt templates (as in temporal cortex). It also binds visual features and performs spatial perspective-taking, which can be linked to the parietal cortex and the dorsal stream of visual processing. Thus, the model also reflects shared perceptual mechanisms in the respective parietal and temporal regions. Similarly, the proprioceptive module does not directly relate to inferior parietal regions which integrate sensorimotor information. Rather, this integration can be seen in the bidirectional, predictive connectivity between the visual and proprioceptive modules. I chose the term "proprioception" in the following for all motor-related encodings of the mirror neuron system because the model is purely perceptual and does not act per se. Thus, it will infer proprioceptive perceptions from observations instead of corresponding motor commands, which is however technically comparable. Finally, the intention module is somewhat related to frontal encodings and kinematic intentions, but the most abstracted part of the model.

In the following chapters, I will first give an overview of the functionality of the model in Section 5.1, clarify important modeling assumptions and notations in Section 5.2, before the basic information processing of visual and proprioceptive stimuli is explained in Section 5.3. In Section 5.4, I will explain how predictive encodings are learned from the processed modal information types, and finally, in Section 5.5, I will explain the inference of action classes from these encodings.

5.1 Technical Model Overview

The action understanding model consists of three modules, each representing an abstraction of action relevant modal or amodal neural activity, and interactions thereof. An overview of the modules, network components and the respective processing steps is shown in Figure 5.1.

(i) The *vision module* processes a number of visually observed, salient features. Each feature is represented by a Cartesian coordinate relative to a global 6 dimensional frame of reference, consisting of origin and orientation. Features may either correspond to joint locations – representing bodily landmarks – or distractors that do not represent bodily characteristics (cf. Section 2.2.1).

As mentioned earlier, the embodied learning approach assumes that visual information about own bodily actions is observed and learned from a self-perceptual perspective (e.g. head centered) and then correlated to other modal or amodal codes. Thus, when optical action patterns of another person are observed, the model has to infer the frame of reference the action is currently observed from to establish the visual correspondence to the learned embodied codes, including action codes in other modalities. This operation is performed in the *perspective-taking* step, which incrementally adapts to the differences between the self-centered frame of reference, and the frame of reference the action is perceived from, driven by top-down expectations that come from the currently activated embodied codes. Perspective-taking thus can be compared to mental imagery of spatial relations, which solves the correspondence problem in the model.

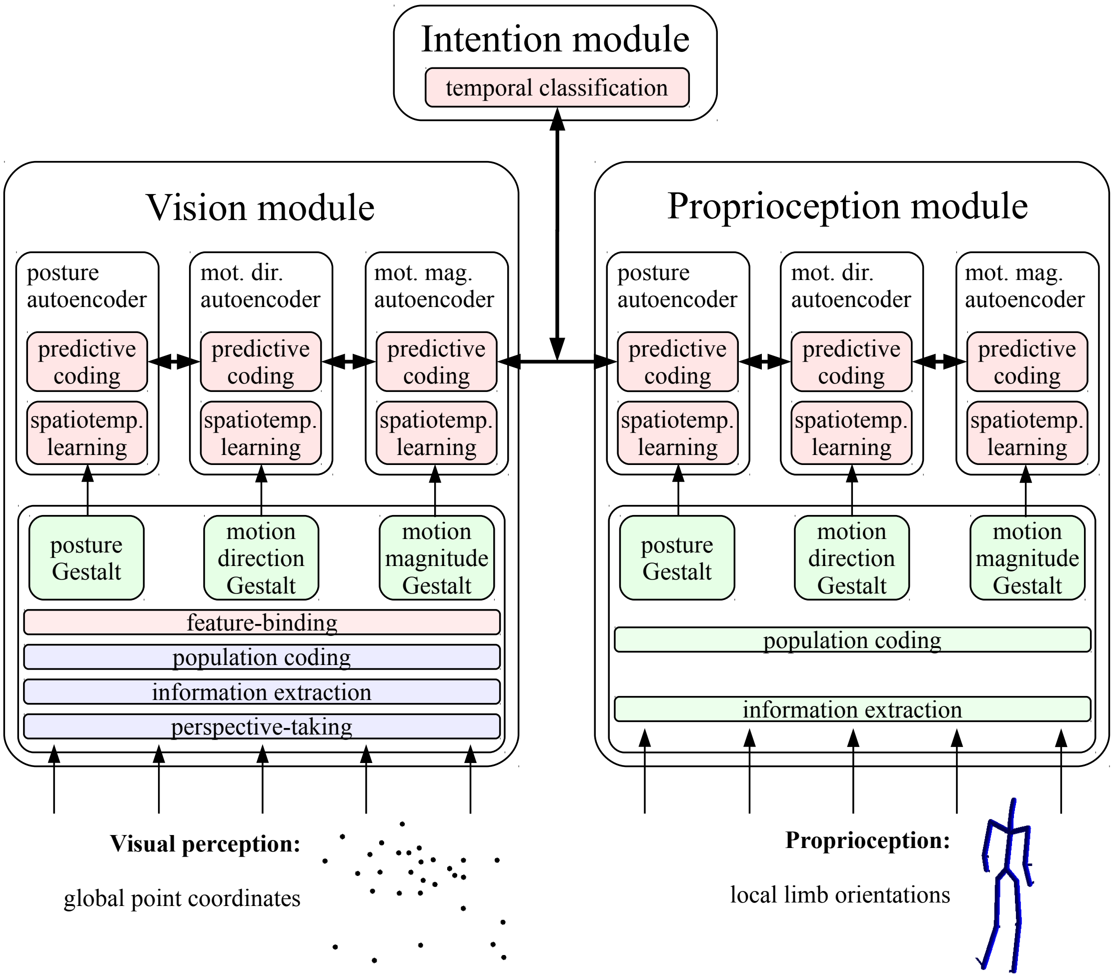


FIGURE 5.1: Overview of the modules of the model and their connectivity. Visual global feature coordinates and proprioceptive limb orientations are processed in two separate modules. The modules extract posture, motion direction, and motion magnitude configurations, while visual perception is supported and adapted by top-down driven perspective-taking and feature binding steps. The submodal configurations or Gestalt percepts are then learned by separate autoencoders, which learn spatio-temporal codes and apply them for predictive coding. The learned codes are classified in an intention module, which is able to top-down bias the predictive encodings. The three main neural modules are able to predict their own activated, compressed codes as well as the codes of other modules. Here, blue boxes represent processing steps applied in the same manner to each *observed* feature. Green boxes represent processing steps applied in the same manner to *bodily* features only. Red boxes represent processing steps where each (either observed or bodily) feature is treated individually.

Subsequently, submodal information of each input feature is extracted and encoded separately by populations of locally receptive cells, indicated by the *information extraction* and *population coding* steps. The submodalities comprise the currently inferred (or spatially imagined) position, motion direction, and motion magnitude of the respective input stimuli. First, subdividing fundamental types of information from the input stimuli in this systematic manner is crucial for the perceptual adaptation processes in the model, as will be explained later and shown in the experiments. Second, encoding the respective types of information in locally tuned cells approximates a normalization of neural activity for each feature, which on the one hand prepares for the subsequent identification of relevant visual features, and on the other hand facilitates equitable learning of compressed, submodal codes.

The feature binding submodule in the visual pathway performs the *selection of action relevant features* out of the set of *all* observed features, as well as the *assignment of the selected features* to the correct neural processing paths. This essentially solves the binding problem in the model: I assume that – during embodied training – each salient visual input stimulus corresponds to a specific joint, and that this distinct assignment from input indices to joint indices is provided. During testing, however, the assignment is not provided, which is why the model is to infer the correct assignment, again by applying top-down expectations. As well, there may be distracting features even with similar motion dynamics, which do not correspond to any learned bodily landmark. Thus, the model is to select the action relevant information as well. Summing up, selection and assignment are identical to an identification of body joints from a set of moving visual features, which humans are capable of in experiments with point light displays (Johansson, 1973; Pavlova, 2012).

In this model, the binding problem is tackled by considering multiple hypothetical associations for each input feature, applying top-down expectations about the observed biological motion dynamics in terms of submodal neural activity to gradually infer and associate the relevant dynamical patterns. Technically, this is realized by a gated connectivity from each *salient* feature population to a number of *bodily* feature populations that are capable of representing multiple, potentially matching stimuli in parallel. Each of those connections – e.g. from all three submodal populations of a salient feature indexed by 1 to the three submodal populations of bodily feature 3, representing the right elbow – is gated by an adaptive selector neuron that represents a non-linear assignment strength. The network will incrementally reinforce the gates that match the locally expected dynamics, while diminishing the connections that associate divergent dynamics. In case there are more salient stimuli than learned bodily features, the connectivity represents both a selection of relevant features as well as an assignment to the correct bodily features.

Taken together, the first two submodules of the visual processing pathway consist of pre-structured, bottom-up, perceptual processing, which applies specific adaptations online to match top-down expectations – resulting in spatial predictive encodings. The expectations are generated by three temporal conditional autoencoders, each combining a *submodality* of the transformed, selected and ordered bodily features into a *whole-body Gestalt percept*. The autoencoders first learn compressed codes without considering temporal or crossmodal dependencies to cover the contingencies of submodal self-perceptions, which essentially models embodied learning of submodal spatial codes. The learned and recognized codes are then used to learn predictions of succeeding of codes. These predictions are learned by identifying temporal features in a short logarithmic history of several recently activated submodal codes, with the objective of minimizing the divergence to the actual observed codes. The identification of suitable temporal features is not bootstrapped by or restricted to specific dependencies. In fact, the predictions may conditionally and temporally depend on the previously activated codes in their own, submodal domain, as well as on all codes of other autoencoders. Taking the proprioceptive codes that are learned in a similar manner into account, this essentially models modal and crossmodal inference processes. For example, the current location of a joint together with the motion direction and magnitude – each represented by a separate, compressed code – are in combination suitable to predict the next position of the respective joint. Conversely, specific positions may be suitable to predict a change in the direction of motion, for example when a specific extreme posture is reached. In sum, the model is expected to learn multi-conditional, temporal dependencies that develop from distributed, compressed, submodal, spatial codes.

Using autoencoders, the bottom-up activated submodal codes can be back-projected onto the corresponding stimuli to obtain biased reconstructions of

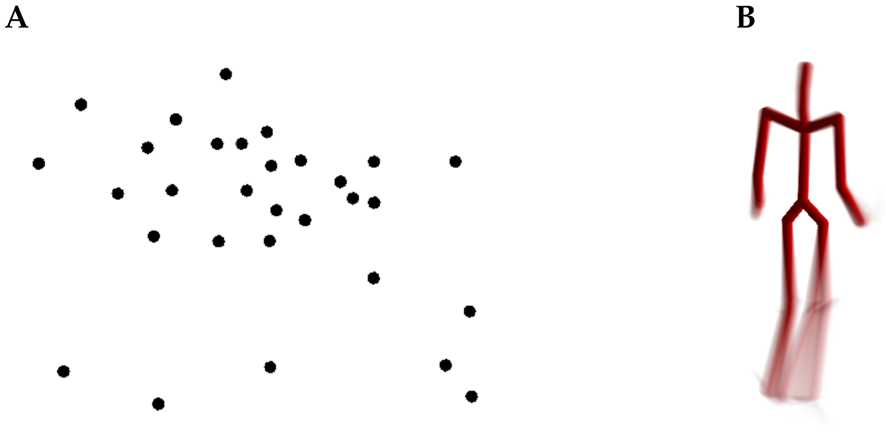


FIGURE 5.2: A point light stimulus (**A**) that pushes the model's internal body scheme towards resonance (**B**), after perspective-taking and feature-binding were successful.

Gestalt perceptions. As a consequence, when a fully trained autoencoder is provided with an imperfect stimulus (e.g. an observed action with unknown identity of the features, shown from an unknown perspective), it is able to infer an expected stimulus that is biased by its embodied, or self-perceptual experience. Although the expectation is distorted as well when a distorted stimulus is presented, it is typically closer to the actual, non-distorted stimulus given that it was encoded during training. Then, adapting the bottom-up processing of the stimulus given the biased reconstruction as target, the perception in terms of perspective, feature selection and assignment is typically pushed towards correspondence, compensating differences in self-perception and observation. Improving the perceptual processing also improves the expectation by itself, such that the respective perceptual parameters are adapted along a path that typically leads to inference.

Similar to bottom-up activated codes, the autoencoders can infer the currently expected actual Gestalt stimulus given a *predicted* code. In case there is no input provided to the overall network, this can be considered a global, modality comprehensive imagination of learned action patterns. In case that only one of the modalities is provided, the code prediction of the other modality is comparable to a simulation of the expected stimulus. Notably, when proprioceptive activations are predicted from visual stimuli (see Figure 5.2), it can in fact be said that the motor system resonates in response to the observation: Since the simulation of submodal encodings is determined both by local, submodal temporal dependencies as well as lateral, crossmodal dependencies, the submodal simulations are partially self-sustaining, and partially striving for global consistency. As a result, the network is able to infer proprioceptive stimuli via (even highly distorted) visual stimuli after a short period of spatial and temporal adjustment, which enables it to 'understand' observed actions in terms of corresponding proprioceptive sensations, or even to simulate and imagine whole body actions in multiple modalities.

In contrast to the visual pathway, proprioceptive stimuli are not encoded by means of global coordinates of joints or distractors. Each proprioceptive input represents the orientation of a specific *limb*, relative to the orientation of the predecessor in the body structure hierarchy (cf. Section 2.2.2). As such, proprioceptive processing assumes knowledge about the hierarchy and length of the limbs. Proprioceptive postural codes thus encode whole-body limb direction constellations in local orientations, instead of positional constellations in a global frame of reference as in the visual pathway. As a result, when only a single limb is moved, no motion is perceived in the successors in the body hierarchy. Analogously to the visual processing pathway, the proprioceptive pathway extracts submodal information from the bodily input features, encodes them separately in populations of locally receptive cells, and develops codes and predictions from the spatial and temporal contingencies. Proprioceptive sensations, however, do not require perceptual inference or spatial imagination by nature, because they cannot be observed from others. Hence, the perspective-taking, feature selection and assignment steps are left out.

All of the six developing modal and submodal encodings – posture, motion direction and motion magnitude of both visual and proprioceptive autoencoders – are further processed by the *intention module*. In the presented setup of the model, the purpose of this module is to identify the type of action that is currently observed, which can be related to kinematic intentions such as walking on a floor or executing a basketball dribble. Technically, the module is supervised to learn distinctions of movements using the constellation of developing visuo-proprioceptive codes, applying a temporal classifier network. The output of this module is an activity vector that represents a confidence value for each of the action classes shown during training. The class output is recurrently connected to all visual and proprioceptive autoencoders and *biases* the temporal prediction of their respective state progress. This enables the autoencoders to converge towards different temporal attractor states during prediction, based on the class currently presented. Thus, given a motion intention, visual and proprioceptive code simulations are pulled into consistent temporal attractors that represent the movement via a top-down influence, while synchronizing each other via the lateral, crossmodal dependencies. As well, when the class is inferred from a (visually) provided input, it allows to stabilize the simulation of corresponding inputs (such as proprioceptions). In context of the mirror neuron system, the model assumes that also a coupling between visual and intentional codes exists (albeit rather indirect) and is also used to enhance inference and embodied simulations.

In the following, all of the mentioned modules and processing steps are detailed including mathematical formulations. At first, however, the modeling paradigms are described as well as the notation of the formalization.

5.2 Methodological Subsumption and Formal Preamble

The neural network model processes its inputs in a specific, pre-structured manner. It implements *learning and structural biases* by means of different modules with different purposeful processing steps, in contrast to most deep neural networks that have a repetitive structure. Some of the involved neural transformations and calculations are *fixed*, some of them are *learned* in a training phase, and some of them are *adapted* in a testing phase. All of the processing steps in the network are performed by artificial *neurons* that are locally connected and communicate via artificial synapses or *weights*. Artificial neural activity models the average firing rate of cells in terms of second generation neural networks (Maass, 1997). Following a strictly *connectionist* approach (Elman, 1998), each of the artificial neurons and weights communicates with connected components only, through defined pathways only.

Furthermore, all network components are *differentiable* with respect to multiple error signals that are propagated backwards along distinct processing paths of the network. This allows for parameter optimization by *gradient descent* on convex error functions (Rumelhart, Hinton, and Williams, 1986). The implementation also considers time delays and recurrences in signal propagation by *backpropagation through time* (Werbos, 1988; Mozer, 1989), and limits the activation functions' minimum derivatives with respect to their input (cf. flat-spot elimination Fahlman, 1988). Hence, the artificial neural network models bottom-up processing by means of stimulus propagation, and top-down processing by means of both error backpropagation as well as recurrent connectivity. Backpropagated errors, however, are (with one exception) not provided, such that the network can be considered to be unsupervised, or, more precisely, *self-supervised*, because it generates its own error signals. Besides the recurrent nature of the network, the propagation of signals involves *axonic modulations* as well as multiplicative, *presynaptic interactions*: Each synaptic input to a neuron is the weighted product of an arbitrary number of axonic outputs (akin to tensors in gain-field networks, cf. Andersen, Essick, and Siegel, 1985; Salinas and Abbott, 2001), while axonic outputs can be modulated by one another via weights as well (comparable to shunting inhibition Eccles, 2013; Blomfield, 1974).

All learning and adaptation is based on the same principle, that is, by *minimizing backpropagated, self-generated prediction errors* via gradient descent. Learning here means encoding specific compressed action patterns given input stimuli by permanently changing weights, while adaptation means altering the processing of inputs in a way such that they consistently match the encoded patterns by temporarily changing weights. Adaptations are restricted to specific transformations, such that the network effectively cannot be subject to self-deception. Further, all learning and adaptation is accomplished *online*, meaning that the sequence of inputs is coherent in time. No batch learning, randomized input sampling, or comparable training methods are applied. While this supports cognitive plausibility, it also makes the network principally prone to the problem of *catastrophic forgetting* (McCloskey and Cohen, 1989). By selecting appropriate coding schemes and learning principles, this problem, also referred to as *recoding problem* in the following, can be effectively avoided, as will be described later.

With this in mind, and with reference to the differentiability of all of the applied components, a mathematical notation is used in the following that focuses on the *results* of the respective neural processing steps rather than the neural connectionism and the transfer functions chosen for the respective components. Error signals, parameter adaptations and learning rules are highlighted specifically. The connectivity, on the other hand, is indicated in distinctive figures, and explained qualitatively. In the following, all processing steps and the corresponding neural network structures are visualized in connectivity diagrams. A legend for these diagrams is presented in Figure 5.3.

Mathematical variables are notated as follows: Superscripts generally annotate which *submodality*, which *modality*, or which *adaptive, learning, or generative process* the variable refers to, and combinations thereof. When superscripts denote exponents instead, this is emphasized explicitly. Thus, superscripts can indicate that a variable exists in multiple contexts. In case that there is only a single, obvious context for a variable, the superscript is dropped. Subscripts generally describe indices either referring to a particular *observed feature*, *bodily feature*, *spatial dimension* or *neuron* in a layer or population. In case there are multiple indices that refer to the same type of component, the order of subscripts is is equivalent to the direction of signal processing, meaning for example that the first index describes where activity originates from, while the second index describes where the activity is directed towards. Vectors and scalars are indicated by lower case letters, while vectors are bold. Matrices and discrete constants are indicated by upper case letters. Related variables may have the same variable character but are distinguished by an indicative, transcribed symbol. For example, a variable transcribed with a dot may indicate a different encoding format, while a tilde may indicate a top-down expectation referring to the variable without tilde.

Thus, variables generally come in the following format:

$$\overset{<\text{indicative symbol}>}{[\text{variable character}]}^{<\text{submodality}>|<\text{modality}>|<\text{reference to process}>}_{<\text{observed feature}>|<\text{bodily feature}>|<\text{neuron/dimension}>} \tag{5.1}$$

where [ ] denotes mandatory characters, $<\ >$ denotes optional characters, and $|$ denotes non-exclusive logical disjunctions.

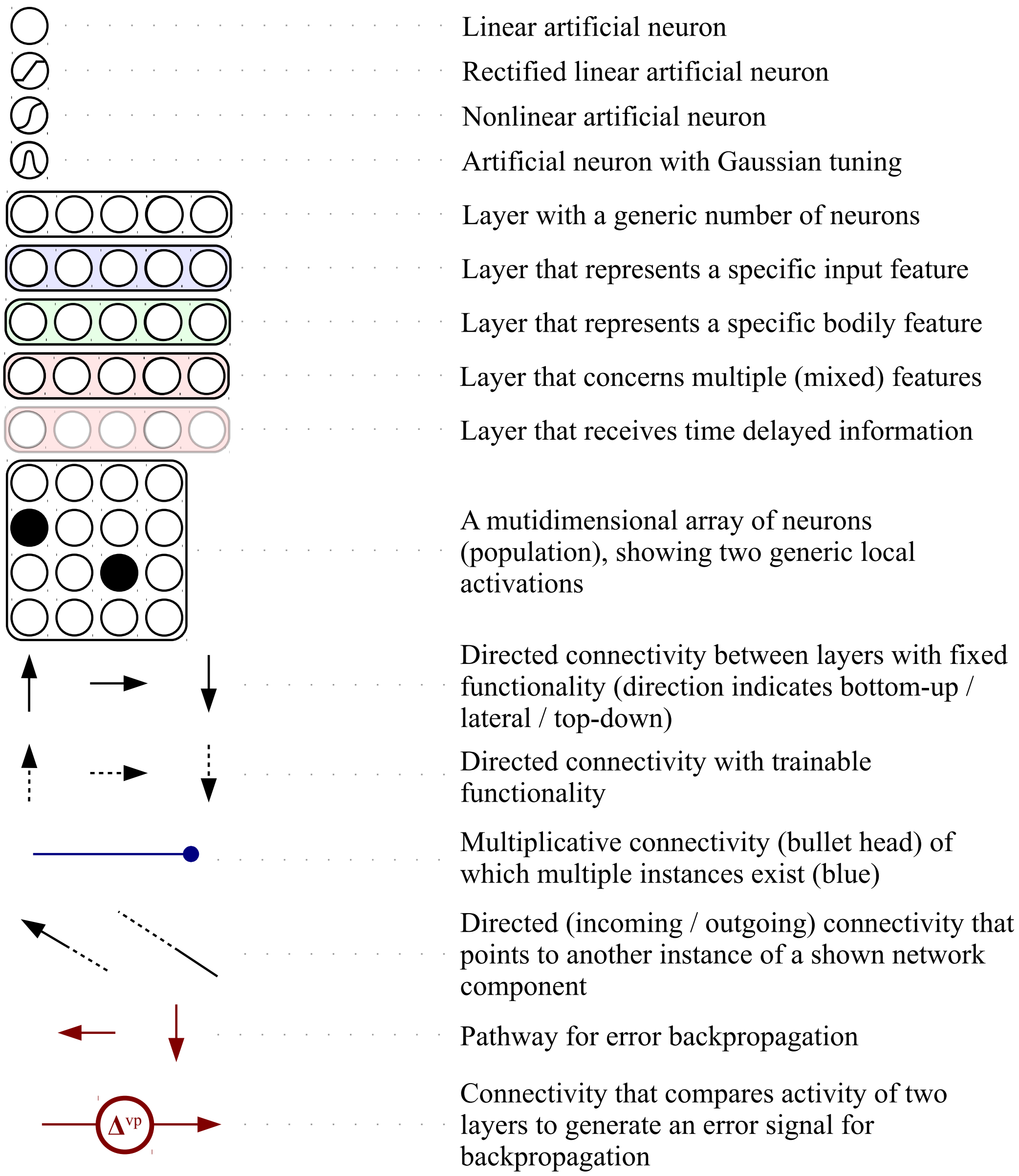

Linear artificial neuron
Rectified linear artificial neuron
Nonlinear artificial neuron
Artificial neuron with Gaussian tuning
Layer with a generic number of neurons
Layer that represents a specific input feature
Layer that represents a specific bodily feature
Layer that concerns multiple (mixed) features
Layer that receives time delayed information
A mutidimensional array of neurons (population), showing two generic local activations
Directed connectivity between layers with fixed functionality (direction indicates bottom-up / lateral / top-down)
Directed connectivity with trainable functionality
Multiplicative connectivity (bullet head) of which multiple instances exist (blue)
Directed (incoming / outgoing) connectivity that points to another instance of a shown network component
Pathway for error backpropagation
Δvp
Connectivity that compares activity of two layers to generate an error signal for backpropagation

FIGURE 5.3: Legend for neural connectivity diagrams. Blue and green layers, as well as their connectivity, exist multiple times (once per observed or bodily feature), although this is typically not shown in the connectivity graphs. Red layers exist only once for the whole (sub)modal processing path.

For example, $\dot{m}^v_{ai}$ denotes the a-th component of the motion magnitude (m) of the i-th visual (v) feature, encoded by a population of neurons ($\dot{\ }$). As another example, g^{vp} denotes the whole body Gestalt perception (g) of a visual (v) posture (p), while $\tilde{\mathrm{g}}^{vp}$ denotes the according expectation generated by the model. An X in a superscript denotes a set that defines in which modalities and submodalities the variable exists (e.g. $\mathbf{c}^X, X \in \{vp, vd, vm, pp, pd, pm\}$), while triple points in subscripts are discrete set-builder notations (e.g. $a_{1\ldots3} = \{a_1, a_2, a_3\}$).

Using this notation, the processing of input features, the extraction of submodal features, and the adaptation of visual perception are outlined in the following.

5.3 Information Processing and Perceptual Inference: Visual and Proprioceptive Pathways

The model processes inputs in two distinct domains that represent different modalities. First, the *visual module* processes a number of N^v point-like, salient visual features in a global frame of reference (cf. Section 2.2.1), and maps them to M^v bodily features. I assume that (i) these saliences are made available to the model such that visual features potentially relevant for action understanding are provided, (ii) *joints* provide the most significant visual features for action understanding, and (iii) the global coordinates of these features – including depth information – can be recognized precisely and coherently from binocular visual streams. Indeed, in my experiments, these inputs originate from motion capture data recorded with multiple cameras, and the resulting stimuli are closely related to experiments with point-light displays introduced by Johansson, 1973. The model is however not limited to the processing of biological stimuli, and can also select subgroups of the provided input coordinates for processing while disregarding irrelevant inputs. Taken together, the input to the visual pathway comes in the form of an arbitrary number of 3D Cartesian *feature coordinates* that may or may not correspond to specific joint coordinates in motion capture data, relative to a *global frame of reference*, which consists of a 3D origin and a 3D orientation. A visual, global coordinate input at time step t is indexed by i and denoted by $\mathbf{x}^v_i$ in the following.

Second, the *proprioceptive module* processes a number of M^p bodily postural features perceived proprioceptively (cf. Section 2.2.2). Here, the model assumes that each of the proprioceptive sensory features (i) is descriptive for a specific *limb*, (ii) represents postural information locally by means of the orientation of the limb relative to its structural predecessor, and (iii) that such features can be extracted reliably from streams of proprioceptive sensations. As a result, a proprioceptive feature comes in the form of a 3D Cartesian unit vector, representing the *limb orientation* in the *local frame of reference* of the predecessor in the body hierarchy. In contrast to the visual features, proprioceptive features do neither encode distances between joints (i.e. limb lengths) directly or indirectly, nor do they encode global coordinates of joints. Rather, proprioceptive features depend on a particular (simplified) body model that provides structural knowledge about the body and its limb configuration. A proprioceptive, local orientation input at time step t is denoted by $\mathbf{x}_i^p$ in the following.

5.3.1 Segregation of Distinct Submodalities

Various types of information feature various types of invariances. For example, the position of a visual feature is not invariant to rotation or translation, while the motion direction of a feature is invariant to translation, but not to rotation, and the magnitude of motion is invariant to both rotation and translation. The model will learn partially invariant codes from these distinct submodal perceptions and use them for perceptual inference. For example, when biological motion is perceived from an unknown viewing angle, it still can identify the constellation of motion magnitudes of the observed points. This can provide a prior expectation about which observed feature corresponds to which bodily feature.

Figure 5.4 shows a connectivity graph for the processing of a single, visual feature. The input at time step t is represented by activity of a layer consisting of three neurons, each directly representing a dimension of the Cartesian input coordinate. Before all other processing steps occur, the network calculates the velocity $\mathbf{v}_i^v$ of each visual feature coordinate:

$$\mathbf{v}_i^v(t) = \mathbf{x}_i^v(t) - \mathbf{x}_i^v(t-1) \ . \tag{5.2}$$

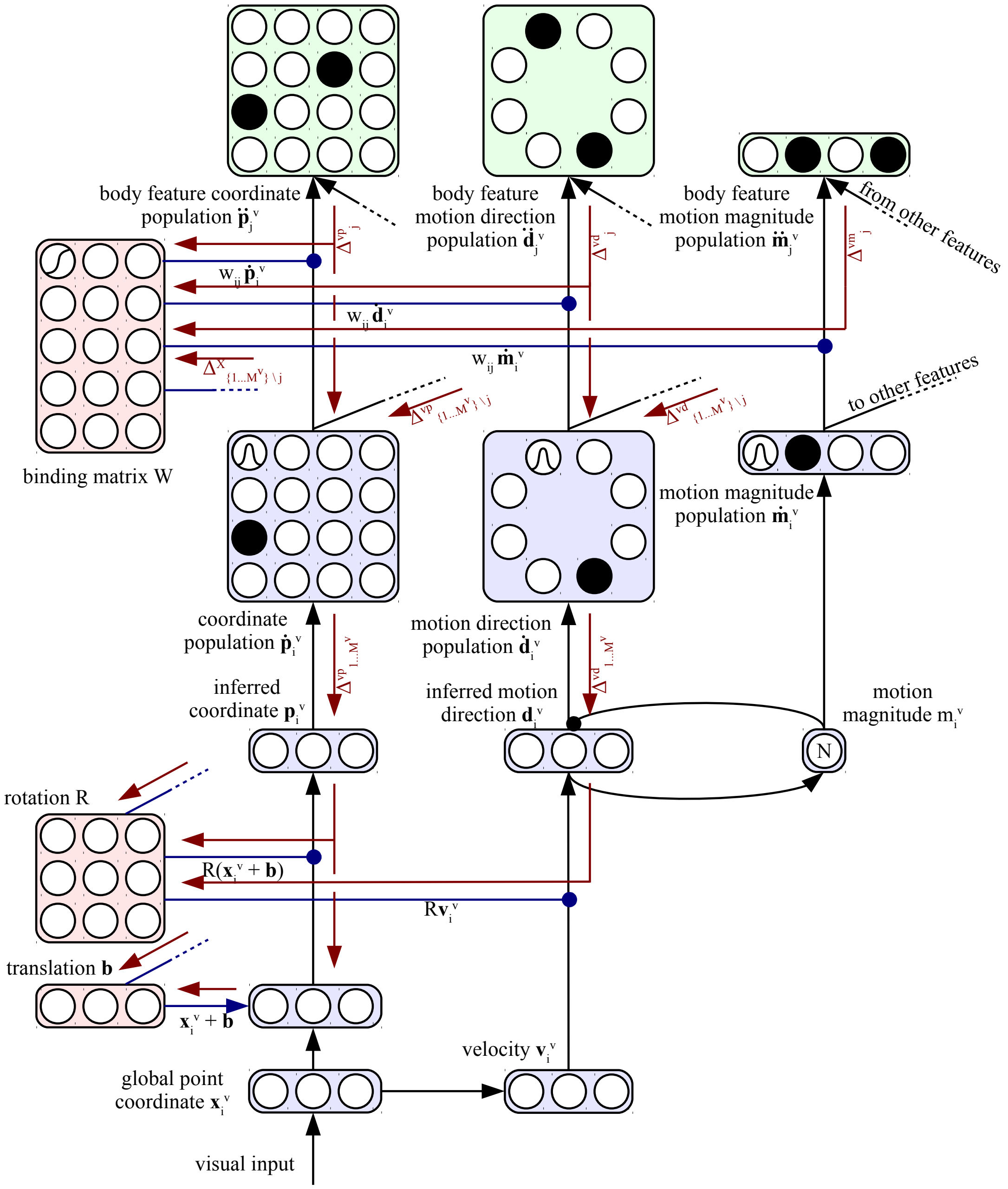

body feature coordinate population $\dot{\mathbf{p}}_j^v$
body feature motion direction population $\dot{\mathbf{d}}_j^v$
body feature motion magnitude population $\dot{\mathbf{m}}_j^v$
from other features
binding matrix W
$w_{ij}\,\dot{\mathbf{p}}_i^v$
$w_{ij}\,\dot{\mathbf{d}}_i^v$
$w_{ij}\,\dot{\mathbf{m}}_i^v$
Δ^{vp}_j
Δ^{vd}_j
Δ^{vm}_j
$\Delta^X_{\{1...M^v\}\setminus j}$
$\Delta^{vp}_{\{1...M^v\}\setminus j}$
$\Delta^{vd}_{\{1...M^v\}\setminus j}$
to other features
coordinate population $\dot{\mathbf{p}}_i^v$
motion direction population $\dot{\mathbf{d}}_i^v$
motion magnitude population $\dot{\mathbf{m}}_i^v$
$\Delta^{vp}_{1...M^v}$
$\Delta^{vd}_{1...M^v}$
inferred coordinate $\mathbf{p}_i^v$
inferred motion direction $\mathbf{d}_i^v$
motion magnitude m_i^v
N
rotation R
$R(\mathbf{x}_i^v + \mathbf{b})$
$R\mathbf{v}_i^v$
translation $\mathbf{b}$
$\mathbf{x}_i^v + \mathbf{b}$
velocity $\mathbf{v}_i^v$
global point coordinate $\mathbf{x}_i^v$
visual input

FIGURE 5.4: Processing cascade from the i-th visual input to the j-th visual bodily feature. Rotation and translation are applied to each feature in the same manner. The feature position, motion direction, and motion magnitude are extracted from each individual input feature, and encoded separately by specially arranged populations of neurons. A neural gating matrix then selects and assigns the observed features to the correct populations that represent *bodily* information. Rotation, translation, selection, and assignment are driven by top-down error signals that stem from self-generated submodal expectations.

From the visual input coordinate and its velocity, three types of submodal information are derived: (i) the *inferred coordinate* $\mathbf{p}_i^v$ of the feature (also referred to as *position* of *posture* of a feature in the following), (ii) the *inferred direction of motion* $\mathbf{d}_i^v$, and (iii) the *magnitude of motion* m_i^v. Two transformations are raised to determine these values: (i) the *rotation matrix* R and the *translation bias* $\mathbf{b}$, both of which are applied to the whole visual percept, that is, all features consistently. The mental rotation matrix models the capability to imagine different perspectives or vantage points on observed visual features, while the translation bias represents the center of the respective coordinate system, which is also the center of rotation. Both of these transformations are applied to the input coordinates, such that

$$\mathbf{p}_i^v(t) = R(t) \cdot \mathbf{x}_i^v(t) + \mathbf{b}(t) \ . \tag{5.3}$$

Analogously, rotation is applied to the calculated velocity of a feature. Since velocities are invariant to translations, the translation bias is not applied in this step. The result of the rotated velocity is subdivided into two components that separate the direction of motion as well as its magnitude, such that

$$m_i^v(t) = \|R(t) \cdot \mathbf{v}_i^v(t)\| \tag{5.4}$$

$$\mathbf{d}_i^v(t) = \frac{R(t) \cdot \mathbf{v}_i^v(t)}{\max\left(m_i^v(t), o\right)} \tag{5.5}$$

where $\|\cdot\|$ is the Euclidean norm. Note that in equation 5.5, there is a point at which the direction of marginal feature motion is not normalized to unit length anymore, defined by the motion direction recognition threshold o. This is to account for the irrelevance of the direction of motion that could be caused by sensor noise.

Proprioceptive features are split up into submodal features in a similar manner, however neglecting the adaptation of perspective, as shown in Figure 5.5. Another difference is that the input coordinates are normalized orientation vectors, which does however not influence the mentioned processing steps.

In sum, the foregone processing steps obtain 3D Cartesian joint coordinates, motion directions, as well as scalar motion magnitudes of the observed features in a specific frame of reference that is determined by the network itself in the

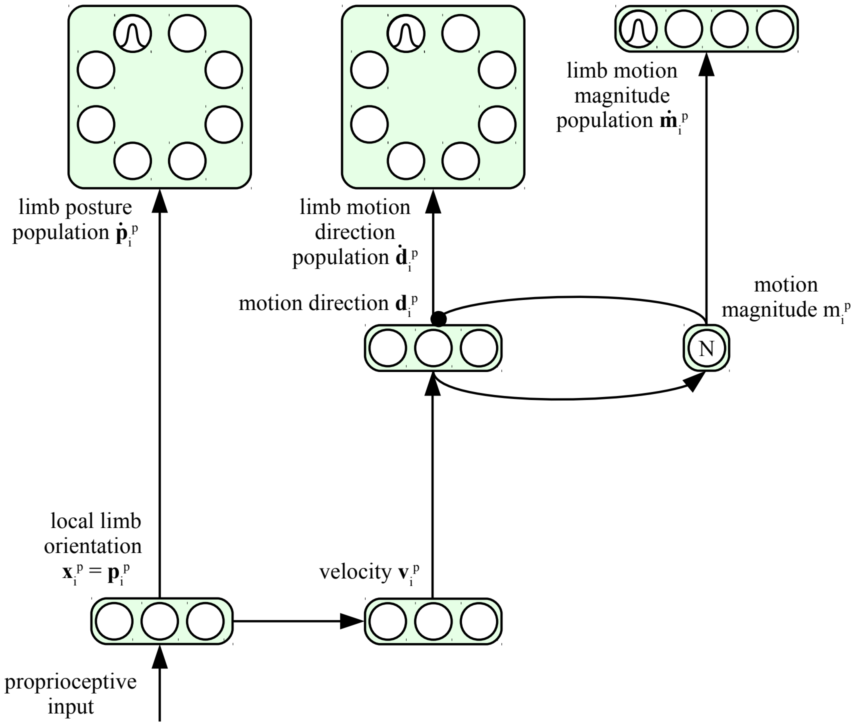

limb posture population $\dot{\mathbf{p}}_i^p$
limb motion direction population $\dot{\mathbf{d}}_i^p$
limb motion magnitude population $\dot{\mathbf{m}}_i^p$
motion direction $\mathbf{d}_i^p$
motion magnitude m_i^p
N
local limb orientation $\mathbf{x}_i^p = \mathbf{p}_i^p$
velocity $\mathbf{v}_i^p$
proprioceptive input

FIGURE 5.5: Processing cascade for a single proprioceptive feature. The information extraction and population coding are comparable to visual features, but stem from limb orientations instead of joint positions. Further, no perspective-taking or feature binding has to be applied, such that no top-down error signals are processed.

visual domain. In the proprioceptive domain, the processing provides 3D joint orientations as well as the motion directions and scalar motion magnitudes. Before I continue to detail further processing steps applied to these types of information, the characteristics of the visual rotation and translation mechanisms are described in the following section.

5.3.2 Perspective-taking for Visual Correspondence

In the model, perspective-taking consists of a translation followed by a rotation of all salient visual features, and aims at establishing the best possible correspondence between the input and top-down expectations. In this sequence of transformations, the translation reflects the origin of the model's internal, imagined frame of reference and also the center of rotation, while the rotation reflects an imagined vantage point.

The translation and center of rotation is determined by the bias neurons $\mathbf{b}$ which can be adapted by gradient descent to minimize top-down error signals in a set E^s, which originate from submodal Gestalt autoencoders as will be described in Section 5.4:

$$b_a(t) = b_a(t-1) - \eta^s \sum_{e \in E^s} \frac{\partial e(t)}{\partial b_a(t)} + \gamma^s[b_a(t-1) - b_a(t-2)], a \in \{x, y, z\} \tag{5.6}$$

$$E^s = \{\beta^{vp}\Delta^{vp}_{1...M^v}(t)\} \tag{5.7}$$

where b_a is a translation on the a-axis, γ^s is the momentum term (Rumelhart, Hinton, and Williams, 1986) used for the translation adaptation, and η^s is the according adaptation rate. The translation biases are initialized at 0 without variance. Since the motion direction of coordinates as well as their magnitude are invariant to translations, the adaptation is influenced only by the weighted error signals $\beta^{vp}\Delta^{vp}_{1..M^v}$ backpropagated along the processing paths that refer to the *coordinates* of the M^v encoded bodily features (see Figure 5.4).

The rotation R is implemented by a 3x3 neural matrix whose activity reproduces an extrinsic z-y-x Euler rotation. It is driven by three Euler angles α_x, α_y, and α_z. Each of the Euler angles results in a rotation around a specific Cartheisan axis and – similar to the translations – is represented by a bias neuron that can be adapted online by gradient descent on a set of errors signals. The adaptation over time follows the rule

$$\alpha_a(t) = \alpha_a(t-1) - \eta^r \sum_{e \in E^r} \frac{\partial e(t)}{\partial \alpha_a(t)} + \gamma^r[\alpha_a(t-1) - \alpha_a(t-2)], a \in \{x, y, z\} \tag{5.8}$$

$$E^r = \{\beta^{vp}\Delta^{vp}_{1...M^v}(t), \beta^{vd}\Delta^{vd}_{1...M^v}(t)\} \tag{5.9}$$

where γ^r implements the momentum term for the adaptation of the rotation, η^r implements the adaptation rate, and E^r is the set of error signals to minimize. The rotation biases are initialized at 0 without variance. In accordance with the connectivity shown in Figure 5.4, these error signals are backpropagated along the neural processing paths that relate to the *coordinates* and *motion* *directions* of all observed features. The positional error signals are represented by $\Delta^{vp}_{1..M^v}$, while $\Delta^{vd}_{1..M^v}$ represents the errors signals for the motion direction features, which are again weighted by factors β^{vp} and β^{vd}. The motion magnitude is invariant to rotation by nature and thus not considered in this process.

The calculation of the gradient for the adaptation of the rotation angles depends on the neural connectivity between the Euler angle bias neurons α_x, α_y, α_z and the rotation matrix layer R, which is exemplified in Figure 5.6. It consists of three sub-modules that represent rotations R_x, R_y and R_z – each representing an axis-specific rotation matrix – and an intermediate module. Accordingly, the matrices of activation functions for the three sub-modules are

$$R_x(\alpha_x(t)) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha_x & -\sin\alpha_x \\ 0 & \sin\alpha_x & \cos\alpha_x \end{pmatrix} \tag{5.10}$$

$$R_y(\alpha_y(t)) = \begin{pmatrix} \cos\alpha_y & 0 & \sin\alpha_y \\ 0 & 1 & 0 \\ -\sin\alpha_y & 0 & \cos\alpha_y \end{pmatrix} \tag{5.11}$$

$$R_z(\alpha_z(t)) = \begin{pmatrix} \cos\alpha_z & -\sin\alpha_z & 0 \\ \sin\alpha_z & \cos\alpha_z & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{5.12}$$

The respective pre-synaptic connection scheme is equivalent to encapsulated matrix-matrix multiplications, resulting in the activation $R = R_x(\alpha_x)R_y(\alpha_y)R_z(\alpha_z)$. Thus, the activity of the rotation layer R is again a rotation matrix, restricting all possible modulations of coordinates and motion directions to rotations.

The rotation of an object centered frame of reference is comparable to the visuo-spatial abilities of perspective-taking and mental rotation: Mental rotation is believed to be an information-driven mental process that rotates observed objects in their environment, while visuo-spatial perspective-taking is believed to be a mental process that rotates the observed environment onto the vantage points of others (Hegarty and Waller, 2004). Similarly, in the model, the whole visual percept is imagined from the perspective of another actor.

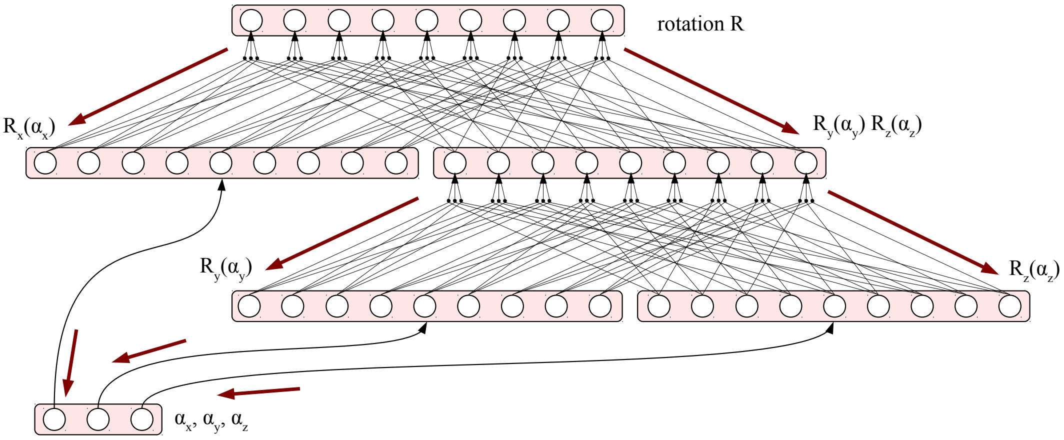

rotation R
$R_x(\alpha_x)$
$R_y(\alpha_y)\ R_z(\alpha_z)$
$R_y(\alpha_y)$
$R_z(\alpha_z)$
$\alpha_x, \alpha_y, \alpha_z$

FIGURE 5.6: Neural rotation matrix R for adaptation of the visual frame of reference. The rotation matrix is determined by two gain-field like presynaptic interaction schemes that resemble matrix-matrix multiplications. Neurons in the axis-specific rotation modules directly represent the elements of the respective rotation matrix, and are driven by bias neurons that directly represent rotation angles accordingly.

Moreover, spatial abilities are thought to be gradual and iterative: Higher degrees of misorientation between observer and model or object result in higher recognition time and lower recognition performance, which has also been observed for biological motion stimuli (Shepard and Metzler, 1971; Shepard and Metzler, 1988; Pavlova and Sokolov, 2000). In the model, gradient descent on top-down expectation errors approximates this iterative inference by approximating a minimal path rotation to an error-optimal target.

Taken together, perspective-taking is driven by top-down, embodied expectations about submodal perceptions that originate from embodied experience. The network utilizes partially invariant types of information that feature different predictive characteristics with respect to the visual perspective. This submodal decomposition is crucial for establishing the visual correspondence. While it has been shown that expectations that consider only the direction of motion are sufficient for perspective-taking (Schrodt et al., 2015), positional information adds further robustness, for example when no motion is observed. On the other hand, motion directions can already provide a suitable guess for the adaptation of the rotation while the translation is not fully converged, which does not generally hold for positional signals. Motion magnitude information is not considered for perspective-taking, but still is very crucial for establishing correspondence by means of feature binding, as will be explained in Section 5.3.4.

After perspective-taking, the model has access to a number of submodal visual features in an inferred frame of reference. These have to be selected and integrated into *Gestalt* perceptions, meaning that specific, relevant features have to be combined in an expected order. In preparation of this, the observed features are encoded by populations of artificial neurons as described in the following.

5.3.3 Visual and Proprioceptive Population Coding

After the extraction of submodal information and inference of a visual frame of reference, each submodal feature is encoded separately by one population of neurons with Gaussian tunings. At his step, the coding scheme is converted from neural activity that directly represents Cartesian vectors and scalars – which potentially cover arbitrary ranges – to topological neural activity that represents local responses to specific stimuli within a limited range. This approach is closer to the actual representation format of the brain and furthermore allows to co-encode stimulus uncertainties (Pouget, Dayan, and Zemel, 2000). It also supports subsequent computations as will be described later in this section. Thus, also the proprioceptive submodal features are encoded by this means.

The response $\dot{p}^{v}_{ai}$ of the a-th neuron in a population that encodes the position (p) of the i-th visually (v) observed feature is described by

$$\dot{p}^{v}_{ai}(t) = \left(r^{vp}\right)^{D^{vp}} \cdot \mathbb{N}\left(\mathbf{p}^{v}_{i}(t), \mathbf{c}^{vp}_{ai}, \sigma^{vp}\right) \tag{5.13}$$

where $\mathbb{N}(\mathbf{l}, \mathbf{m}, n)$ is the density of the multivariate normal distribution at $\mathbf{l}$ with mean $\mathbf{m}$ and a D^{vp}-dimensional diagonal covariance matrix $n \cdot I$. It specifies the response of the neuron to the current Cartesian, positional stimulus $\mathbf{p}^{v}_{i}$. Each neuron in a submodal feature population has an individual tuning or center $\mathbf{c}^{vp}_{ai}$ as well as a response breadth or variance σ^{vp}. The centers are evenly distributed in the expected range of the submodal stimuli, such that neighboring neurons are placed at a center distance of r^{vp}. The center distance determines the variance by

$$\sigma^{vp} = \zeta^{vp} \cdot (r^{vp})^2 \tag{5.14}$$

where $\zeta^{vp} \in (0, 1]$ is the *continuity factor*. Scaling the density function by r^{vp} to the power of the dimension D^{vp} of the feature ensures that the sum of activity in the population is approximately 1 for stimuli in the expected range. The continuity factor modulates the breadth of the cell tunings in a way that approximately preserves the sum of activity in the population (see Figure 5.7). It parameterizes a trade-off between an evenly covered input space (for $\zeta^{vp} = 1$) and a high derivative of single cell activity with respect to local stimuli changes (e.g. for $\zeta^{vp} = 0.1$). A high derivative is advantageous for online learning of generative codes, because it effectively relaxes the problem that an activated code is continuously overfitted to the current input and thus no actual learning takes place, which can happen particularly when using high learning rates and sequential learning in autoencoders (cf. catastrophic forgetting problem, McCloskey and Cohen, 1989).

Analogously to the coordinate populations, directional motion features are represented by scaled Gaussian activity via

$$\dot{d}^{v}_{ai}(t) = \left(r^{vd}\right)^{D^{vd}} \cdot \mathbb{N}\left(\mathbf{d}^{v}_{i}(t), \mathbf{c}^{vd}_{ai}, \sigma^{vd}\right) \tag{5.15}$$

while motion magnitude features are represented by

$$\dot{m}^{v}_{ai}(t) = (r^{vm})^{D^{vm}} \cdot \mathbb{N}\left(m^{v}_{i}(t), c^{vm}_{ai}, \sigma^{vm}\right) \tag{5.16}$$

The arrangement of the centers is set up to be in accordance with the dimension, range, and configuration space of the respective submodal stimuli. Consequently, neurons that encode visual positional features ($D^{vp} = 3$) are arranged evenly on a 3D grid in a specific range. In contrast, neurons that encode visual directional motion ($D^{vd} = 3$) as unit vectors are arranged on the surface of a unit sphere, while neurons that encode visual motion magnitudes ($D^{vm} = 1$) are distributed linearly, as also indicated in Figure 5.4.

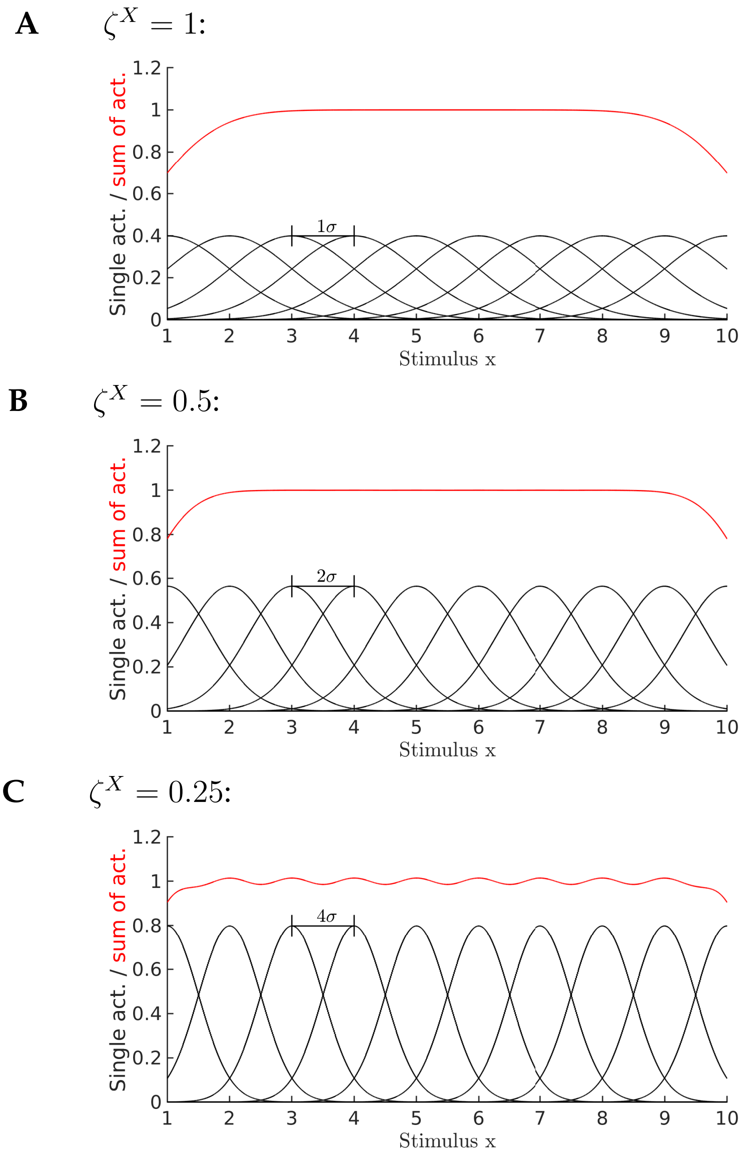


FIGURE 5.7: Influence of the continuity factor ζ^X on cell tunings and the resulting sum of activity in populations. When a submodal stimulus x is within a specified range (here $x \in [0..10]$), the resulting sum of activity (red) in a population of Gaussian tuned neurons that represent the stimulus converges to about 1 in the inner area of the range and drops at the border areas of the range. Lowering the continuity factor $\zeta^X \in [0, 1]$ results in increasingly narrow Gaussian tunings (black) of the cells (from **A** to **C**). As a result, changes in the stimulus result in higher changes in the resulting population activity. While the sum of activity increases in outer regions of the covered population range, the energy in inner regions becomes increasingly inconsistent.

The principle of encoding proprioceptive features is generally comparable ($D^{pp} = D^{pd} = 3$, $D^{pm} = 1$), with the exception that both proprioceptive motion direction features as well as proprioceptive postural features are represented by unit vectors and thus the according population neurons are arranged on the surface of a sphere, as indicated in Figure 5.5. The uniform arrangement of centers on a unit sphere can be approximated by distributing the centers evenly on the surface of a cube and projecting the centers on the sphere by normalizing their coordinates to unit length. Note that small motion signals activate the respective Gaussian neurons tuned to directional motion rather uniformly, since the normalization of their driving signal is limited by the recognition threshold o (see Section 5.3.1).

Not alone because of the different nature of the six submodal stimuli types, it is important to apply an encoding scheme that results in comparable codes. Here, specific stimuli are not under- or over-represented by means of their level of activity, since their representation is de-linearized by the Gaussians. This ensures that all submodal stimuli are represented approximately by the same magnitude of neural activity also if a stimulus lies between the centers of two neighbouring Gaussians. Given that there are predictions about the current population encoded submodal stimulus, the deviations from the actual stimulus are thus expected to be relatively comparable across the submodalities of a single feature, as well as across the features. Respective predictions are generated by higher layers of the network and used both for learning codes from the stimuli as well as adapting the frame of reference. Thus, given that the stimulus contingencies are learned with comparable precision, all error signals originating from different bodily features have approximately the same top-down influence on the adaptive mechanisms described before. Assuming that the variance in the population coded submodal stimuli over the training set is comparable across features, the speed of learning the contingencies of each individual feature is comparable as well, resulting in overall balanced submodal Gestalt codes. To a certain degree, the approach is similar to individual input scaling based on variance per input dimension, which is often performed for autoencoders (cf. Scholz and Vigário, 2002), but requires additional preprocessing of the data set.

Further, using population codes, multiple conflicting stimuli can be encoded in parallel, which can then be disambiguated by adaptive mechanisms, which will be used to tackle the binding problem. Note that all observed features must be handled equally and without prior knowledge by the network as long as they have not been identified and assigned to bodily features. Thus, also the ranges of the feature populations *must not* differ per feature, but they must be set uniformly for each submodality.

In summary, Cartesian vectors can be encoded by populations of neurons with Gaussian responses that satisfy specific energy constraints and thus support the development of balanced representations and the adaptive perceptual processes. Vice versa, population encoded activity can be decoded by means of trilateration to obtain the corresponding Cartesian stimuli, which is used in the model evaluations to visualize them accordingly. This procedure is explained in the following.

Decoding Population Encoded Stimuli

A population coded stimulus can be decoded by calculating the distance of the stimulus to the center of the respective neurons and approximating the intersection of spheres with the respective radii that originate from the respective centers (see Figure 5.8). Trilaterations can usually be solved in closed form. In case of the proposed model, however, populations contain between 8 and 64 neurons, such that the solution is over-determined. Thus, an iterative, gradient-based optimization process is used here for the multilateration problem (cf. Sirola, 2010). Note that the following method is generic, thus, a standard notation is used with superscripts that do not refer to (sub)modalities of the network but represent powers.

Generally, the distance d_a of a stimulus s to the center of a Gaussian tuned neuron with index a as described before can be derived from its activation x_a by

$$d_a(t) = \sqrt{-2\sigma \log\left(x_a(t)(2\pi\sigma)^{D/2}\right)} \tag{5.17}$$

where σ is the variance and D is the dimension of the center. Deriving the stimulus directly is not possible. However, the stimulus can be approximated iteratively by minimizing the deviation of the given stimulus distances $d_a \forall a$

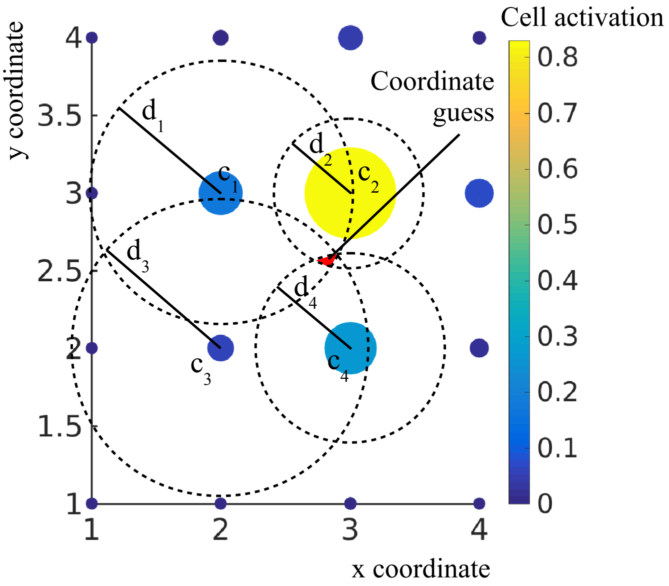


FIGURE 5.8: Example for multilateration of a stimulus $(2.8, 2.6)$ encoded by a population of 4x4 Gaussian tuned cells with centers/tunings $c_i \in (1...4, 1...4)$. The color and size of the dots represents the level of activity of a locally tuned neuron. From these, the distances d_i to the stimulus can be inferred, indicating that the stimulus is expected to be placed on a corresponding circle around the respective center (dashed lines). Here, the actual stimulus is approximately in the intersection of the circles (red area). Generally, the stimulus can be approximated by iterative optimization processes.

from the distance of a *guessed stimulus* $\tilde{\mathbf{s}}$ to the respective centers $\mathbf{c}_a \forall a$. This results in the objective function $E(\tau)$ for minimization over sub-steps in time τ

$$E(\tau) = \sum_a {}^1\!/{}_2 \left(d_a(t) - \|\tilde{\mathbf{s}}(\tau) - \mathbf{c}_a\|\right)^2 \tag{5.18}$$

The derivative of E with respect to the stimulus guess $\tilde{\mathbf{s}}$ leads to the gradient descent based iterative update rule

$$\tilde{\mathbf{s}}(\tau) = \tilde{\mathbf{s}}(\tau-1) + \eta \cdot \sum_a w_a(t) \cdot (d_a(t) - \|\tilde{\mathbf{s}}(\tau) - \mathbf{c}_a\|) \cdot (\tilde{\mathbf{s}}(\tau-1) - \mathbf{c}_a) \tag{5.19}$$

where η specifies the adaptation rate for the iterative optimization. w_a weights the influence of each neuron in the process. Strictly following the gradient results in $w_a = 1 \forall a$. However, the influence of each neuron is set relative to its activation, such that neurons closer to the stimulus obtain a higher weight in the optimization process. This approach is expedient since changes in the activity of neurons distant to the stimulus result in much greater changes in the guessed stimulus than changes in activity of neurons close to the stimulus. If, for example, the population activity is influenced by i.i.d. noise or other imprecisions, weighting by

$$w_a(t) = \frac{x_a(t)}{\max_b x_b(t)} \tag{5.20}$$

results in an improvement of the approximation. Alternatively, the distance could be used. Since learning algorithms typically prefers to eliminate high errors in activations (i.e. neurons with high activity close to the stimulus) over low errors (i.e. neurons with low activity far from the stimulus), weighting is useful.

The last coordinate guess from time step $t-1$ is used as initial guess for the minimization. The minimization is complete when the gradient length $\|\tilde{\mathrm{s}}(\tau)\|$ falls below a specified value that defines the precision of convergence, or when τ exceeds a specific iteration limit. The precision has to be defined relative to the expected range of the stimulus, which differs across the submodal types of information.

Decoded stimuli are used in the evaluations to visualize the submodal Gestalt perceptions that are expected by the model and come in the form of population coded activity. Besides perspective-taking, these expectations also drive the binding of input features into Gestalt percepts, which is described in the following.

5.3.4 Visual Feature Binding

As described in Section 3.3, the binding problem concerns the selection and integration of separate visual features in the correct combinations, which is assumed to be related to attention (Treisman, 1998). Thus, for solving this problem, an approach is chosen that is comparable to directing attention selectively to observed features that match expected relative positions and motion dynamics, and by integrating these local features into a global Gestalt context. The model assumes that each bodily feature is processed by a specific neural processing path, such that features observed in an arbitrary order have to be put into the correct (or expected) order. Features that do not match the Gestalt context have to be neglected. To do so, the model utilizes partially invariant top-down expectations that come from the separate submodal encodings.

Both the selection of features relevant for the recognition of biological motion as well as the assignment to the respective neural processing paths are handled consistently by an adaptive, gated connectivity matrix from *observed features* $i \in \{1...N^v\}$ to *bodily features* $j \in \{1...M^v\}$, $M^v \leq N^v$, resulting in

$$\ddot{\mathbf{p}}^v_j(t) = \sum_{i=1}^{N^v} w_{ij}(t) \cdot \dot{\mathbf{p}}^v_i(t) \tag{5.21}$$

$$\ddot{\mathbf{d}}^v_j(t) = \sum_{i=1}^{N^v} w_{ij}(t) \cdot \dot{\mathbf{d}}^v_i(t) \tag{5.22}$$

$$\ddot{\mathbf{m}}^v_j(t) = \sum_{i=1}^{N^v} w_{ij}(t) \cdot \dot{\mathbf{m}}^v_i(t) \tag{5.23}$$

where $\ddot{\mathbf{p}}^v_j$, $\ddot{\mathbf{d}}^v_j$, and $\ddot{\mathbf{m}}^v_j$ represent the population encoded activations of the j-th assigned – or bodily – submodal feature in the position, motion direction and motion magnitude domains respectively, $\dot{\mathbf{p}}^v_i$, $\dot{\mathbf{d}}^v_i$, $\dot{\mathbf{m}}^v_i$ represent the according activity of the i-th unassigned – or observed – submodal feature, and w_{ij} represents the corresponding assignment strength $\in (0, 1)$. The assignment strength is implemented by a non-linear neuron with logistic activation function, such that

$$w_{ij}(t) = \frac{1}{1 + \exp\left(-w^b_{ij}(t)\right)} \tag{5.24}$$

where w^b_{ij} is the activity of an adaptive bias neuron. The biases are adapted with respect to the weighted error signals that originate from all submodal bodily features, as also indicated in Figure 5.4:

$$w^b_{ij}(t) = w^b_{ij}(t-1) - \eta^w \sum_{e \in E^w} \frac{\partial e(t)}{\partial w^b_{ij}(t)} + \gamma^w [w^b_{ij}(t-1) - w^b_{ij}(t-2)] \tag{5.25}$$

$$E^w = \{\beta^{vp} \Delta^{vp}_{1...M^v}(t), \beta^{vd} \Delta^{vd}_{1...M^v}(t), \beta^{vm} \Delta^{vm}_{1...M^v}(t)\} \tag{5.26}$$

where γ^w is the momentum and η^w is the adaptation rate for the assignments. The assignment biases w_{ii}^b that map from the i-th observed feature to the i-th bodily feature are initialized at 1000 without variance (resulting in $w_{ii} \rightarrow 1$), while all other assignments are initialized at -1000 without variance (resulting in $w_{ii} \rightarrow 0$) during training, since the assignment is assumed to be known to the model during self-observation. During testing, all assignment biases are initialized at -10 without variance, resulting in an initial, subtle mixture of all possible assignments, since the assignment is unknown and to be inferred by the model.

Obviously, neither the selection nor assignment are restricted to specific connectivity patterns in this approach: One observed feature may be assigned to multiple bodily features and vice versa. Observed features may not get assigned at all as well. Furthermore, the assignment can adopt continuous values between 0 and 1, such that also mixtures of assignments may occur. Interestingly, as evaluated extensively in exploratory studies, none of these eventualities do need further attention when using the population coding scheme described before. In fact, inconsistent assignments seem rather beneficial for the process of binding observed features to embodied representations, as I will explain in the following.

First, observed distractor features can be neglected by the model simply by not assigning them to any bodily feature. This results in a selection of the features that match the expected biological motion dynamics best. Second, when multiple observed features are assigned to a single bodily feature, the activities in the respective bodily population is summed up and the respective stimuli are maintained in parallel. Then, there are either multiple, distinct local activations in the bodily population or overlapping activations. When the mentioned top-down error signals are applied to adapt the assignment, however, only a single, local activation is expected. Furthermore, a specific overall energy is expected in the population. Thus, there is an implicit tendency to 'decide' for only one of the stimuli (the best matching one) by increasing its assignment weight, while at the same time decreasing the weights of the others. The matching furthermore considers the different submodalities: For example, two features may have the same motion direction, but they are nonetheless disambiguated via the postural differences. Analogously, features may be located at ambiguous positions that do not clearly reveal their identity, in which case the motion dynamics can decide. As a result, the network will gradually infer the single association that best matches the expected dynamics, while neglecting features that do not match any of the expectations. In the process, the weight of a selected feature will converge towards 1 to match the expected energy in the population, while all other assignment weights converge towards 0. Note that it is also not possible that the weighted sum of two *different*, population coded stimuli results in an expected stimulus (in contrast to Cartesian encodings). Thus, there is no theoretical possibility that the algorithm will converge towards pseudo assignments that consist of mixtures of observed features.

While the process of selection and assignment is applied to individual input features even in hypothetical, parallel state, it is not possible to completely minimize the error with respect to the expectancy on this basis only. If the observed perspective does not correspond the learned, egocentric perspective, the global minimum of the error can only be minimized by adaptations of the global frame of reference, handled by the herein before mentioned perspective-taking processes (see Section 5.3.2), adapting all observed features in the same manner. Thus, both processes have to be applied in parallel.

Clearly, however, there is a dependency of the two processes on each other. Assuming that the correct rotation and translation have not been inferred so far, it is difficult to test hypothetical associations against given expectations, since in turn, these originate from a different (egocentric vs. allocentric) perspective. Thus, there is a strong likelihood that none of the observed features is matching. Moreover, the expectations initially come from the codes of higher layers that are activated by the imperfect (non-rotated, non-translated, non-selected, non-assigned) stimuli and are thus imperfect as well. To resolve this dependency, the segregation of the inputs into different submodal perceptions that feature different invariances to spatial transformations is crucial. For the magnitude of coordinate motion, both rotation and translation do not play a role in a global frame of reference. Thus, the magnitude expectations should provide the best initial guess to drive the assignment of the features, improving the assignment of the position and motion direction features likewise. The magnitude error is not suited to provide signals for rotation and translation. Given a (preliminary) assignment, however, there is a less arbitrary error signal for the rotation that comes from the expected motion directions, as well as an error signal for both rotation and translation that comes from the expected posture, which finally lead to overall convergence of the perceptual adaptation.

Note that the proposed sequence in which information is extracted and adaptations are applied in the network is carefully considered: Adaptations of the assignment and selection have no influence on the submodal information determined beforehand, i.e. they do not result in false perceptions of motion. Likewise, the adaptation of the model's rotation does not result in a perceived magnitude of motion, while the translation has no influence on the direction of motion. Thus, perceptual adaptation is not confused with actual point light motion, which would interfere in the process.

All things considered, the perceptual adaptation and extraction of information works on several specifically pre-structured, locally and globally complementary domains. It derives an embodied representation from observed features by transforming global perspectives and establishing the corresponendes of selected, single features to Gestalt-like configurations. Thus, self-representations can be activated independent of the perspective on observed actions. In the process, it is crucial to segregate and facilitate submodal types of information, and learn expectations from them with different top-down predictive characteristics. In the following, I will describe the development of embodied, spatially and temporally predictive codes that provide suitable error signals for the solution to the correspondence and binding problems described here.

5.4 Predictive Coding and Embodied Simulation: A Temporal Conditional Autoencoder

This stage of the model serves the purpose (i) to learn the embodied contingencies of submodal perceptions in compressed, generative codes, (ii) learn temporal predictions of the submodal codes in dependency on other submodal codes for embodied simulation, and (iii) to generate submodal Gestalt expectations for perceptual inference by means of perspective-taking and feature-binding. In the following, I introduce a temporal conditional autoencoder that serves as the building block for encoding the six described visual and proprioceptive submodalities, which is shown in Figure 5.9. The autoencoders basically learn compressed, spatially predictive codes of their inputs as described in Section 5.4.1, and temporally predictive codes from the developing spatial codes as described in Section 5.4.2, which are modulated by top-down intention codes.

5.4.1 Spatial Representation

Each set of submodal body feature populations is joined into a *submodal Gestalt* vector g^X. That is, both the visual and proprioceptive modules provide separate Gestalt vectors for the positional / postural, and the respective motion direction and motion magnitude submodalities. The features within each set are concatenated in a specified order. As described in Section 5.3.4, the order in which the body features are assigned from the observed features can be inferred by adaptive, error-minimizing processes. I denote these Gestalt vectors

$$\mathbf{g}^{vp}(t) = (\ddot{\mathbf{p}}_1^v(t), ..., \ddot{\mathbf{p}}_{M^v}^v(t)) \tag{5.27}$$

$$\mathbf{g}^{vd}(t) = \left(\ddot{\mathbf{d}}_1^v(t), ..., \ddot{\mathbf{d}}_{M^v}^v(t)\right) \tag{5.28}$$

$$\mathbf{g}^{vm}(t) = (\ddot{\mathbf{m}}_1^v(t), ..., \ddot{\mathbf{m}}_{M^v}^v(t)) \tag{5.29}$$

$$\mathbf{g}^{pp}(t) = (\ddot{\mathbf{p}}_1^p(t), ..., \ddot{\mathbf{p}}_{M^p}^p(t)) \tag{5.30}$$

$$\mathbf{g}^{pd}(t) = \left(\ddot{\mathbf{d}}_1^p(t), ..., \ddot{\mathbf{d}}_{M^p}^p(t)\right) \tag{5.31}$$

$$\mathbf{g}^{pm}(t) = (\ddot{\mathbf{m}}_1^p(t), ..., \ddot{\mathbf{m}}_{M^p}^p(t)) \tag{5.32}$$

To learn distributed, predictive encodings of actions, each submodal Gestalt perception is encoded separately by one spatially and temporally predictive, generative autoencoder. Each autoencoder entails a *code fusion* vector $\mathbf{f}^X$, which can be seen as a compressed, non-linear representation of a submodal Gestalt, and which is the result of fusing temporal *code predictions* $\tilde{\mathbf{f}}^X$ with spatial *code observations* $\dot{\mathbf{f}}^X$, based on a submodal stimulus reliability $q^X \in [0, 1]$:

$$\mathbf{f}^X(t) = \tanh\left(q^X(t)\dot{\mathbf{f}}^X(t) + (1 - q^X(t))\tilde{\mathbf{f}}^X(t) + \mathbf{b}^X(t)\right) \tag{5.33}$$

$$X \in \{vp, vd, vm, pp, pd, pm\}$$

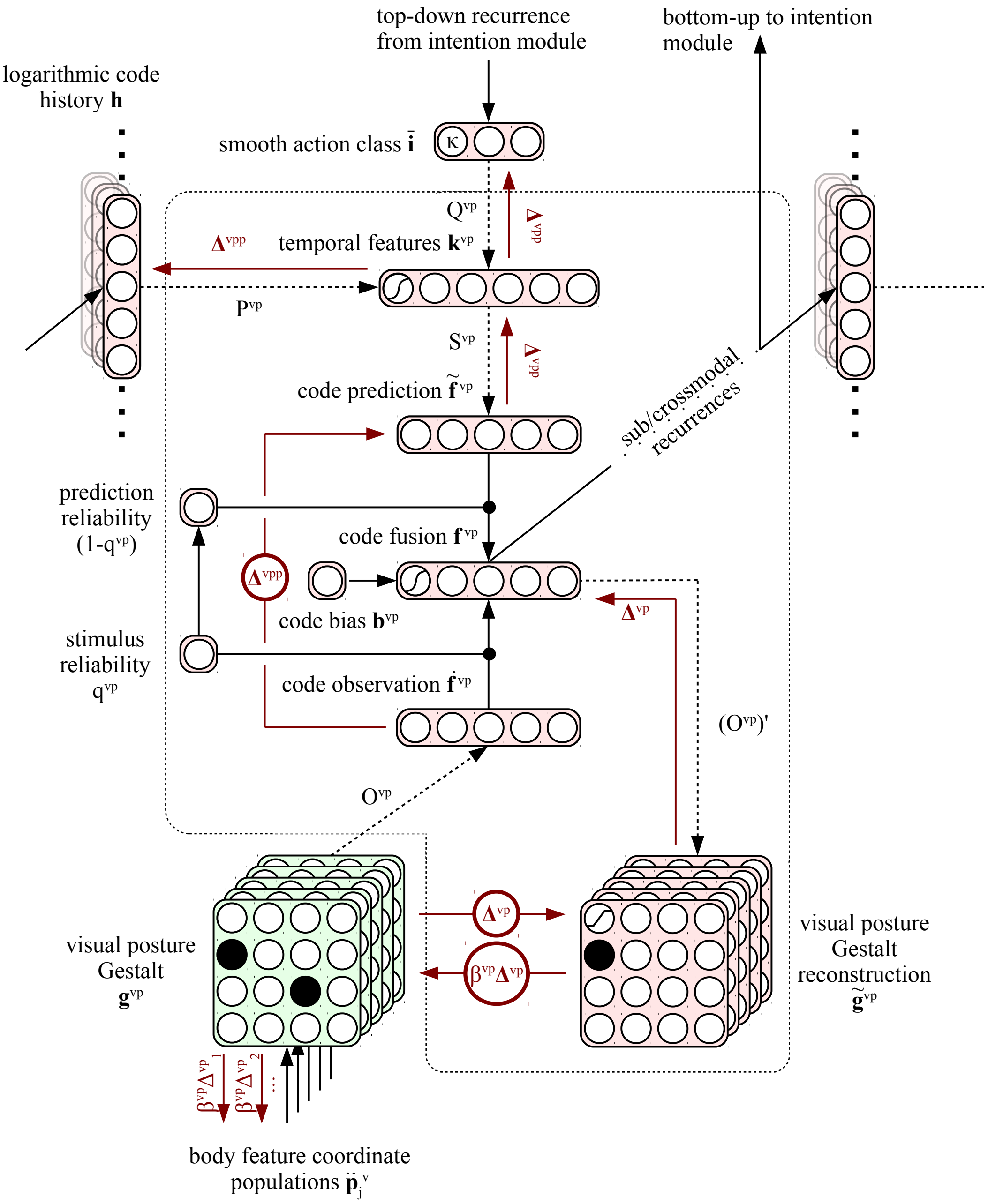

top-down recurrence from intention module
bottom-up to intention module
logarithmic code history **h**
smooth action class $\bar{\mathbf{i}}$
κ
Q^{vp}
Δ^{vpp}
temporal features $\mathbf{k}^{vp}$
Δ^{vpp}
P^{vp}
S^{vp}
Δ^{vpp}
code prediction $\tilde{\mathbf{f}}^{vp}$
sub/crossmodal recurrences
prediction reliability $(1\text{-}q^{vp})$
code fusion $\mathbf{f}^{vp}$
Δ^{vpp}
code bias $\mathbf{b}^{vp}$
Δ^{vp}
stimulus reliability q^{vp}
code observation $\dot{\mathbf{f}}^{vp}$
$(O^{vp})'$
O^{vp}
visual posture Gestalt $\mathbf{g}^{vp}$
Δ^{vp}
$\beta^{vp}\Delta^{vp}$
visual posture Gestalt reconstruction $\tilde{\mathbf{g}}^{vp}$
$\beta^{vp}\Delta^{vp}_1$
$\beta^{vp}\Delta^{vp}_2$
...
body feature coordinate populations $\ddot{\mathbf{p}}_j^v$

FIGURE 5.9: Connectivity scheme of a submodal temporal conditional autoencoder (here, in the visual posture submodality). The autoencoder learns compressed spatial codes of submodal Gestalt perceptions that come in the form of grouped feature population activity. It then learns temporal predictions of the codes by extracting temporal features from a logarithmic history of activated spatial codes across modalities, and by application of a top-down kinematic intention bias. Predicted and observed codes can be fused based on presence and reliability of sensory stimuli. The autoencoder generates spatial and temporal prediction errors for learning and perceptual adaptation.

where $\mathbf{b}^X$ is a trainable bias. The code observation is activated by the respectively observed submodal Gestalt via a weight matrix O^X:

$$\dot{\mathbf{f}}^X(t) = O^X(t)\mathbf{g}^X(t) \tag{5.34}$$

To generate submodal Gestalt expectations, a *Gestalt reconstruction* $\tilde{\mathbf{g}}^X$ that decodes the fused code vector via the transposed transformation $\left(O^X\right)'$ is obtained via a rectified linear function

$$\tilde{\mathbf{g}}^X(t) = \min\left(\max\left(\left(O^X(t)\right)' \cdot \mathbf{f}^X(t), 0\right), u^X\right) \tag{5.35}$$

The autoencoders thus use tied or shared weights to reduce the parameter complexity of the spatial encoder/decoder (cf. LeCun, 1989; LeCun et al., 1990), meaning that the same weights are used for activation of the code observation as well as the generation of input reconstructions. Since the reconstruction directly reflects scaled Gaussian activity as described in Section 5.3.3, the range of the respective neural activity is limited to reflect the codomain of the respective Gaussians, which is realized via the activity maximum u^X of the rectified linear units:

$$u^X = \zeta^X \left(\frac{\pi}{2\sigma^X}\right)^{D^X/2} \tag{5.36}$$

To represent sensory contingencies and form submodal encodings for action recognition, the code biases $\mathbf{b}^x$ and tied activating and generative weights O^X are trained to minimize the reconstruction error with respect to the Gestalt input which serves as target signal. The model assumes full sensory reliability ($q^X = 1$) during this spatial training procedure, such that the code fusions, as well as the Gestalt reconstructions, are driven by the current perceptions only and not by code predictions. The difference to the Gestalt reconstruction is backpropagated over the transposed weights to the codes, as indicated in Figure 5.9. Then, the objective for minimization is

$$\begin{aligned}\Delta^X(t) &= {}^1\!/\!_2 \left\|\mathbf{g}^X(t) - \tilde{\mathbf{g}}^X(t)\right\|^2 &(5.37)\\ &= \{\Delta^X_{1\ldots M^Y}(t)\} &(5.38)\end{aligned}$$

where Y is the respective modality. According to that, weights in O^X are updated according to the rule

$$o_{ij}^X(\tau) = o_{ij}^X(\tau - 1) - \eta^{Xc} \frac{\partial \Delta^X(t)}{\partial o_{ij}^X(t)} + \gamma^{Xc}[\partial o_{ij}^X(\tau - 1) - \partial o_{ij}^X(\tau - 2)] \tag{5.39}$$

where o_{ij}^X is an element of O^X, and γ^{Xc} and η^{Xc} are the respective momentum and learning rates for spatial learning in autoencoders. Biases $\mathbf{b}^X$ are updated analogously. The spatial weights and code biases are initialized by a normal distribution with mean 0.0 and variance 0.1.

A significant difference of this formulation to the gradient update rules for adaptive network components should be noted: Weight updates for adaptation are generally applied in the current time step t, whereas weight updates for training are applied at *random* time steps $\tau \geq t$, that is, after a random delay. The learning time step τ is obtained by the following, algorithmic approach: Each trained weight in the network maintains a list of future updates with size $s \leq \Theta$, where Θ is the temporal horizon for the application of learning gradients. At each time step t, the current gradient based update for the weight is appended to the list. Instead of a direct application, however, a random element of the list is chosen with a probability of s/Θ, or else no update is chosen. Thus, the list will eventually reach the size Θ, and a random update is applied at each time step after an expected delay of Θ time steps. A similar practice was applied by Mnih et al., 2013. Here, it pursues the purpose of breaking the temporal correlation between gradients to further decrease the problem of recoding (like population coding) without having to shuffle the order of the inputs, which would violate the online learning criterion (see Chapter 5.2). When applying gradient updates at random time steps, an immediate influence of the update on subsequently activated codes is effectively circumvented.

Taken together, each autoencoder learns compressed, non-linear, spatial codes via gradient descent on an error signal produced by itself, and thus the learning algorithm can be considered to be unsupervised. Except for the non-linearities, this procedure is comparable to finding the principle components of the Gestalt stimuli, effectively finding representative factors for the data while reducing the input dimensionality (cf. Jolliffe, 1986). In the process, the bias vector $\mathbf{b}^X$

is supposed to approximate the average code over the training set, thus indirectly representing a general, perceptual bias for a submodal Gestalt. The code bias serves as input to the *fused* code, and thus biases both observed and predicted codes equally. For observations, it represents the best initial guess of a code when no salient feature is currently assigned, or grouped into a Gestalt, such that hardly any input from the Gestalt populations is forwarded to the code observation. The bias is thus useful to obtain an instant expectation of the currently observed stimulus, which can help to assign features and infer the perspective concurrently. Conversely, if no bias would be present, there also would be no driving signal for assigning features to the Gestalt perceptions, since there would be no expectation error. Thus, learning such perceptual biases is of utmost importance.

While the deviations between submodal Gestalt stimuli and the respective Gestalt reconstructions are propagated over the weights of the autoencoders to learn compressed, generative codes, the same error signal is used for adaptation of the rotation, translation, and binding of visual features as shown in Figure 5.4. To accomplish this, the errors are backpropagated not via the fused codes, but directly over the Gestalt population, weighted by the factors β^X, and along the separate visual bodily feature populations, their inherent submodal structures, and in consequence also weighted along the assigned, visually observed features as described in Section 5.3.4. Direct propagation avoids that the error signal is bottle-necked or distorted by the compressed code f^X of the autoencoder. Note that here, the reconstruction can be considered the target instead of the input to an autoencoder, however resulting in the same quadratic error signal for minimization, such that the perception is able to adapt to the expected stimulus.

After separate spatial representations for the submodal sensory contingencies have been learned, the autoencoders learn to predict the temporal progress of the respective codes, which is described in the following.

5.4.2 Temporal Representation

Besides learning compressed, generative, spatial codes from the population encoded Gestalt stimuli, the autoencoders also learn temporal representations that predict the progress of the developing codes, and – indirectly via the generative weights – can also predict the progress of the multimodal Gestalt stimuli. To this end, the overall network architecture develops predictive attractors in the distributed submodal encodings that also drive and incorporate action intentions as argued in Section 2.2.3. The predictive structures are essentially established via recurrent, bi-directional connectivity between the autoencoders and the intention module: Each submodal autoencoder predicts the progress of its own fused codes from recently activated codes within the submodality, as well as recently activated codes of other submodal autoencoders, while the prediction is being biased by the continuous inference of an action class as will be described in Section 5.5. Biasing via the class follows the purpose to determine the activated attractor and to resolve ambiguities in the predictions such that different action types can be simulated consistently, considering that the submodal codes for all actions are encoded in the same structures. *Simulation* here refers to the reenactment of learned embodied codes. These distributed, submodal simulations are partially self-preserving, and partially influenced by the crossmodular connectivity. As a result, simulation goes beyond a simple, static mapping from visual to proprioceptive stimuli and vice versa. Rather, the method establishes consistency over time between distributed modal and submodal domains that mutually synchronize their activity until an overall consistent, learned attractor is reached.

Furthermore, by setting the submodal stimulus reliabilities q^X accordingly, the network obtains the capability of selectively triggering modal simulations which are constantly biased or temporarily primed by observations and rely on the learned embodied codes. The simulation concerns several aspects: Maybe most importantly in the context of action understanding, it enables to infer or imagine the progress of proprioceptive sensations in terms of whole body posture, motion directions, and motion magnitudes when observing actions visually, while at the same time inferring the type of action as part of the distributed simulation. The same principle of embodied simulation also results in the networks ability to simulate consistent whole-body motion in the visual, proprioceptive, and intention domains without any sensory stimulus. Thus, the method shows how predictive encodings yield inference as well as simulation capabilities that can be flexibly adapted and used in multiple directions.

The predictive mechanism applied here assumes that a specific time window of the recently activated spatial codes, together with a constant action bias, is sufficient for predicting the next code in each submodality. Thus, each autoencoder learns a non-linear transformation from a *logarithmic history* of a set of submodal fused codes (including its own codes) to identify *temporal features* which are suitable for prediction. Similar to relevance of the submodal encodings for perceptual inference, the submodal encodings feature different predictive characteristics also with respect to the lateral inference of submodal codes. The logarithmic code history $\mathbf{h}$ serves as lateral input to each autoencoder (additional to the bottom-up input from the Gestalt populations) and is set up by the following tuple-builder notation:

$$\mathbf{h}(t) = \left(\frac{\sum_{d=2^{i-1}}^{2^i-1} \mathbf{f}^X(t-d)}{2^{i-1}} \mid i \in \{1, ..., \theta\}, X \in \{vp, vd, vm, pp, pd, pm\} \right) \quad (5.40)$$

where θ parameterizes the temporal horizon for the prediction of submodal codes. For example, $\theta = 3$ results in the code history vector

$$\begin{aligned} \mathbf{h}(t) = \Bigg(& \mathbf{f}^{vp}(t-1); \frac{\sum_{d=2}^{3} \mathbf{f}^{vp}(t-d)}{2}; \frac{\sum_{d=4}^{7} \mathbf{f}^{vp}(t-d)}{4} \\ & \mathbf{f}^{vd}(t-1); \frac{\sum_{d=2}^{3} \mathbf{f}^{vd}(t-d)}{2}; \frac{\sum_{d=4}^{7} \mathbf{f}^{vd}(t-d)}{4} \\ & ... \\ & \mathbf{f}^{pm}(t-1); \frac{\sum_{d=2}^{3} \mathbf{f}^{pm}(t-d)}{2}; \frac{\sum_{d=4}^{7} \mathbf{f}^{pm}(t-d)}{4} \Bigg) \end{aligned}$$

The approach essentially averages the last codes that were activated in exponentially (in this case to base 2) increasing time windows. Thus, past submodal information can be preserved with decreasing precision without having to consider each of the time steps separately (cf. Oord et al., 2016). Figure 5.10 qualitatively describes the information that the logarithmic history extracts from the codes. Note that possibly not all of the submodalities carry relevant information for the prediction of other modalities, thus the set of used submodalities can be restricted a priori, which is however not done here. Here, a learning algorithm is raised to identify the indicative temporal features within the set of all

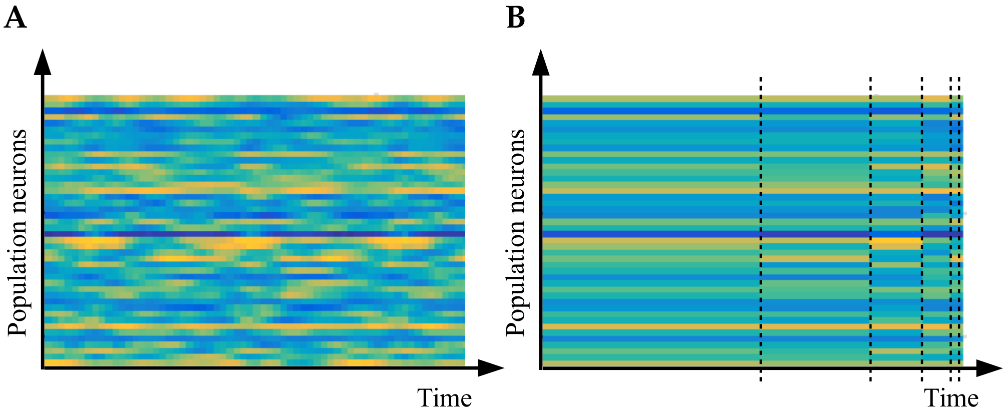


FIGURE 5.10: Comparison between the history of activated spatial codes in an autoencoder (**A**) and their logarithmic representation (**B**) that serves as lateral input to an autoencoder at each time step. The activity is color coded, ranging from 0 (blue) to 1 (yellow). In the logarithmic representation, exponential time windows of the codes are averaged, such that the autoencoder receives a short history (here, six inputs for $\theta = 6$) of the recently activated codes (gathered over 63 time steps) with decreasing precision backwards in time.

embodied submodal codes. Thus, the code history inputs to all autoencoders are identical.

For learning temporal predictions, the logarithmic code history is transformed non-linearly to a temporal feature vector $\mathbf{k}^X$ via a transformation matrix P^X. Furthermore, the temporal features are basically biased by a top-down action class vector $\bar{\mathbf{i}}$, provided by the intention module, via a transformation matrix Q^X. The intention bias comes as a result of classifications of submodal codes as described in the next chapter. The top-down influence of the action class is stabilized by low-pass filtering. Since the (partially) predicted codes may determine the predicted class itself in the long run, this avoids that the network fluctuates quickly between different classes. Formally, this results in

$$\mathbf{k}^X(t) = \tanh\left(P^X(t)\mathbf{h}(t) + Q^X(t)\bar{\mathbf{i}}(t)\right) \tag{5.41}$$

$$\bar{\mathbf{i}}(t) = \kappa\bar{\mathbf{i}}(t-1) + (1-\kappa)\mathbf{i}(t) \tag{5.42}$$

$$X \in \{vp, vd, vm, pp, pd, pm\} \tag{5.43}$$

where κ is the intention dynamics parameter (slow dynamics for $\kappa \to 1$), and $\mathbf{i}$ is the current intention inference of the model. The transformation from logarithmic codes to biased temporal features is expected to extract meaningful hidden states for the code prediction $\tilde{\mathbf{f}}^X$ in the local submodality X, while the transformation from the class provides slow dynamic biases. The resulting hidden features are transformed again via a linear transformation S^X to yield the code prediction

$$\tilde{\mathbf{f}}^X(t) = S^X(t)\mathbf{k}^X(t) \tag{5.44}$$

Altogether, the connectivity of the predictive component of each autoencoder is comparable to a multilayer perceptron with specific input properties.

The involved transformation matrices are learned while the model is driven by self-perceptions with full sensory certainty ($q^X = 1$). Thus, the activated fused codes stem from the bottom-up perceptions alone, and they are not influenced by the current prediction. Given a predicted code $\tilde{\mathbf{f}}^X$ and an observed code $\dot{\mathbf{f}}^X$, an error signal Δ^{Xp} is constructed that is backpropagated over, and used to train the weight matrices P^X, Q^X, and S^X:

$$\Delta^{Xp}(t) = {}^1\!/_2 \left\| \tilde{\mathbf{f}}^X(t) - \dot{\mathbf{f}}^X(t) \right\|^2 \tag{5.45}$$

$$p_{ij}^X(\tau) = p_{ij}^X(\tau-1) - \eta^{Xp}\frac{\partial \Delta^{Xp}(t)}{\partial p_{ij}^X(t)} + \gamma^{Xp}[p_{ij}^X(\tau-1) - p_{ij}^X(\tau-2)] \tag{5.46}$$

$$q_{ij}^X(\tau) = q_{ij}^X(\tau-1) - \xi\eta^{Xp}\frac{\partial \Delta^{Xp}(t)}{\partial q_{ij}^X(t)} + \gamma^{Xp}[q_{ij}^X(\tau-1) - q_{ij}^X(\tau-2)] \tag{5.47}$$

$$s_{ij}^X(\tau) = s_{ij}^X(\tau-1) - \eta^{Xp}\frac{\partial \Delta^{Xp}(t)}{\partial s_{ij}^X(t)} + \gamma^{Xp}[s_{ij}^X(\tau-1) - s_{ij}^X(\tau-2)] \tag{5.48}$$

where $p_{ij}^X \in P^X$, $q_{ij}^X \in Q^X$, $s_{ij}^X \in S^X$, η^X is the learning rate for code predictions and γ^{Xp} is the respective momentum. Again, gradients for individual weights are applied after random delays. All respective weights are initialized by a normal distribution with mean 0 and variance 0.1.

As explained, the learning rule identifies predictive temporal features from all learned codes and the classes. Given that the codes can be predicted also without considering the class, it is not guaranteed that the model will use the class during learning, particularly considering that the class dimensionality is much lower than the logarithmic code history dimensionality. To avoid that the model loses its capability to bias the prediction given a class, a multiplier ξ is applied to the learning rate of the class biases. Given that $\xi > 1$, the learning algorithm will prefer to use the class over the last codes for the minimization of code prediction errors.

Predicting the codes instead of the input stimulus effectively reduces the number of free parameters, and still the corresponding input stimuli can be inferred via the generative weights $(O^X)'$ from the code fusion vector $\mathbf{f}^X$. In contrast to Long-Short-Term-Memories (Hochreiter and Schmidhuber, 1997), which specify the temporal horizon implicitly by unfolding the network a specific number of steps through time, this approach explicitly defines the temporal horizon. Note that here, no recurrences or delays have to be resolved for training the prediction mechanisms, since the error signals are not propagated (through time) across the autoencoders.

However, the predictive connectivity scheme results in feed-forward recurrences on multiple, hierarchical levels. First, there is a submodal recurrence inside of each autoencoder, such that the last activated codes influence the currently predicted codes locally, offering the possibility to learn self-consistent simulations of submodal codes. Second, there are lateral recurrences across submodalities, for example, from the visual motion direction autoencoder to the visual posture autoencoder. Applying crossmodal predictions in this manner effectively reduces the problem that predictions could get stuck, because they induce implicit causal dependencies in the predictions. Without these intramodal recurrences, predictions could stagnate when a code is activated for a longer period of time: As a descriptive example, a limb may move into one direction for a longer time before it changes its direction rather abruptly, making it hard to predict the actual turning point of the limb solely from the history of motion direction codes. The current posture, however, is constantly changing, and can finally indicate a change in the motion direction at a certain position. Third, there are lateral, crossmodal recurrences from the visual module to the proprioceptive module and vice versa, which implement crossmodal, causal inferences in a similar manner. Fourth and last, there is a recurrent, top-down influence of the action intention. The action intention is activated by the (simulated or observed) codes themselves, and in turn

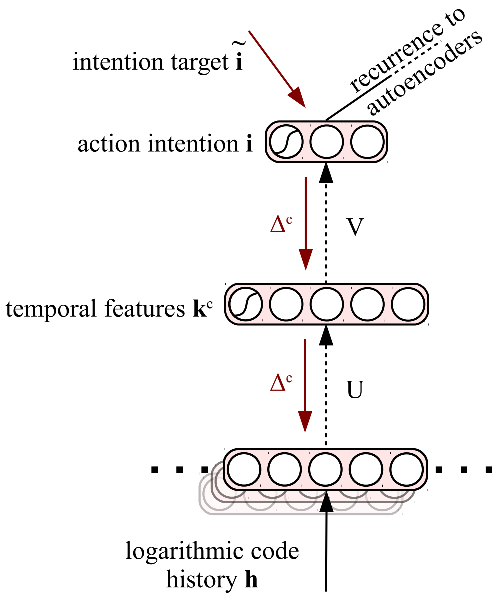

intention target $\tilde{\mathbf{i}}$
recurrence to autoencoders
action intention $\mathbf{i}$
Δ^c
V
temporal features $\mathbf{k}^c$
Δ^c
U
logarithmic code history $\mathbf{h}$

FIGURE 5.11: The temporal action classifier of the model. Again, a logarithmic time window of all submodal spatial codes is used fro classification. The module generates recurrent top-down action class inferences that bias the simulation of submodal codes. Apart from the input representation, the approach is technically equivalent to a multilayer perception.

influences the submodal code predictions in subsequent time steps. How these motion intentions are learned is described in the following.

5.5 Action Inference and Top-down Biasing: A Temporal Classifier

The submodal spatial Gestalt codes are classified by the top-most module of the network, which in turn biases the prediction of submodal codes. Here, the task of the module is to identify the kinematic intention from the observed or simulated codes. To some extent, this can be compared to frontal encodings of the mirror neuron system (see Section 2.2.3). Analogously, however, the module could also infer other properties of the observed action like for example emotional states.

The classifier network is shown in Figure 5.11. Analogously to the prediction of codes in the autoencoders, first, non-linear temporal features are extracted from the logarithmic history of activated submodal codes via a matrix U

$$\mathbf{k}^c(t) = \tanh\left(U(t)\mathbf{h}(t)\right) \tag{5.49}$$

The resulting features are then again mapped non-linearly via a matrix V onto the class or intention $\mathbf{i}$:

$$\mathbf{i}(t) = \frac{1}{1 + \exp\left(V(t)\mathbf{k}^c(t)\right)} \tag{5.50}$$

The dimension of the intention vector reflects the number of classes observed during training, and each neuron's activity $\in (0, 1)$ represents the recognition of a class respectively. Note that the class is not low-pass filtered here (but in the autoencoders), such that a precise momentary error signal can be obtained for training. The parameters of the classifier are learned in analogy to the code predictions:

$$\Delta^c(t) = {}^1\!/\!_2 \left\| \tilde{\mathbf{i}}(t) - \mathbf{i}(t) \right\|^2 \tag{5.51}$$

$$u_{ij}(\tau) = u_{ij}(\tau - 1) - \eta^i \frac{\partial \Delta^c(t)}{\partial u_{ij}(t)} + \gamma^i [u_{ij}(\tau - 1) - u_{ij}(\tau - 2)] \tag{5.52}$$

$$v_{ij}(\tau) = v_{ij}(\tau - 1) - \eta^i \frac{\partial \Delta^c(t)}{\partial v_{ij}(t)} + \gamma^i [v_{ij}(\tau - 1) - v_{ij}(\tau - 2)] \tag{5.53}$$

where $u_{ij} \in U$, $v_{ij} \in V$, η^i is the learning rate and γ^i is the momentum. In contrast to all other learning and adapting parameters, the classes are supervised by a teaching signal $\tilde{\mathbf{i}}$ which provides the currently observed motion class to the model during training.

When the classifier has been completely trained, it can for example detect postural features, or sequences of directional motion indirectly via the Gestalt codes to infer the type of action. Depending on which stimulus modalities are available, this class can also help to infer one modality from another by biasing the predictions in the respective autoencoders, or it can be used bias all predictions for simulation of a particular action when no stimulus is available. Thus, the intention module can be seen as providing the driving signal that activates multimodal, embodied, intention-specific simulations. Amongst others, the top-down influence of the classes on the code predictions is evaluated in the experiments, which are introduced in the following.

Chapter 6

Experimental Results

In the following chapters, I will evaluate the proposed action understanding model in several respects. First, I will introduce the properties and format of the motion capture stimuli used for the model evaluations in Section 6.1. I will continue with explaining the topological parameters of the network in Section 6.2. The subsequent chapters concern the learning performance of the different network components, as well as the validation of the learned encodings and adaptive components in several scenarios. In this approach, learning and action observation are consecutive, assuming complete information during learning, and incomplete as well as imperfect information during observation. The procedure for training the network is set forth in Section 6.3. I evaluate how spatial generative submodal codes are learned from self-perceptions in Section 6.3.1, how temporal code predictions are learned subsequently in Section 6.3.2, and how in parallel the classification of action types develops from the codes in Section 6.3.3. In the following action understanding experiments, first I will systematically evaluate the visuo-spatial and perceptual inference abilities of the network in Section 6.4. Then, I will show how the model is able to selectively infer and simulate submodal perceptions and intention codes to understand observed actions in 6.5. Section 6.6 demonstrates the embodied simulation capabilities of the network and shows how self-sustaining, multimodal simulations can be primed by observations or imagined selectively without sensory stimulation given a constant top-down bias. Finally, in Section 6.7, the model is evaluated when facing completely new actions, and in tracking their orientation over time.

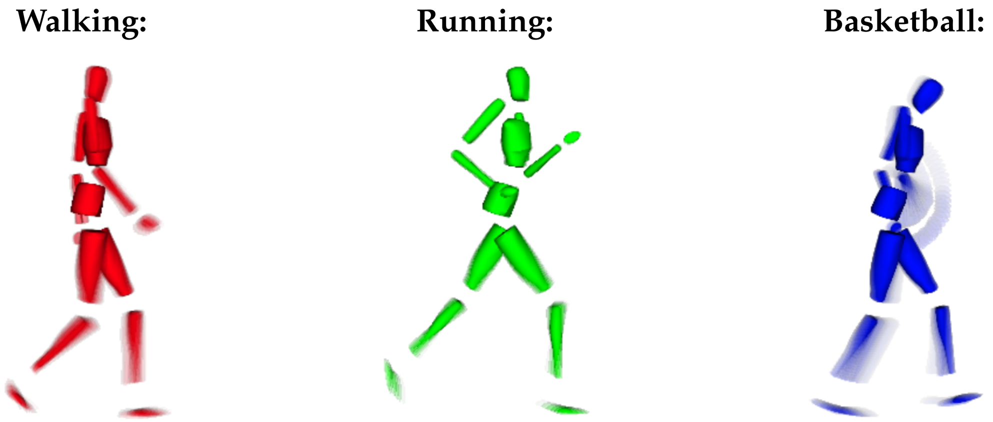


FIGURE 6.1: Example snapshots of the stimuli used for training and testing the model. Three short and idealized motion captures are used for training, while altogether 15 long and more complex captures are used for testing.

6.1 Experiment Stimuli

The model is evaluated on the basis of motion captures of the Carnegie Mellon University (CMU) Graphics Lab Motion Capture (MoCap) Database[1]. The CMU database was chosen since it is one of the most prominent public motion capture databases, and thus a number of tools is available and the results of this thesis may potentially be compared to other approaches. Furthermore, most other databases are less comprehensive, or focus on the emotional properties of an actor, social interactions, or similar.

The CMU motion tracking data was recorded with 12 high-resolution infra-red cameras at 120 Hz using 41 tracking markers attached to human subjects. Each of the tracking markers provided a 3D bodily landmark position, which was then mapped to a skeleton template that defines limb lengths and a limb hierarchy. For the evaluations in this thesis, the recorded 3D landmark positions of all subjects were matched to a simplified version (several bones were manually removed) of the skeleton of subject 9, such that limb lengths of all motion capture recordings were normalized. However, skeletal normalization is not strictly necessary for the model since it partially applies information that is invariant

[1] see http://mocap.cs.cmu.edu/ as of 08.01.2018

to variances in body morphology (cf. Schrodt and Butz, 2014). The implementation for reading out, processing, and displaying the motion capture data was based on *AMC-Viewer* by Jim McCann[2].

TABLE 6.1: CMU MoCap database trials chosen for training and testing the action understanding model.

Subject	**Trial**	**Length (time steps)**	**Description**
Training			
35	7	reduced to 260	walking
9	3	reduced to 92	running
6	2	reduced to 115	basketball dribble, right-handed
Testing 1		(similar)	
5	1	598	walking
6	1	494	walking
10	4	549	walking
12	1	523	walking
2	3	173	running
16	46	136	running
35	19	160	running
35	22	167	running
6	2	721	basketball dribble, right-handed
6	3	527	basketball dribble, right-handed
6	4	396	basketball dribble, **left-handed**
6	5	385	basketball dribble, right-handed
Testing 2		(dissimilar)	
55	2	2180	Lambada dance
49	3	1504	Jumping up and down
40	11	6020	Waiting for the bus

Recordings from subjects performing three different movements (*walking*, *running* and *basketball dribbling*) were utilized for training, as exemplified in Figure 6.1. These types of movements have been selected for training and most of the evaluations for several reasons: As explained before, this thesis does not consider action goals or modal end states, but rather continuous, kinematic actions, and consequently, continuous, cyclic motion captures were used. Only a handful of the motion captures in the CMU database are loopable straightforward, including the selected ones. Furthermore, although the CMU database provides numerous different actions and respective descriptions for them, there are only a few, cyclic actions classes for which multiple comparable examples,

[2] see http://www.cs.cmu.edu/~jmccann/ as of 08.01.2018

performed by multiple actors exist. There is also no annotation but only descriptions of the trials themselves, and variances across trials with similar description are partially substantial. Nonetheless, the choice of data allowed to train the model on a single, presumably typical trial per action class, and to validate the learned encodings given a number of trials obtained from other subjects.

The exact selection of all trials and subjects from the CMU database can be seen in Table 6.1. As shown, 467 frames (3.9 seconds) of training data faced 4829 frames (40.2 seconds) of test data in the same classes, and beyond that, 9704 frames (80.9 seconds) of test data in other classes in the experiments, resulting in a ratio of 1:31.12. While typically, neural networks are trained on relatively extensive training sets, and validated on smaller test sets, the idea behind the approach pursued here is to replicate the typically greater variance of potentially observable movements in comparison to the personally executable movements in real world situations.

The type description of a movement trial provides the target signal $\tilde{\mathbf{i}}$ for learning the classification. For training, a short episode of each trial was selected and manually edited to form a continuous cycle. Thus, the model was able to learn cyclic action attractors without terminal state. Four other trials of each class remained untouched and served as *similar* (same class) testing and validation set for the model. Where possible, the test trials were performed by other subjects than the training trials. For basketball trials, only one subject was available in the database. However, the trials of the same subject differ significantly in this case. For example, one of the basketball dribblings was performed left-handed, while the others were performed right-handed. Additionally, a lambada dancing trial, a jumping trial, as well as a waiting-for-the-bus trial were selected for evaluation of the model during the observation of *dissimilar* movements. These trials are used exclusively in the last experiment in Section 6.7 to evaluate the model's capability to understand unknown actions.

Every second frame of a trial was used as input to the model while starting with a random (odd or even) frame number, effectively splitting up and doubling the data by parity. From the 30 limbs provided by the skeleton template, $M^v = 15$ starting- or end-point positions were selected as visual inputs, while $M^p = 16$ limb orientations were selected as proprioceptive inputs, as shown in

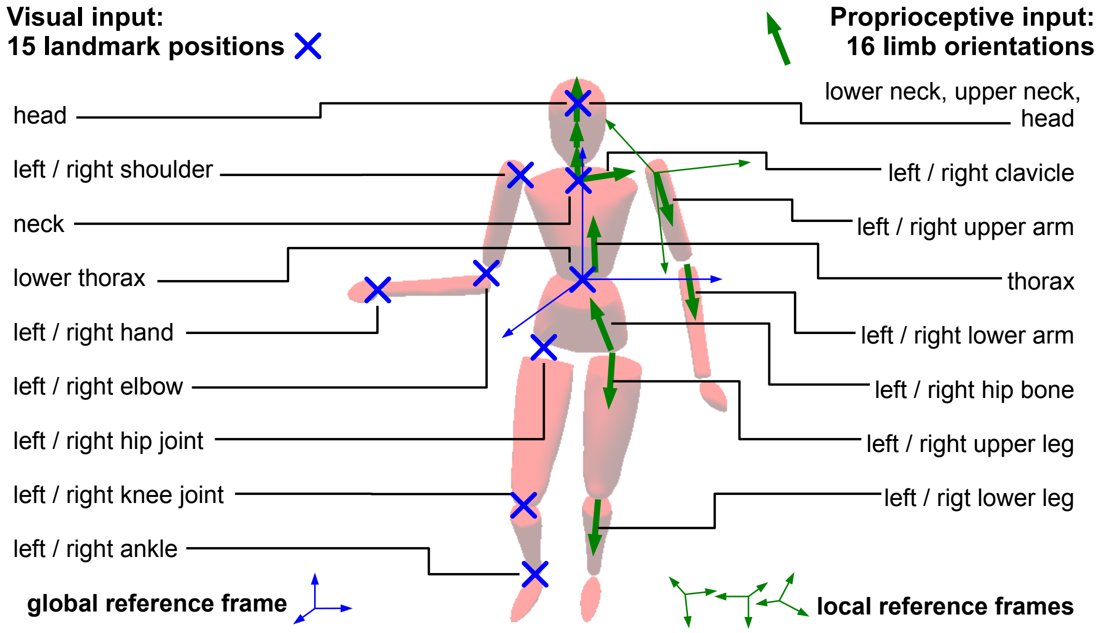


FIGURE 6.2: An overview of visual and proprioceptive inputs to the model. Visual inputs represent 3D bodily landmark positions (and possibly distractor coordinates) in a *global* frame of reference provided by the lower thorax. Proprioceptive inputs correspond to 3D orientations of limbs in a *local* frame of reference provided by the predecessor in the body hierarchy. In the experiments, 15 visual inputs as well as 16 proprioceptive inputs have been selected manually from the data.

Figure 6.2. In the selected data and its interpretation, subjects move on the x-z-plane (ground) into z direction, while y is directed upwards and x is directed to the right of the actor's movement direction. Visual inputs to the model were adjusted for the *per-frame* x-z-coordinate, as well as the *average* y-coordinate of the skeleton's root (i.e. lower thorax). As a consequence, the model did not observe locomotion on the floor, but it did observe ground-relative upwards/downwards dynamics of all bodily landmarks. The orientation of the visual inputs remained untouched. Proprioceptive inputs are relative to their local predecessor in the body hierarchy and thus their coordinates were not altered altogether. Finally, all inputs to the network were exponentially smoothed (with a smoothing factor of 0.9) to account for noise in the recorded positions, and to extrapolate transitions between movement trials presented in succession.

The visual inputs can potentially be transformed globally via a rotation offset matrix A^{data}, and a translation offset vector $\mathbf{b}^{\text{data}}$. Furthermore, the visual features can potentially be provided in arbitrary order to test the adaptive feature binding capabilities of the network. As well, an arbitrary number of distractors can be added to the inputs. Proprioceptive features are not transformed, permuted, or filled up with distractors. They are either fully available (e.g. during training), or completely absent to test the simulation capabilities of the model. In the following, I will specify the topological parameters of the network chosen for the experiments.

6.2 Topological Network Parameters

All of the results in this thesis were obtained using networks with the same architectural parameters, which are shown in Table 6.2, and which are explained in the following. Parameters that refer to the adaptive perceptual and learning components, the sensation of specific modalities, or teacher signals are mentioned separately. If not stated otherwise, evaluations are generic and refer to a single network instance that highlights the learning and perceptual characteristics of the network in an appropriate way. The replicability of these results was confirmed using multiple networks. Furthermore, statistics of multiple, independently trained network instances are raised.

The visual pathway of the network is configured to process up to 30 global 3D input coordinates, thus consisting of 30 of the visual processing paths shown in Figure 5.4 from Section 5.3. 15 of the input features come from the selected bodily features defined above. The others are either disabled or filled up with distractor features, depending on the experiment. The neural information extraction of each visual feature path results in three populations for the position, motion direction and motion magnitude of the respective coordinate. Following the specification of the motion capture inputs, the proprioceptive pathway is configured to receive 16 local 3D orientations of body limbs as defined in the last section.

As mentioned earlier in Chapter 5.3.3, methods are applied to avoid catastrophic forgetting of motion encodings during training. Discontinuities are

TABLE 6.2: Architectural parameters chosen for the experiments.

Parameter	Description
$N^v = 30$	Maximum number of visual inputs
$M^v = 15$	Number of visual bodily features
$N^p = 16$	Number of proprioceptive features
64	Dimension of visual coordinate populations
27	Dimension of visual motion direction populations
8	Dimension of visual motion magn. populations
27	Dimension of proprioceptive posture populations
27	Dimension of proprioceptive motion direction populations
8	Dimension of proprioceptive motion magn. populations
$\zeta^{vp} = 0.2$	Continuity of visual coordinate populations
$\zeta^{vd} = 1$	Continuity of visual motion direction populations
$\zeta^{vm} = 0.3$	Continuity of visual motion magn. populations
$\zeta^{pp} = 0.5$	Continuity of proprioceptive posture populations
$\zeta^{pd} = 1$	Continuity of proprioceptive motion direction populations
$\zeta^{pm} = 0.3$	Continuity of proprioceptive motion magn. populations
$[-103.6, 71.4]$	Stimuli range of visual coordinate population (cm)
$[-1, 1]$	Stimuli range of visual motion direction population
$[0, 6.16]$	Stimuli range of visual motion magn. population (cm)
$[-1, 1]$	Stimuli range of proprioceptive posture population
$[-1, 1]$	Stimuli range of proprioceptive motion direction population
$[0, 0.29]$	Stimuli range of proprioceptive motion magn. population
40	Dimension of spatial codes
40	Dimension of temporal code features
$\theta = 6$	Temporal horizon for logarithmic code history
$o = 0.01$	Motion direction recognition threshold

induced in the autoencoders' inputs by specifically parameterized population coding. Particularly in the domain of visual positional inputs, the difference between coordinates of two consecutive time steps is rather marginal, and the configuration space of each single feature is typically limited to a small subspace of the range covered by the representing population, resulting in even less discontinuity. By contrast, the motion direction and magnitude entail some intrinsic discontinuities (e.g. the direction of arm and leg motion is changing suddenly at extreme postures). Accordingly, the *resolutions* and *continuity factors* of submodal populations are set up to support discontinuity and nonetheless maintain the precision of the encoded information: Visual coordinate populations are configured to contain 64 neurons (4 centers on each dimension) with a continuity factor of 0.2, while proprioceptive limb orientation populations contain 27 neurons (3 per dimension) with a continuity of 0.5. All directional motion populations are configured to contain 27 neurons and a continuity factor of 1. The centers of all populations representing limb orientations or motion directions were projected to unit spheres to cover the input space optimally. One dimensional magnitude populations are configured to contain 8 neurons with a continuity factor of 0.3.

The *ranges* over which the centers of the respective submodal and feature specific population neurons were distributed were determined heuristically according to the motion capture data and the applied skeleton. In particular, the visual receptive field was set to cover just about the height of the modified skeleton in arbitrary orientations, resulting in 175 cm^3, and the perceived motion magnitude of visual features was limited to 6.16 cm per time step, which equals about 3.7 m/s.

The precision with which spatial codes and temporal code predictions can be learned also depends on the dimensionality of the representing layers. For all experiments in this thesis, all spatial codes as well as the temporal features extracted from them (in the autoencoders as well as the classifier network) were configured to be 40 dimensional. The temporal features were extracted from a logarithmic code history with $\theta = 6$, taking into account the last 63 time steps. How networks that use these topological parameters are trained is detailed in the following.

6.3 Training Evaluations

The goal of the training procedure is (i) to develop generative spatial encodings of embodied, visual and proprioceptive self-representations of actions, (ii) to learn distinctions in the identity of the actions, and (iii) to form stable, predictive attractors of multimodal state sequences for the encountered actions that allow for inferences and embodied simulations.

Training is embodied in the sense that the model has unobstructed access to visual and proprioceptive information, meaning that proprioceptive stimuli are

TABLE 6.3: Setup of the network components during embodied training.

Parameter	Description
$\eta^s = 0$	Adaptation rate of the origin of the visual FOR
$\gamma^s = 0$	Momentum of the adaptation of the origin
$\eta^r = 0$	Adaptation rate of the orientation of the visual FOR
$\gamma^r = 0$	Momentum of the adaptation of the orientation
$\eta^w = 0$	Adaptation rate of the feature selection and assignment
$\gamma^w = 0$	Momentum of the feature selection and assignment
$\mathbf{b}^{\text{data}} =$ egocentric	Data translation offset
$A^{\text{data}} =$ egocentric	Data rotation offset
$w_{ij} =$ provided	Feature binding weights
$\tilde{\mathbf{i}} =$ provided	Classification target
$\kappa = 0.9$	Intention dynamics parameter
$q^{vp/d/m} = 1$	Visual stimulus reliabilities
$q^{pp/d/m} = 1$	Proprioceptive stimulus reliabilities
$\Theta = 1500$	Horizon for applying learning weight updates

available, and that visual stimuli are perceived from an egocentric view point (i.e. $\mathbf{b}^{\text{data}}$ and A^{data} are neutral elements with respect to the reference frame of the motion capture data). Consequently, the adaptation of the model's internal visual frame of reference is disabled during training. Analogously, the model has full access to the identity and grouping of the visually perceived features (represented by the feature binding weights w_{ij}). Thus, the adaptive feature binding is disabled as well. The training procedure furthermore assumes that intention states are available during self-observation and learning. The relevant model parameters for the overall training procedure are shown in Table 6.3.

As explained in Chapter 5.4.2, the classification of actions as well as the prediction of their progress is based on the learned spatial Gestalt codes. That is, if the weights that represent Gestalt codes are modified during training, also the weights that represent predictions and classifications would have to be modified, since their target mapping is changing. To avoid that these aspects interfere, training is subdivided into two phases: In the first training phase, spatial codes are learned from visual and proprioceptive inputs. This phase is evaluated in Chapter 6.3.1. In the second phase, the network learns attractor state sequences as well as classifications from the learned spatial codes, which is evaluated in Chapters 6.3.2 and 6.3.3, respectively. Specific parameters for these phases of training are detailed separately in the following evaluations.

6.3.1 Embodied Learning of Submodal Spatial Codes

TABLE 6.4: Parameters for learning and adaptation chosen for spatial training.

Parameter	Description
$\eta^{vpc} = 0.0005$	Learning rate of visual posture codes
$\eta^{vdc} = 0.001$	Learning rate of visual motion direction codes
$\eta^{vmc} = 0.01$	Learning rate of visual motion magnitude codes
$\eta^{ppc} = 0.0005$	Learning rate of proprioceptive posture codes
$\eta^{pdc} = 0.0005$	Learning rate of proprioceptive motion direction codes
$\eta^{pmc} = 0.0005$	Learning rate of proprioceptive motion magnitude codes
$\gamma^{Xc} = 0.9$	Momentum of learning spatial codes
$\eta^{Xp} = 0$	Learning rate of temporal code predictions
$\gamma^{Xp} = 0$	Momentum of learning temporal code predictions
$\xi = 5$	Learning rate multiplier for class biasing
$\eta^{i} = 0$	Learning rate of spatial code classifications
$\gamma^{i} = 0$	Momentum of learning spatial code classifications

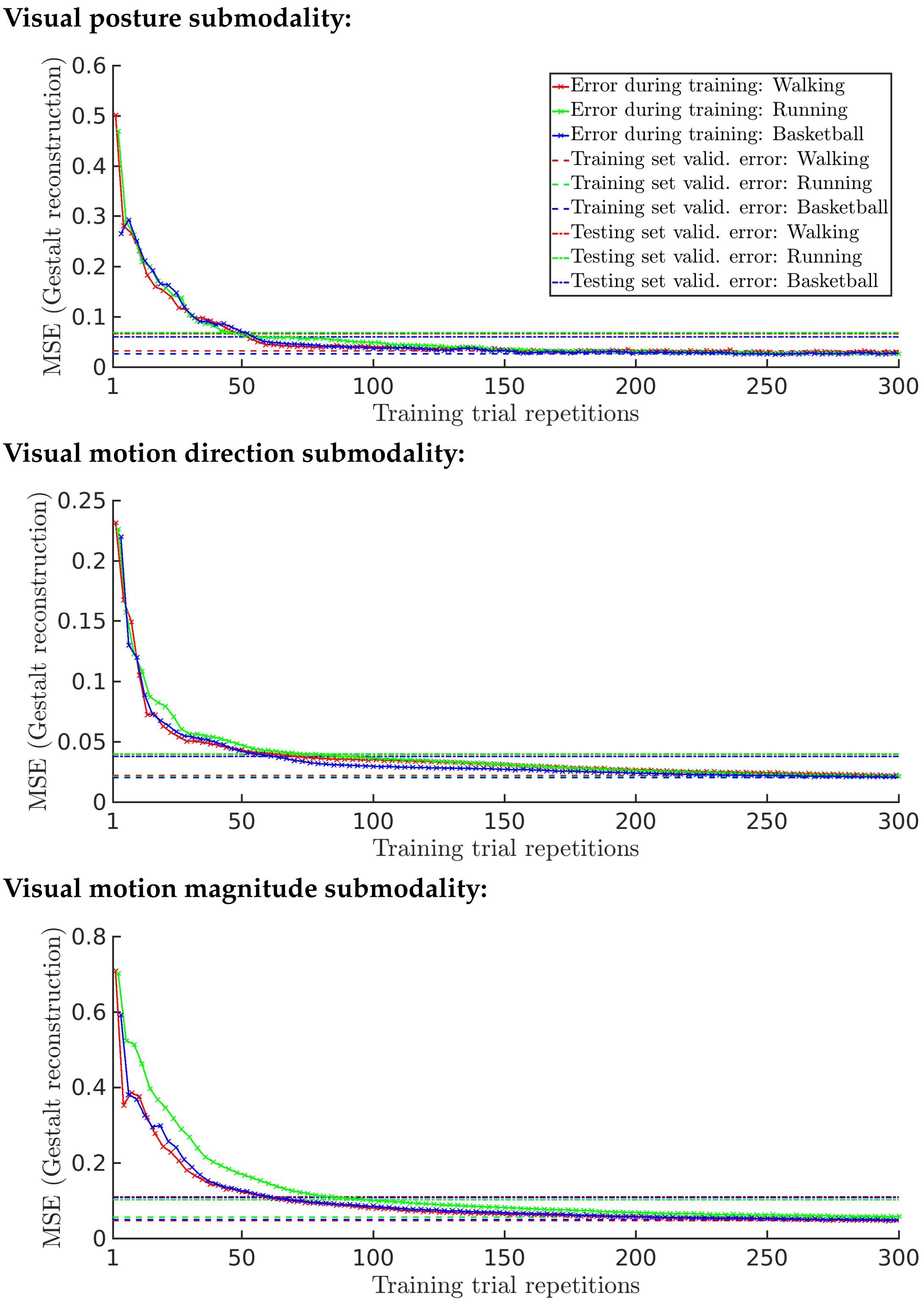
In this experiment phase, spatial Gestalt codes in the six submodalities are learned by adapting the weight matrices O^X (see Figure 5.9 from Section 5.4.1) based on the reconstruction errors in the respective autoencoder modules. No error signals are propagated between the autoencoders, such that the codes develop independently from each other, yet representing different aspects of the same data. Table 6.4 shows an overview of the parameters chosen for this phase of training. Learning rates and momentum were determined heuristically, taking into account the different continuities of the submodal encodings. As shown in the table, the learning of temporal predictions and motion classes is disabled.

Each of the three training trials is presented to the model for 500 consecutive time steps, starting at a random time step of the respective trial. The movements are presented in fixed order, and including random, approximated transitions between them. Altogether, each of the 3 training movements is shown

100 times, resulting in 150000 time steps of spatial training. After the training procedures, all training and test trials are presented to the model without weight updates to test for generalization and recoding.

Figure 6.3 and Figure 6.4 show that the Gestalt reconstruction errors Δ^X for all submodal autoencoders ($X \in \{\mathrm{vp}, \mathrm{vd}, \mathrm{vm}, \mathrm{pp}, \mathrm{pd}, \mathrm{pm}\}$) gradually decreased during spatial training. As also shown in the figures, the network did not tend to recoding, as the average error of a trial during training was comparable to the average error of the same trial when learning was disabled. Similarly, the network did not tend to strong overfitting to the training set, as the errors between training and test trials were sufficiently comparable. Furthermore, the error signals were comparable across the different types of movements and different network instances. However, the reconstruction errors of the autoencoders converged to somewhat different absolute levels in comparison across the submodalities. Nonetheless, the developing encodings seem sufficiently balanced, and, for perceptual adaptations, differences in error levels can be compensated by error weighting, as described in Section 5.3.

Figure 6.5 shows a qualitative evaluation of the learning progress: Each submodal autoencoder first quickly learns to reconstruct an average over time of its population encoded inputs. In this early phase of training, changes in the reconstructed population activity mostly represent adaptations of this average, rather than different parts of the observed actions. Thus, it can be said that each autoencoder first learns only a single code. For a clearer and more comprehensive visualization of the model's expectations, a body display can be decoded from the proprioceptive submodal population activity reconstructions, and a point light display can be decoded from the visual reconstructions (see Section 5.3.3), which was done for Figure 6.6. Consistently, using the underdeveloped model expectations, hardly any rhythmic motion can be seen in the body and point light displays. In fact, motion is again predominantly the result of changes in the learned average of visual and proprioceptive body postures, motion directions, and motion magnitudes for the three movement classes trained on. Interestingly, as exploratory evaluations revealed, these average expectations are already sufficient to bootstrap perspective-taking to a certain degree. They are approximately represented by the code layers' biases $\mathbf{b}^X$, which represent submodal, general action templates.

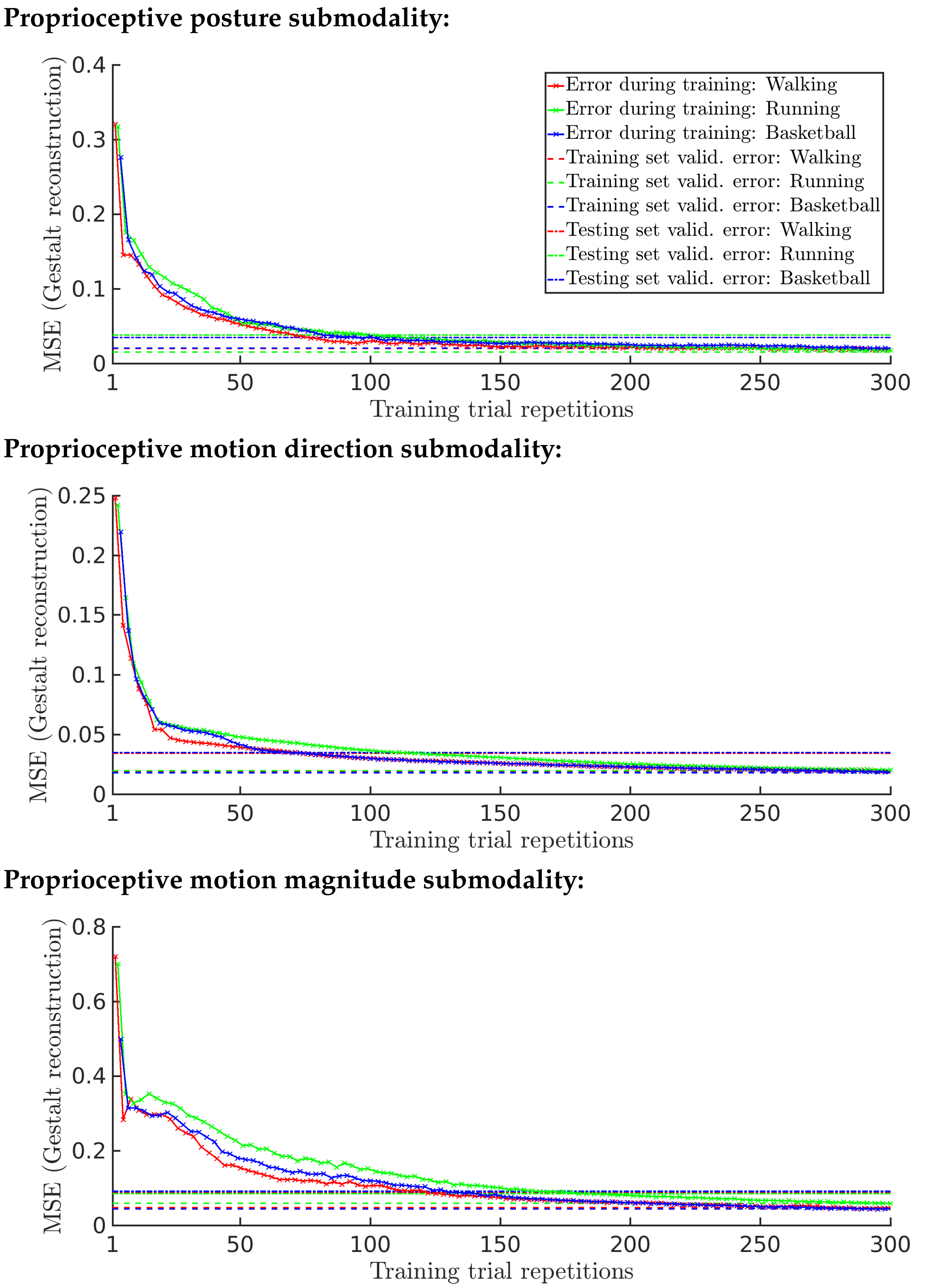


Visual posture submodality:

Visual motion direction submodality:

Visual motion magnitude submodality:

FIGURE 6.3: Visual spatial reconstruction errors of posture, direction, and magnitude Gestalt perceptions for the three different actions during training. The results were obtained by averaging over four independently trained network instances.



Proprioceptive posture submodality:

Proprioceptive motion direction submodality:

Proprioceptive motion magnitude submodality:

FIGURE 6.4: Proprioceptive spatial reconstruction errors of posture, direction, and magnitude Gestalt perceptions for the three different actions during training. The results were obtained by averaging over four independently trained network instances.

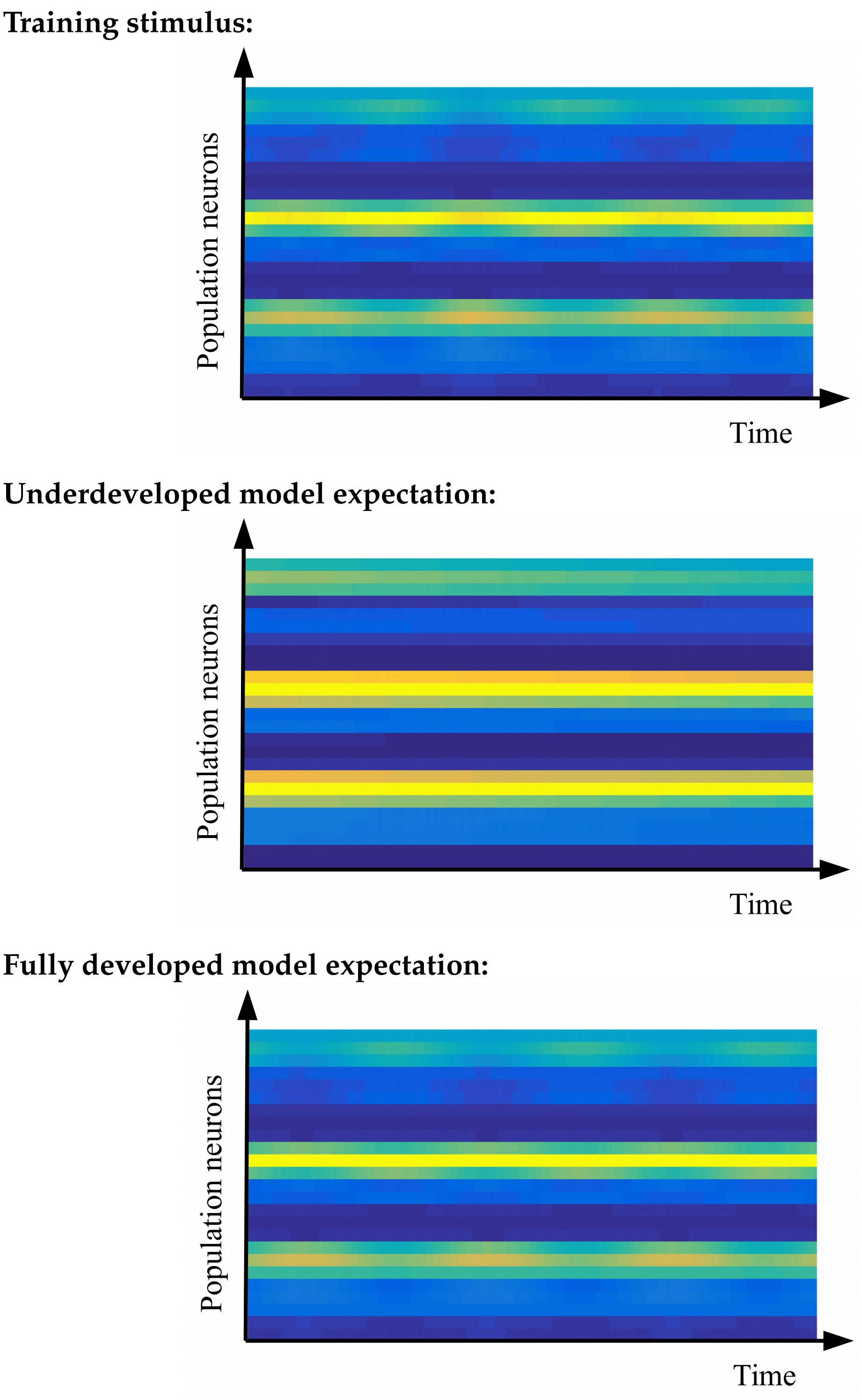


FIGURE 6.5: Qualitative comparison between a population encoded stimulus, a rudimentarily trained stimulus reconstruction, and a fully trained stimulus reconstruction. The plots show an example of the color coded activity of a single, proprioceptive, postural feature population over 200 time steps.

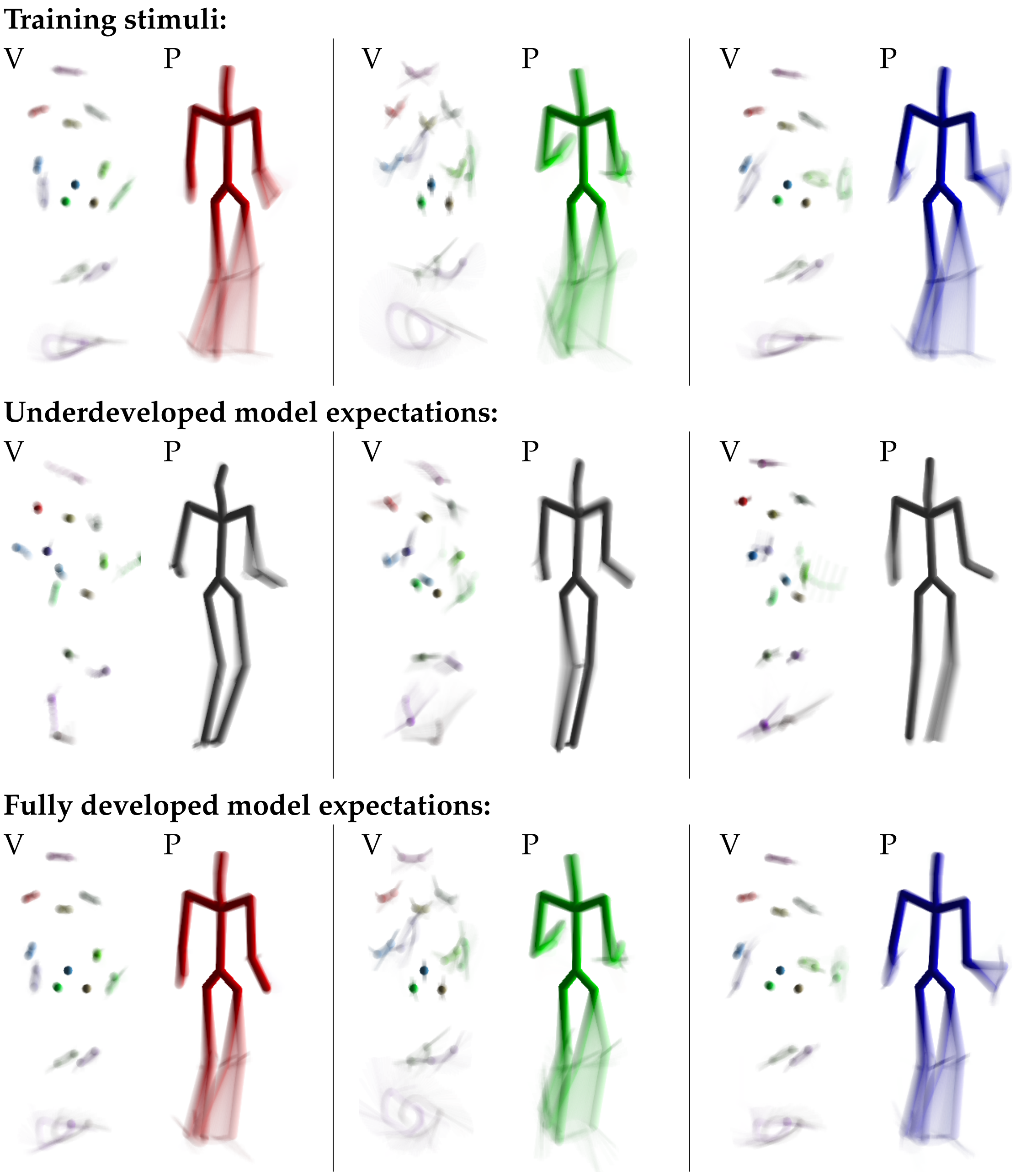

Training stimuli:
V
P
V
P
V
P
Underdeveloped model expectations:
V
P
V
P
V
P
Fully developed model expectations:
V
P
V
P
V
P

FIGURE 6.6: Qualitative comparison between visual (V) and proprioceptive (P) stimulus displays (first row), rudimentarily trained reconstruction displays (second row), and fully trained reconstruction displays (third row). The plot shows the average stimulus/expectation of body postures (point-lights for V and stick figures for P) and feature motion (indicated by lines and their length). The provided/inferred motion class is color coded in the stick figures (red for walking, green for running, and blue for basketball). No classes have been learned for the underdeveloped expectations (thus gray stick figures), which represent average submodal perceptions. Fully developed expectations encode the class as well as cyclic body motion.

Later during training, deviations from the average input push the model towards developing distinct codes for the segments of the movements. This is particularly hard for the model in the postural modalities, where deviations from the average are relatively marginal in terms of population encoded activity. Since the code layer of an autoencoder learns codes for all of the features jointly, discontinuities in a single feature input have an influence on learning the reconstruction of other features. This mutual influence can be seen as a positive effect for several reasons. For example, when the (cyclic) movement of multiple bodily features is correlated, but only one of them deviates strongly from the average, learning distinct codes is driven mainly by this feature. However, once distinct codes are developed, they can more easily incorporate the slight variabilities of correlated features. In a similar manner, the above effect applies also to learning multiple movement classes in the same code layer. In exploratory experiments, spatial training on three movements was more accurate than training on a single movement, given the same training time per class. Thus, correlations between features and action classes result in accelerated training.

Furthermore, learning whole-body Gestalt codes has advantages for robust recognition: Given that for example the observed upper body is recognized precisely, but the observed lower body is not represented in the code manifold, the best matching code is activated nonetheless, effectively interpreting or complementing (in the case of missing features) the observation by means of the embodied experiences. However, since codes are activated by a linear combination of weight vectors that stem from the individual features, the codes can also potentially recognize (and generate) actions not seen during training to a certain degree by combining the weight vectors accordingly. Taken together, the linear combination and dimensionality reduction of submodal feature vectors into non-linear, whole-body Gestalt codes provides recognition robustness, as well as generalization to a certain extent.

In conclusion, the network was able to learn balanced spatial encodings for three complex, cyclic, whole body movements in multiple distributed, modal and submodal domains. Typical problems in online training of autoencoders are circumvented by the proposed training mechanisms. The learned spatial codes provide suitable perceptual biases for action recognition. Furthermore, the codes are generative in that they can be used to construct momentary expectations that cover the complete input manifold seen during training. The codes, however, do not incorporate temporal dependencies, nor are they suitable for crossmodal inference and synchronization of the separate submodalities to infer or simulate consistent, multimodal expectations over time. In the following chapter, the learning of temporally predictive codes will be described and evaluated.

6.3.2 Embodied Learning of Temporal Predictive Encodings

TABLE 6.5: Parameters for learning and adaptation chosen for temporal training.

Parameter	**Description**
$\eta^{Xc} = 0$	Learning rate of visual/proprioceptive Gestalt codes
$\gamma^{Xc} = 0$	Momentum of learning spatial codes
$\eta^{vpp} = 0.0001$	Learning rate of visual posture code predictions
$\eta^{vdp} = 0.0001$	Learning rate of visual motion dir. code predictions
$\eta^{vmp} = 0.00005$	Learning rate of visual motion magn. code predictions
$\eta^{ppp} = 0.0001$	Learning rate of propr. posture code predictions
$\eta^{pdp} = 0.0001$	Learning rate of propr. motion dir. code predictions
$\eta^{pmp} = 0.00005$	Learning rate of propr. motion magn. code predictions
$\gamma^{Xp} = 0.9$	Momentum of learning temporal code predictions
$\xi = 5$	Learning rate multiplier for class biasing
$\bar{\mathbf{i}}$ = teacher forcing	Class bias in autoencoders

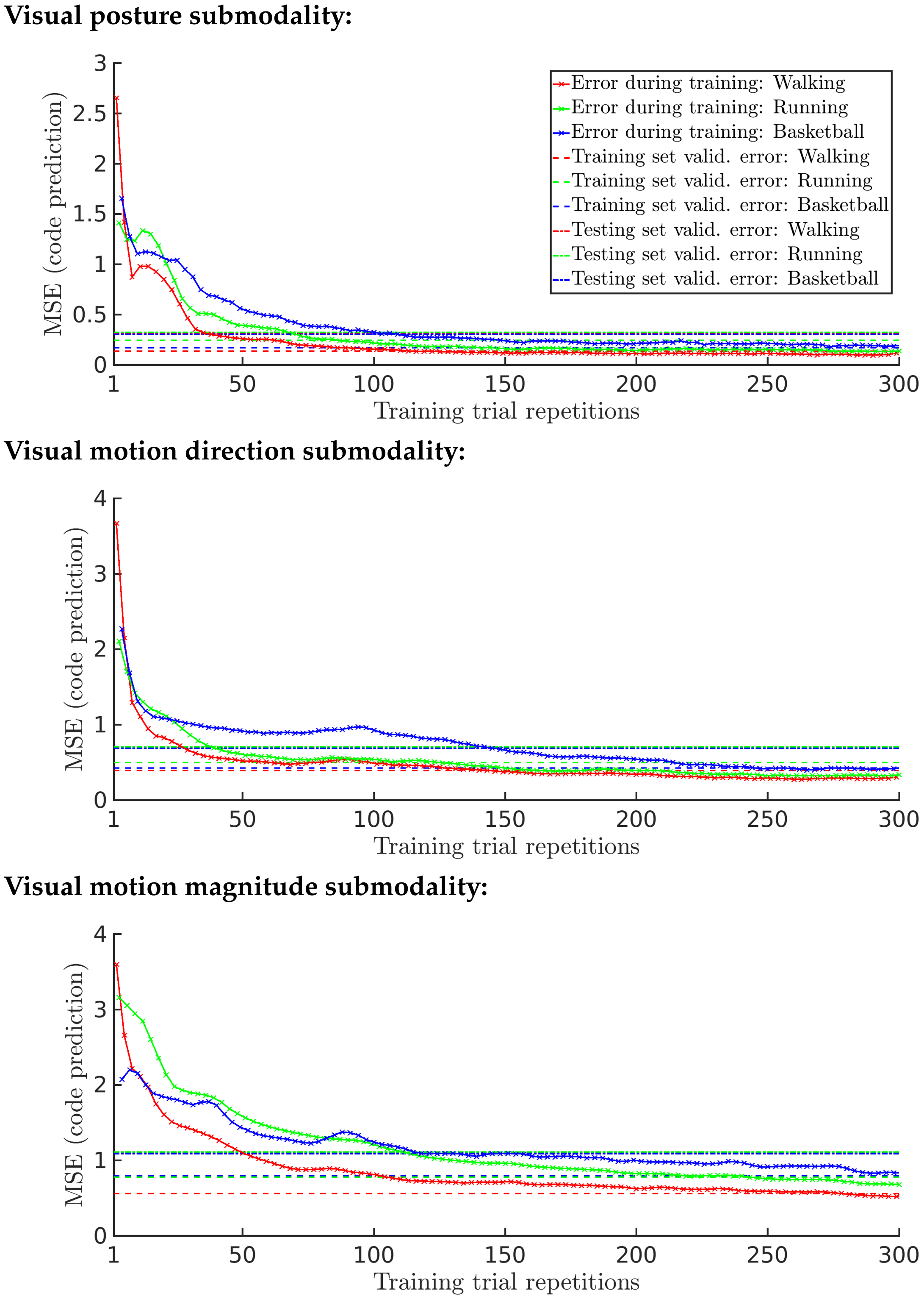
In this experiment, the training of temporal Gestalt code predictions that enable crossmodal inference and lead to distributed simulation abilities are trained and evaluated. The input sequence is the same as during spatial training. The parameters chosen for this experiment are shown in Table 6.5. Spatial training is disabled, and the learning rates and momentum for the autoencoder weight matrices P^X, Q^X, S^X were determined heuristically. The learning rate for class biasing is multiplied by $\xi = 5$ to reinforce the influence of classifications on the prediction. To ensure that the predicted codes are constantly biased by the correct classes instead of the inferred classes, which are developed in parallel (see Section 6.3.3), the top-down classification input of the autoencoders is overridden with the correct classes (cf. teacher forcing, Williams and Zipser, 1989).

Even if a sufficiently trained model classification could be used during temporal training, teacher forcing is preferable since is avoids that the autoencoders utilize the imperfections / dynamics in the classification.

Figure 6.7 and Figure 6.8 show the code prediction errors Δ^{Xp} in this phase of training. They show that in all submodalities, temporally predictive features were successfully extracted from the logarithmic history of spatial codes as well as the provided action classes to predict subsequent codes. Qualitatively, the results are mostly comparable to those during spatial training. There was, however, partially more imbalance in the error levels in comparison across the classes, partially indications of slight recoding, and again, the errors of different submodal encoders ended up at slightly different levels. Compensation is not necessary in this case, since the prediction errors are not used during testing.

During temporal training, the fused Gestalt codes were activated solely from the visual and proprioceptive input data (see $q^{vp/d/m}$ and $q^{pp/d/m}$ in Table 6.3). The error signal for learning was then generated by comparing the activated spatial code to the predicted code that uses the logarithmic history of preceded codes. As shown in the Figure 6.7 and Figure 6.8, the error converged to a specific level. However, successful convergence during temporal training does not ensure that the developed code predictions form stable and self-preserving attractors when the model is driven only by its own predictions.

Several problems can occur. First, the sequence of predicted codes may come to a standstill, given that the current predicted codes converge to the last predicted codes. Acquiring the predictions from multiple sources of information circumvents this problem as far as possible: For example, any significant motion magnitude necessarily determines that the posture (code) in the next time step differs from the last. The training data does contain bodies in motion only, thus, given that the posture code prediction uses the magnitude codes, this problem is effectively avoided. Although it is not explicitly declared which information the autoencoders use and extract their predictions from, it can be expected that they facilitate the most explanatory codes on average, thus identifying the conditional multimodal dependencies of biological motion. Second, the code prediction may diverge over time from the submodal codes encountered during spatial training due to an accumulation of prediction errors. For this reason, the autoencoders predict the *absolute* values of codes, instead of the difference to

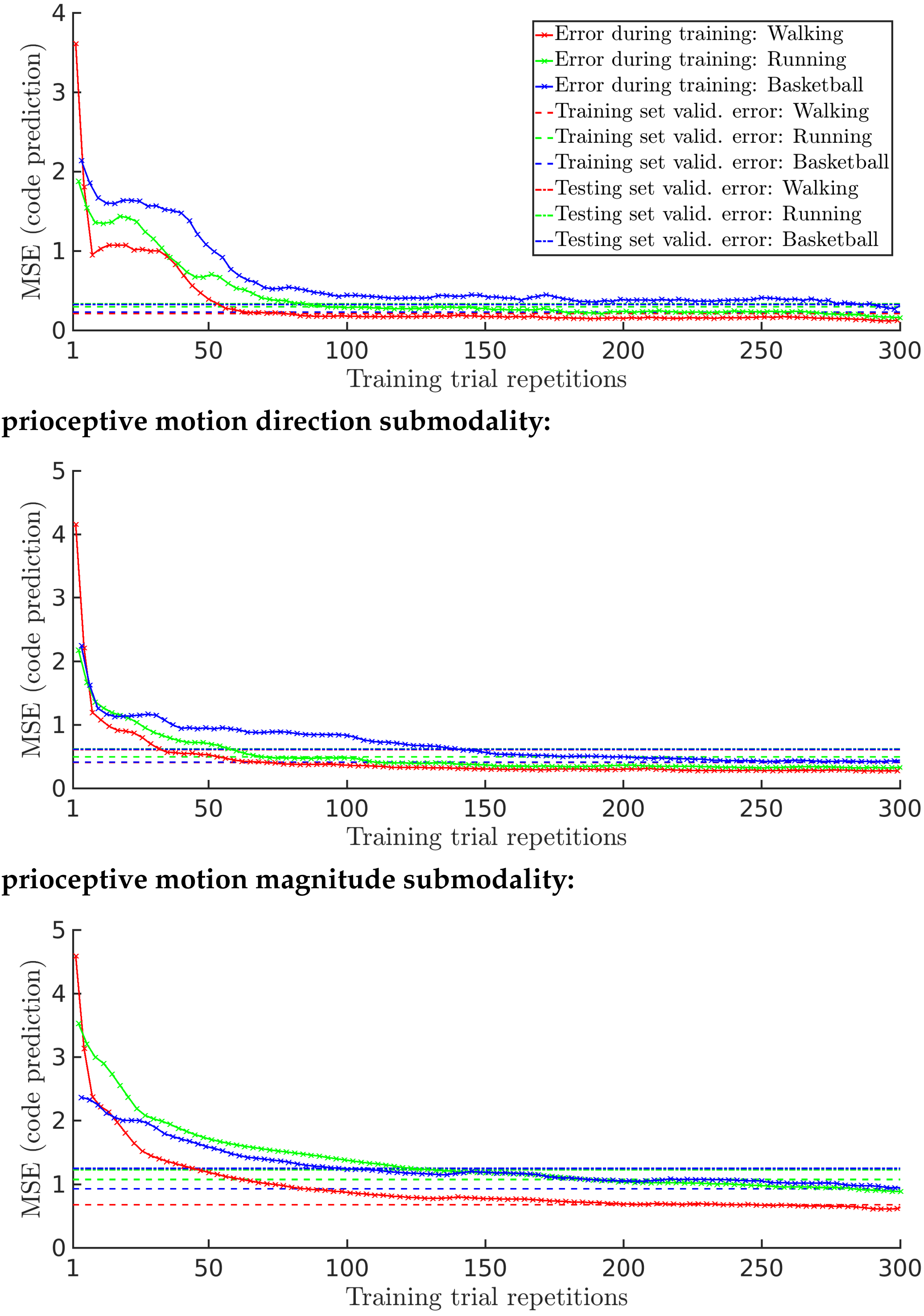


Visual posture submodality:

Visual motion direction submodality:

Visual motion magnitude submodality:

FIGURE 6.7: Visual temporal code prediction errors of the posture, direction, and magnitude Gestalt encoders for the three trained actions. The results were obtained by averaging over four independently trained network instances.



Proprioceptive posture submodality:

Proprioceptive motion direction submodality:

Proprioceptive motion magnitude submodality:

FIGURE 6.8: Proprioceptive temporal code prediction errors of the posture, direction, and magnitude Gestalt encoders for the three trained actions. The results were obtained by averaging over four independently trained network instances.

the last code. Thus, each code prediction does not assume the correctness of the previous, submodal prediction, but relies on all previous predictions and their logarithmic history. In the logarithmic history, prediction errors are smoothed out to a certain degree, further decreasing the chances of divergence. Finally, and thirdly, the non-linearities in the fused codes limit the predictions to a specific codomain by themselves, and tolerate also larger prediction errors in the outer ranges (i.e. -1 or 1). As a result, stable, precise, and self-preserving, multimodal attractor states can be learned.

Taken together, the spatial submodal encodings developed in the last chapter were successfully linked via temporal predictive encodings that extract short-term, potentially multiconditional dependencies. As will be evaluated later, the model is now able to selectively and consistently simulate submodal Gestalt perceptions.

6.3.3 Embodied Learning of Kinematic Intentions

TABLE 6.6: Parameters for learning and adaptation chosen for classifier training.

Parameter	Description
$\eta^i = 0.0004$	Learning rate of spatial code classifications
$\gamma^i = 0.9$	Momentum of learning spatial code classifications

In parallel to the identification of temporal features for code predictions, the action classifier is trained. The learning rates and momentum for training the classifier weight matrices U and V were determined heuristically as shown in Table 6.6.

Figure 6.9 shows the unnormalized, analogous output of the classifier during training. Starting from a prior classification of about $(0.5, 0.5, 0.5)$, the maximum based classification for the walking and running training trials was correct after training 35 trials (about 12 per action class). However, the basketball movement was constantly confused with the walking movement before about 80 repetitions (20 per class). This learning behavior is allegeable seeing that the basketball and the walking trial are very similar except for the right arm movement (the data does not contain context by means of a ball). Eventually, the

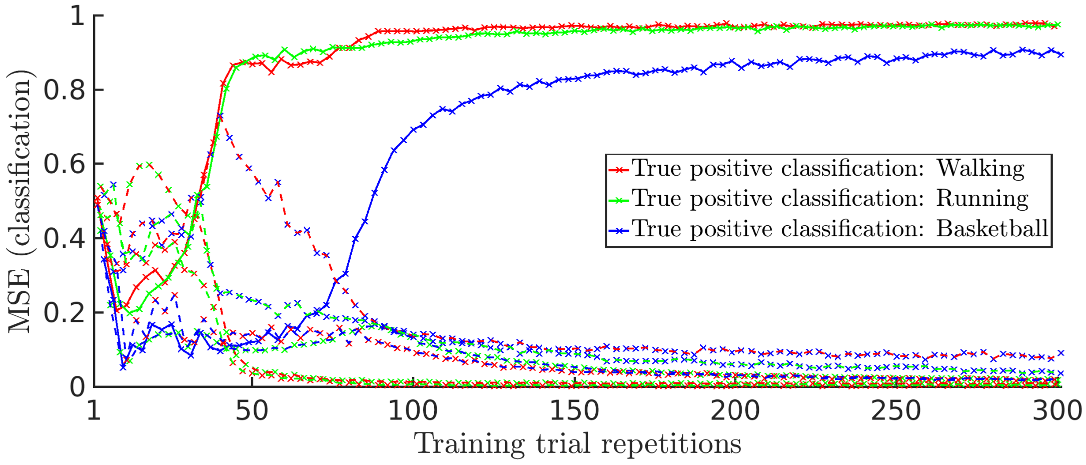


FIGURE 6.9: The classifier output during training. The color of crosses indicates which class was currently presented to the network. Dashed lines denote the output of *a false* classifier, while straight lines denote the output of *the correct* classifier. The respective class is color coded. While the true-positive classifications (crossed straight lines) improve for the walking and running classes, the basketball trial is initially interpreted as walking (dashed red line with blue crosses). The correct classification for basketball develops later during training. The results were obtained by averaging over four independently trained network instances.

network managed to figure out the discriminative features in the multimodal codes also for basketball dribbling, such that the final performance of the classifier was acceptable for all movement classes. Recoding was not present in the classifier, as the classification performance on the training trials was about equal when training was disabled.

For these results, the dimension of the Gestalt codes was most crucial. As also acknowledged in previous work (Schrodt et al., 2015), distinct classifications are substantially facilitated by the development of distinct Gestalt codes for each of the classes, which again, is facilitated by the capacity of the encoders. Since the spatial Gestalt codes were trained without supervision, the *intrinsic variances* in the activated spatial codes had to reveal (also non-linearly separable) class-distinct features. Consequently, an important factor was the use of the logarithmic code history. It allowed to consider the short-term progress of motion for classification, as was implicitly also applied in previous work, where separate motion patterns represented short sections of each movement (Schrodt et al., 2015).

Taken together, the intention module was able to extract the motion class from the developed, multimodal Gestalt codes in a supervised manner. Further evaluations will show that the classification is robust to variances in postural control, and that it extracts meaningful, submodal features for classification.

6.4 Test Evaluations 1: Visuo-Spatial Abilities and Perceptual Inference

TABLE 6.7: Parameters for learning, perception, and adaptation chosen for all perceptual inference experiments.

Parameter	**Description**
$\eta^{Xc} = 0$	Learning rate of visual Gestalt codes
$\eta^{Xc} = 0$	Learning rate of proprioceptive Gestalt codes
$\gamma^{Xc} = 0$	Momentum of learning spatial codes
$\eta^{vX} = 0$	Learning rate of visual posture code predictions
$\eta^{pX} = 0$	Learning rate of propr. posture code predictions
$\gamma^{Xp} = 0$	Momentum of learning temporal code predictions
$\eta^{i} = 0$	Learning rate of spatial code classifications
$\gamma^{i} = 0$	Momentum of learning spatial code classifications
$\bar{\mathbf{i}} =$ inferred by model	Class bias in autoencoders
$\tilde{\mathbf{i}} =$ not provided	Classification target
$q^{vp/d/m} = 1$	Visual stimulus reliabilities
$q^{pp/d/m} = 0$	Proprioceptive stimulus reliabilities

In the following experiments, the goal is to reproduce various situations in which the model observes a person's actions visually, recognizes their bodily features and binds them together in the correct order, takes the perspective of the observed actor, and understands their action both by simulating respective proprioceptions and by inferring the corresponding action class. Consequently, visual stimuli are assumed to be available ($q^{vp/d/m} = 1$), while proprioceptive stimuli are unavailable ($q^{pp/d/m} = 0$), as shown in Table 6.7. In contrast to the training phases, all learning is disabled, in that the relevant learning rates (and momentum) are set to 0. Also, the action class is not provided in all of the following experiments neither via teacher forcing nor as target signal for backpropagation, such that the model runs completely unsupervised, and classes or kinematic intentions are inferred by the model.

Seeing the complexity of the task and the architecture, interdependencies between the involved perceptual adaptation, inference, and simulation processes are to be expected. All of the modeled components are thus evaluated both separately and in parallel in the following sections and chapters. The first evaluation of the model's functionality on data from the first test set, consisting of trials of walking, running, and basketball similar to the training trials, evaluates the spatial visualization abilities in terms of perspective-taking and feature binding via the implemented perceptual adaptation mechanisms.

6.4.1 Perspective-Taking

TABLE 6.8: Parameters for perception and adaptation chosen for the isolated evaluation of perspective-taking.

Parameter	**Description**
$\eta^s = 0.01$	Adaptation rate of the origin of the visual FOR
$\gamma^s = 0.85$	Momentum of the adaptation of the origin
$\eta^r = 0.01$	Adaptation rate of the orientation of the visual FOR
$\gamma^r = 0.85$	Momentum of the adaptation of the orientation
$\eta^w = 0$	Adaptation rate of the feature selection and assignment
$\gamma^w = 0$	Momentum of the feature selection and assignment
$\mathbf{b}^{\text{data}} =$ allocentric	Data translation offset
$A^{\text{data}} =$ allocentric	Data rotation offset
$w_{ij} =$ provided	Feature binding weights

In this task, a fully trained model is visually presented with all 12 trials of the first test set, while the correct feature assignment is provided, and the perspective in terms of spatial orientation and translation is to be inferred by the model (see Table 6.8 for the parameters). Each motion capture trial is first transformed by a random, three-dimensional, constant rotation offset, followed by a translation offset, before serving as input to the model. Thus, in other terms, the model perceives the data from an unknown, allocentric view-point, and is to transfer it into its known, egocentric frame of reference. As qualitative measure for this transformation progress, I define the Orientation Difference (OD) by

$$\mathrm{OD}(t) = {}^{180}/_{2\pi} \cdot \mathrm{acos}\left(\mathbf{tr}(A^{\mathrm{data}}(t)A^{\mathrm{model}}(t)) - 1\right) \quad [°] \tag{6.1}$$

$$A^{\mathrm{model}}(t) = \begin{pmatrix} \alpha_x(t) & 0 & 0 \\ 0 & \alpha_y(t) & 0 \\ 0 & 0 & \alpha_z(t) \end{pmatrix} \tag{6.2}$$

where A^{data} is a constant per trial, global rotation matrix applied to all visual inputs, A^{model} is the dynamic, currently inferred rotation matrix of the model (see Section 5.3.2), and $\mathrm{tr}(...)$ is the trace of the resulting matrix multiplication. The OD describes the minimal amount of rotation about an arbitrary 3D axis to transform the currently derived orientation $A^{\mathrm{data}}A^{\mathrm{model}}$ to the encoded, egocentric orientation in degree. An OD of 180° might for example be caused by a top-down inversion of the walker, or by inverting the walking direction of the walker, when the model has not adapted so far. For measuring the translation with respect to the learned egocentric view – the translation difference (TD) – the Euclidean distance is used:

$$\mathrm{TD}(t) = \left\|\mathbf{b}^{\mathrm{data}}(t) - \mathbf{b}^{\mathrm{model}}(t)\right\| \quad [\mathrm{cm}] \tag{6.3}$$

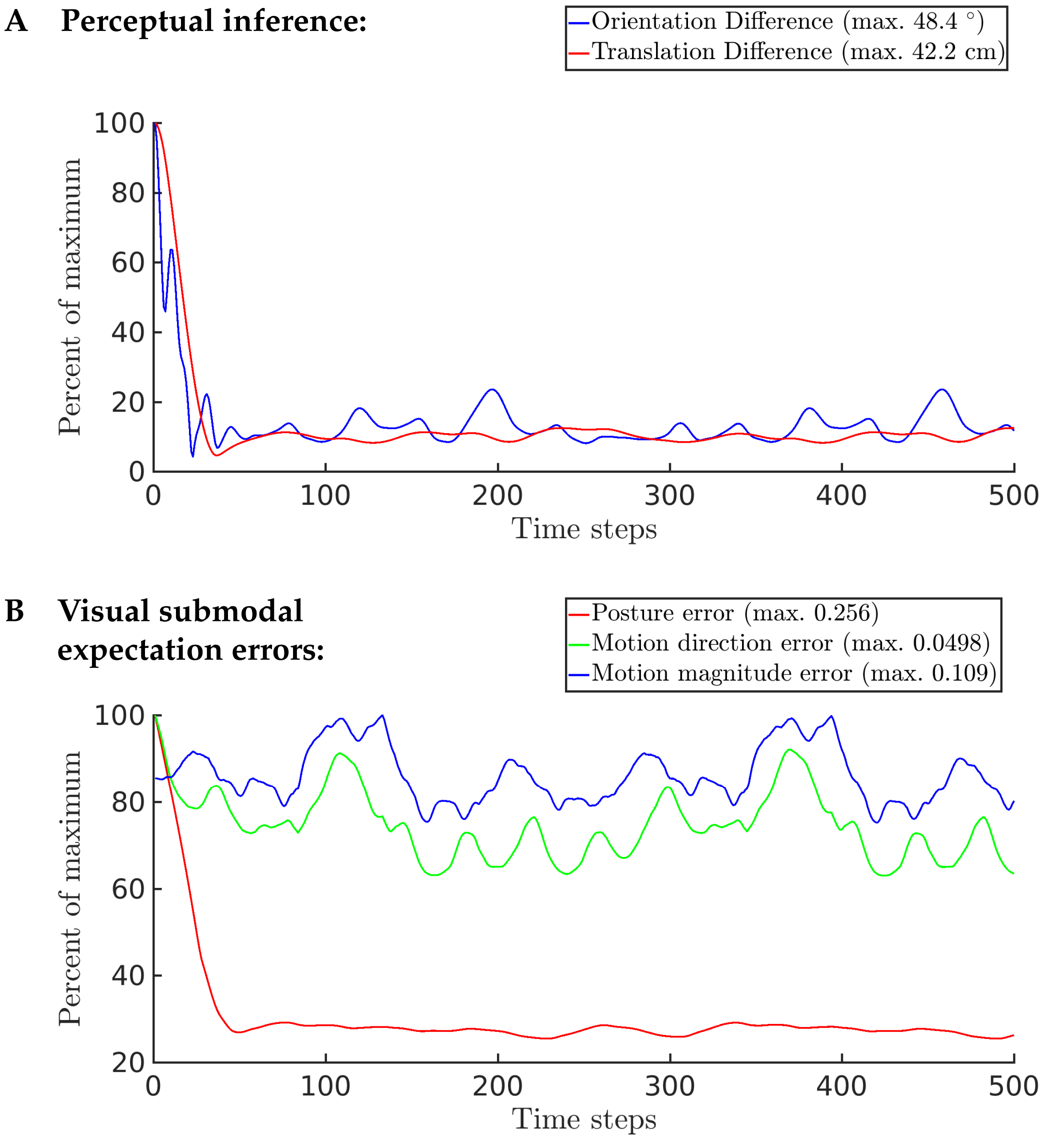
$$\mathbf{b}^{\mathrm{model}}(t) = \begin{pmatrix} b_x(t) \\ b_y(t) \\ b_z(t) \end{pmatrix} \tag{6.4}$$

where $\mathbf{b}^{\mathrm{data}}$ is the constant per trial offset applied to the data, and $\mathbf{b}^{\mathrm{model}}$ is the momentary adaptation of the model. All TD measures are provided in cm.

For the following evaluations, data orientation offsets are sampled equally distributed in the whole OD space (OD $\in [0, 180]°$). Sampling equally distributed orientations instead of orientation differences would result in a considerably higher probability that orientations with about 90° OD are shown, which is avoided to allow a systematic evaluation of the relation between the initial OD and the performance of the model (see Section 6.4.4). Translations offsets are sampled equally distributed in a specified range (TD $\in [0, 56]$ cm, that is, within a 48.5 cm^3 volume). Since the receptive field volume of 175 cm^3 is just as large as the actual stimuli, the TD range is configured to be smaller to ensure that most of the observed features lie within the receptive field.

Figure 6.10 **A** shows a typical example of the development of the OD and TD over time when a walking movement is presented after training, given the above parameterization, an initial (and maximum) OD of about 48°, and an initial TD of about 42 cm. As shown, both the TD and the OD quickly converge in roughly 40 time steps, or in terms of the frame rate of the data, in 0.66 seconds. The remaining OD of about 5° and TD of about 4 cm is largely the result of the differences between the training and the testing data: The model constantly minimizes the error between the input and its own expectations via global rotations and translations, and thus, also subject-specific differences in postural control lead to errors that are minimized to a certain degree, resulting in imprecisions in the perceptual adaptation.

Note that the model has no information about the OD nor the TD. The spatial reconstruction error Δ^{vp} in the visual postural submodality (see Figure 5.9 in Section 5.4) is minimized by the model via the self-supervised, concurrent adaptation of the rotation matrix A^{model} and the translation bias $\mathbf{b}^{\text{model}}$, while the spatial reconstruction error Δ^{vd} in the visual motion direction submodality is minimized by the adaptation of the rotation matrix A^{model} only (see Figure 5.4 in Section 5.3). Figure 6.10 **B** shows the relative descent of these error signals in the example trial. First, it shows that – in relation to the initial (and maximum) value before adaptation – the postural error decreases the most, as can be expected, since the posture submodality is the only one that strongly responds to both changes in orientation and translation, while the motion direction submodality is invariant to changes in the translation. Motion magnitudes are completely invariant to both rotation and translation. Consequently, the motion magnitude error is not, and cannot be regressed by the model via perspective-taking. Second, variances in the error signals are predominantly prevalent in the motion direction and magnitude submodalities, and they are typically higher than in the postural submodality because of the more volatile nature of the types of information they encode. Different variances result from the different test trials, which is largely determined by the similarity to the training trials.



A Perceptual inference:

B Visual submodal expectation errors:

FIGURE 6.10: Comparison of the progress of perceptual inference and submodal expectation errors. The derived orientation and translation difference measures (**A**) decrease simultaneously with the visual motion direction and posture expectation errors (**B**). Different submodal types of information have different characteristics of inveriance to spatial transformations. Motion directions are influenced by rotations only, and thus the error decreases less significantly than the posture error, which is influenced both by rotation and translation adaptations. The motion magnitude submodality is invariant to both translation and rotation and thus does not change on average in the process. The graphs are normalized with regard to the respective maximum defined in the legend.

The progress of perspective-taking in this example run is furthermore visualized in Figure 6.11, showing, besides the point-light stimulus, the convergence of the model's internal frame of reference – driven by adaptations of origin and orientation of the visual module – to the frame of reference of the data. In the process, when the origin is not yet completely inferred, the orientation is slightly overshooting the actual target orientation, as shown in the figure at time step $t = 10$. This is because a rather harsh momentum and learning rate was selected for the experiments, which however pays off in later experiments when the feature binding is *not* provided.

In systematic evaluations of all trials and multiple networks, the model showed comparable performance and succeeded in 100% of all trials of the test set. Table 6.9 shows the mean and variance of the OD and TD for all tested trials after convergence. These values were very stable across independently trained network instances. The reason for the comparatively high final OD and its variance in the first walking trial was – besides a less upright posture in comparison to the training trial – that the walker turns to the right at the end of the trial, which the model was able to *track*, while the reference frame of the data was not compensated for the body orientation, as stated earlier. Similarly, the first basketball trial repeatedly shows rather high deviations from the normal forward direction, which explains the high variance in the remaining OD. Interestingly, the model was also capable of taking the perspective of the left-handed basketball player, highlighting its robustness and generalization. In this case, the distracting features of the left arm may be accountable for the variance in the OD after convergence.

Furthermore, a dependency of the time of convergence on the initial degree of OD was noticed, which coincides with findings about mental rotation (Shepard and Metzler, 1971; Shepard and Metzler, 1988). This dependency will be evaluated in more elaborate tasks in Section 6.4.4. In previous versions, the model did rely on motion direction information only, and often converged to mirrored perspectives in the perspective-taking task when the initial OD was above 90° (Schrodt et al., 2015). In comparison to these results, the addition of postural submodalities speeds up the convergence significantly, and provides

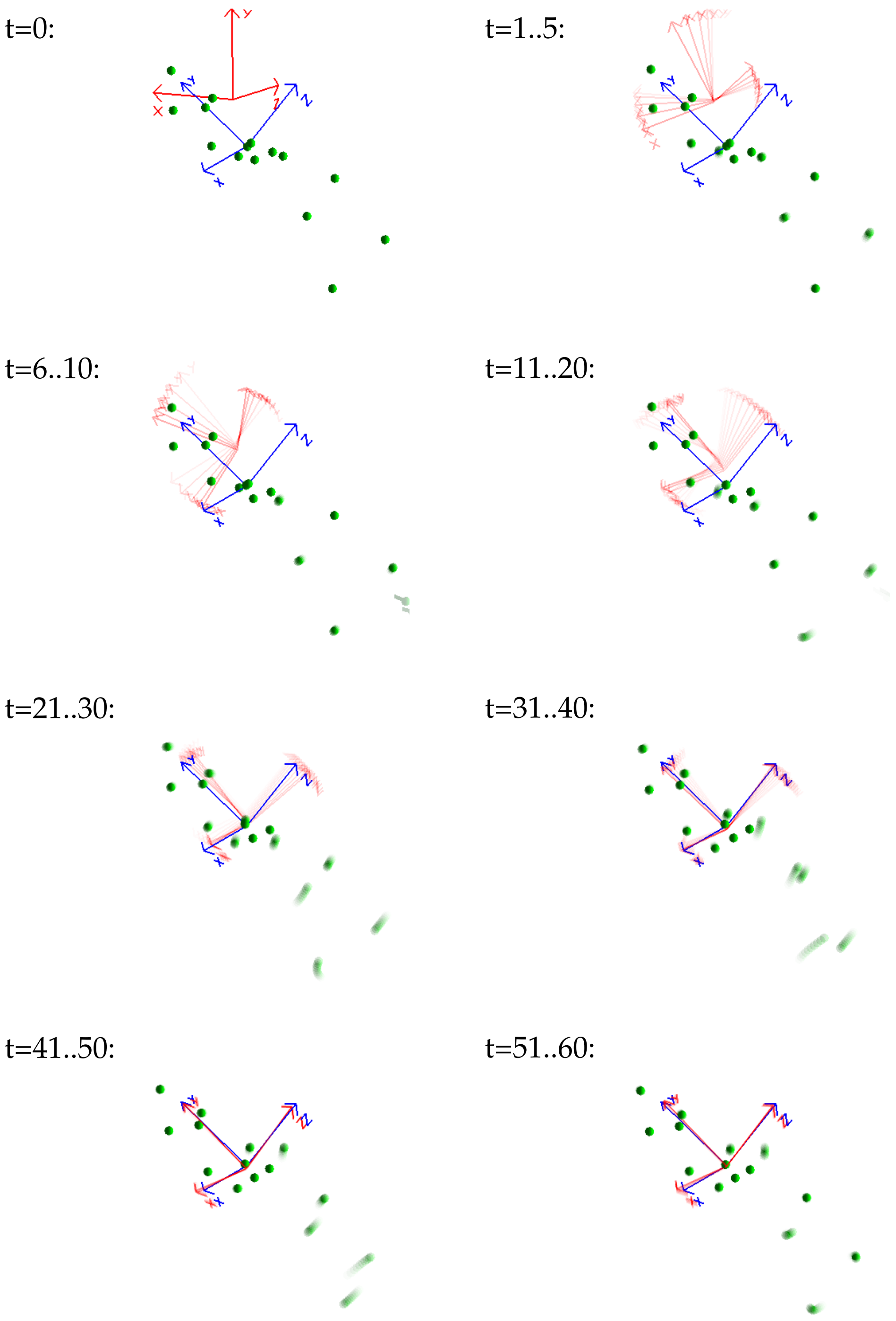


FIGURE 6.11: Progress of perspective-taking given a point-light stimulus. Orientation and translation are inferred in parallel. The model's inferred frame of reference (red) converges to the frame of reference of the data (blue). Each subplot shows the linearly weighted moving average of 10 time steps of the stimulus presented to the network, where green dots represent visual coordinate inputs to the network.

TABLE 6.9: The derived measures for perspective-taking performance after convergence in the isolated perspective-taking task. Shown are the mean and variance of OD (in degree) and TD (in cm) for the respective trials of the test set, as well as on average over all test trials. The results were obtained by averaging over four independently trained network instances.

Subject	**Trial**	**Class**	**Mean OD**	**Var. OD**	**Mean TD**	**Var. TD**
5	1	walking	11.8	2.79	4.74	0.051
6	1	walking	3.75	0.53	2.49	0.18
10	4	walking	6.51	0.351	3.27	0.132
12	1	walking	5.87	0.49	4.52	0.0497
2	3	running	4.1	0.811	1.54	0.199
16	46	running	4.4	1.95	1.03	0.00682
35	19	running	9.61	1.02	1.15	0.0349
35	22	running	6.7	0.725	1.61	0.0553
6	2	basketball (r)	7.22	5.84	1.38	0.167
6	3	basketball (r)	4.97	2.06	1.78	0.0945
6	4	basketball (l)	6.4	5.96	1.72	0.0928
6	5	basketball (r)	4.16	0.659	1.72	0.176
		average	6.29	1.93	2.25	0.103

the model with the necessary information to also infer the orientation of top-down inverted walkers robustly, given that knowledge about the feature identity is provided.

Taken together, the evaluation shows that the model was able to smoothly and robustly take the perspective of an observed actor, given that knowledge about the identity and collocation of bodily features was available. The precision of perspective-taking depended on the similarity in postural control between the embodied actions and the observed actions.

6.4.2 Feature Binding

In this task, the model's feature binding capabilities are evaluated separately from the adaptive perspective-taking components. The parameters for this experiment are shown in Table 6.10. Perspective-taking is disabled, and the test trials are presented to the model without an offset in rotation or translation with respect to the egocentric frame of reference. The feature binding biases w^b_{ij} are reset to -10 for each tested trial, resulting in assignment strengths of $w_{ij} \to 0$,

TABLE 6.10: Parameters for perception and adaptation chosen for the isolated evaluation of feature-binding.

Parameter	Description
$\eta^s = 0$	Adaptation rate of the origin of the visual FOR
$\gamma^s = 0$	Momentum of the adaptation of the origin
$\eta^r = 0$	Adaptation rate of the orientation of the visual FOR
$\gamma^r = 0$	Momentum of the adaptation of the orientation
$\eta^w = 1$	Adaptation rate of the feature selection and assignment
$\gamma^w = 0.9$	Momentum of the feature selection and assignment
$\mathbf{b}^{\text{data}}$ = egocentric	Data translation offset
A^{data} = egocentric	Data rotation offset
w_{ij} = not provided	Feature binding weights
$\beta^{vp} = 1$	Posture expectation error top-down weighting
$\beta^{vd} = 4$	Motion dir. expectation error top-down weighting
$\beta^{vm} = 0.125$	Motion magn. expectation error top-down weighting

such that effectively all observed visual features are initially unassigned. Note that it is not necessary to permute the order of the inputs for the evaluation, since the model loses its knowledge about the correct assignment at this point.

Using a very low initial assignment strength means that almost no activity is forwarded from the submodal populations to the respective submodal code layers in the autoencoders. The codes are then activated almost exclusively by the learned biases that should represent an average submodal perception of the training set. Thus, a prior expectation of the model is present and *guides* the feature assignments initially. The initial assignment strength furthermore decides on the *initial speed* of feature binding, since it marginally activates the codes by mixtures of all possible feature constellations, producing rather chaotic initial error signals with magnitudes proportional to the prior assignment strength. Increasing the initial strength can lead to exceptionally fast but less robust inference (more incorrectly assigned features), since the binding weights are quickly adapted to an initial guess, focusing more on momentary perceptions. The same holds when both feature binding and perspective-taking run in parallel. In this case, the model more often converges to local optima by means of inverted perspectives and mirrored feature assignments. Thus, by parameterizing the initial assignment strength accordingly, the network can be tuned to adapt its perception more or less aggressively. Here, the initial assignment strength is set relatively low such that the model does not react to the resulting perceptual error signals immediately, but integrates them for a short period of time before decisively assigning features.

The prior expectation activated by the code biases alone does typically not lead to the same scale of activity (i.e. length of the code fusion vector $\mathbf{f}^X$) as seen during training. Therefore, the reconstruction error signal (in this case the mismatch between almost no feature input and a marginal prior expectation) can be lower *before* binding visual features than after. However, not selecting any feature did not form an attractor for the gradient descent based feature binding using the above parameterization.

Also to bootstrap the convergence of the feature weights – represented by neurons with logistic activation function – it is crucial to limit the minimum absolute derivative with respect to their bias (cf. flat-spot elimination Fahlman, 1988), in this case, to 0.1. Thus, when a feature weight is saturated (i.e. 1 or 0), the assignment is still flexible and able to respond to submodal errors. Furthermore, the adaptation rate for feature binding has to be exceptionally high to obtain optimal results.

As also shown in the parameters, and as indicated before, the submodal error signals which are minimized for feature binding are individually weighted by the parameters β^{vp}, β^{vd}, and β^{vm}. The weightings were determined heuristically according to their relevance for feature binding, but they are also used in all other experiments. In particular, they also influenced the perspective-taking results, but did not turn out to be crucial or interfere with them, such that they are mentioned just here. Exploratory studies revealed that the magnitude error signal can be very useful for feature binding in the short term after presenting a new action, but less helpful or even obstructive in the long run. The motion direction signals are useful in a short to medium term, and still support the correct assignment slightly in the long run. The postural errors are helpful in the long run and the most crucial component for feature binding and Gestalt recognition.

To measure the progress of feature binding in these experiments, I define the Feature Binding Error (FBE) as the sum of Euclidean distances between the model's assignment of a bodily feature and the correct assignment:

$$\text{FBE}(t) = \sum_{j=1}^{M^v} \sqrt{(w_{jj}(t) - 1)^2 + \sum_{i=1, i \neq j}^{N^v} w_{ij}(t)^2} \tag{6.5}$$

resulting in FBE $\in (0, N^v \cdot M^v)$, while the initial FBE is close to the number of bodily features M^v before binding. A more discrete measure for the correctness of feature binding is provided by counting the number bodily features for which the maximum feature weight is *not* the correct weight, termed Incorrect Assignments (IA), and defined by the function

$$\text{IA}(t) = \sum_{j=1}^{M^v} \min\left(|(\text{argmax}_{i=1}^{N^v} w_{ij}(t)) - j|, 1\right) \tag{6.6}$$

Figure 6.12 **A** shows an example of how these measures typically develop during the observation of a trial of the test data set. First, it can be seen that the FBE slowly but continuously starts to decrease in the beginning (the FBE is theoretically able to drastically increase given that incorrect features are assigned). At the same time, the IA already quickly drops. This result indicates that the model very early obtains information about the correct direction in which to adapt the weights for some but not all of the observed features.

The early information is likely provided by the motion direction autoencoder, as suggested by Figure 6.12 **B**, showing the development of the visual submodal reconstruction errors Δ^X, $X \in \{vp, vd, vm\}$ to be minimized in the same example trial. Initially, the motion direction error is dominant, while the posture and magnitude errors are relatively low. This suggests that the motion direction autoencoder provides a rather strong perceptual bias. As soon as the model starts to minimize the motion direction error, that is, the first features are assigned, the posture and magnitude errors *increase* both to a relatively high level, since the still incomplete assignments activate severely distorted perceptions, leading to error signals that override the errors provided by the code biases as explained before. In the further progress of feature binding, also the posture and magnitude errors decrease to the expected level (cf. Figure 6.10 **B** in the

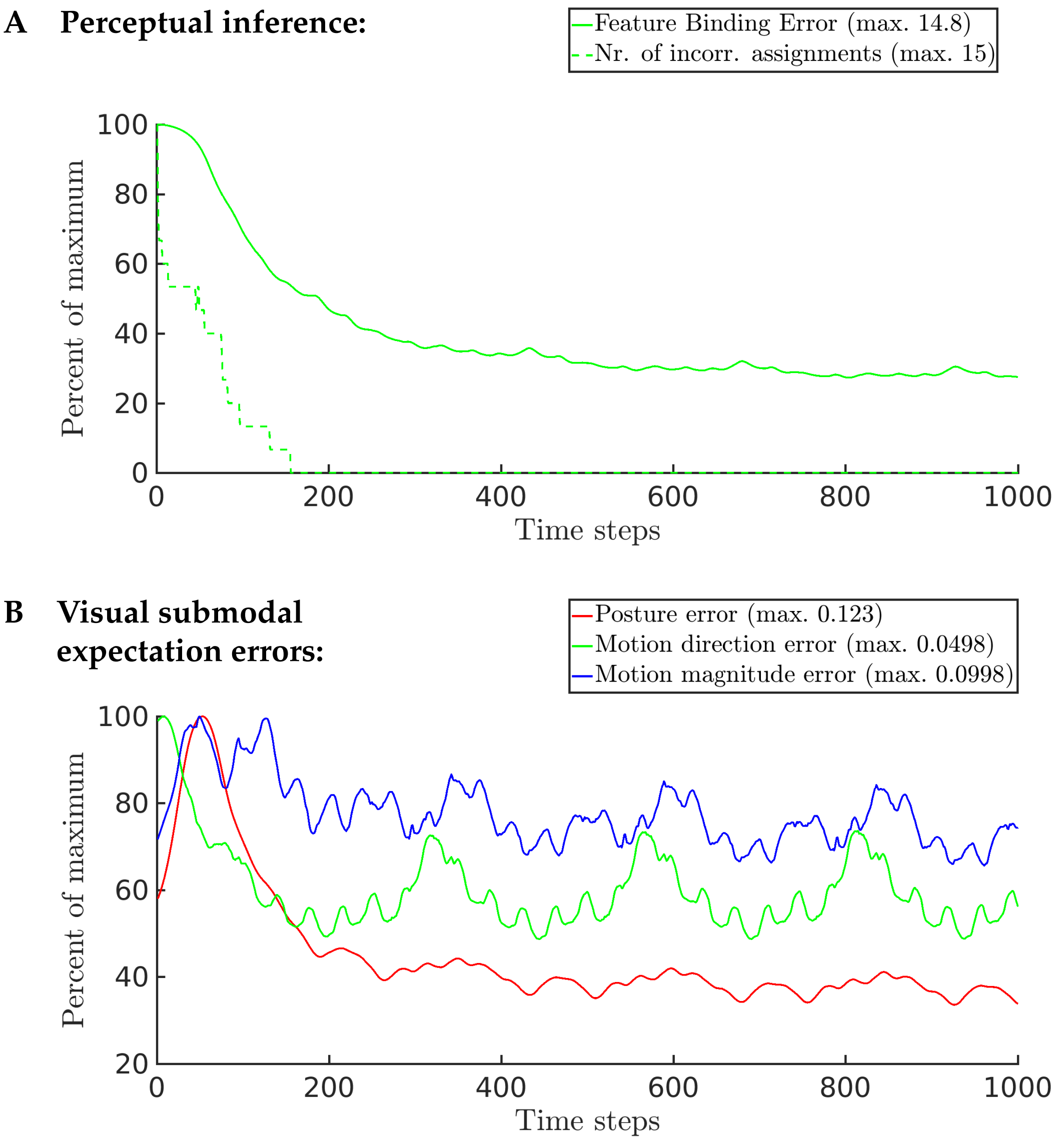


FIGURE 6.12: Comparison of the progress of feature binding with the submodal expectation errors. The feature binding errors (**A**) decrease simultaneously with all submodal expectation errors (**B**). Initially, the driving signal is the motion direction error, which helps the posture and motion magnitude errors to overcome a local minimum. The graphs are normalized with regard to the respective maximum defined in the legend.

perspective-taking experiment). The model identifies all features correctly after about 120 time steps (2 seconds), after which the model is able to stabilize its selection further by minimizing all three reconstruction errors.

Note that the probability to guess the correct assignment by chance in a single shot would be $1.3{\cdot}10^{12}$ in this scenario, and that the model cannot independently keep up and test assignments in parallel, but only iteratively optimizes them using the described heuristics, using parallel and also conflicting population activity. Again, the model has no information about the derived error measures FBE and IA. Nonetheless, these errors very steadily and robustly decrease in the progress of feature binding by means of gradient descent on the submodal reconstruction errors of the autoencoders, although the submodal errors do *not* continuously decrease.

The progress of this feature binding example is also qualitatively illustrated in Figure 6.13, showing the momentary input to the model, the inferred body Gestalt perception, as well the current body Gestalt expectations in the three visual submodalities. The illustration shows that in the beginning, the model expectation (mainly generated by the code biases) is rather vague and spread over the receptive field, while the inferred perception begins to expand from the origin of the visual frame of reference (which is a result of decoding the perceived feature positions from the Gestalt populations that do not get a relevant amount of input yet). At 50 to 149 time steps, the expectation clearly represents a walker-shaped figure, and also slight feature motion can be observed. Still, the inferred perception is rather skewed and imprecise, but more related to the guiding expectation than before. Finally, after convergence, all features are assigned correctly and with sufficient strengths. Thus, both the inferred perception as well as the expectation resemble the embodied experiences (of walking in this case). As interpreted by the population decoder, the inferred perception is still slightly skewed, which results from the remaining expectation errors given the test trial.

Thus, in a nutshell, feature binding continuously improves the inferred perception in the autoencoders. The inferred perception activates the submodal code and in turn generates the submodal expectation as driving signal for feature binding. In the progress, both the inferred perception, as well as the expectations steadily improve towards the learned stimulus, while the expectation is

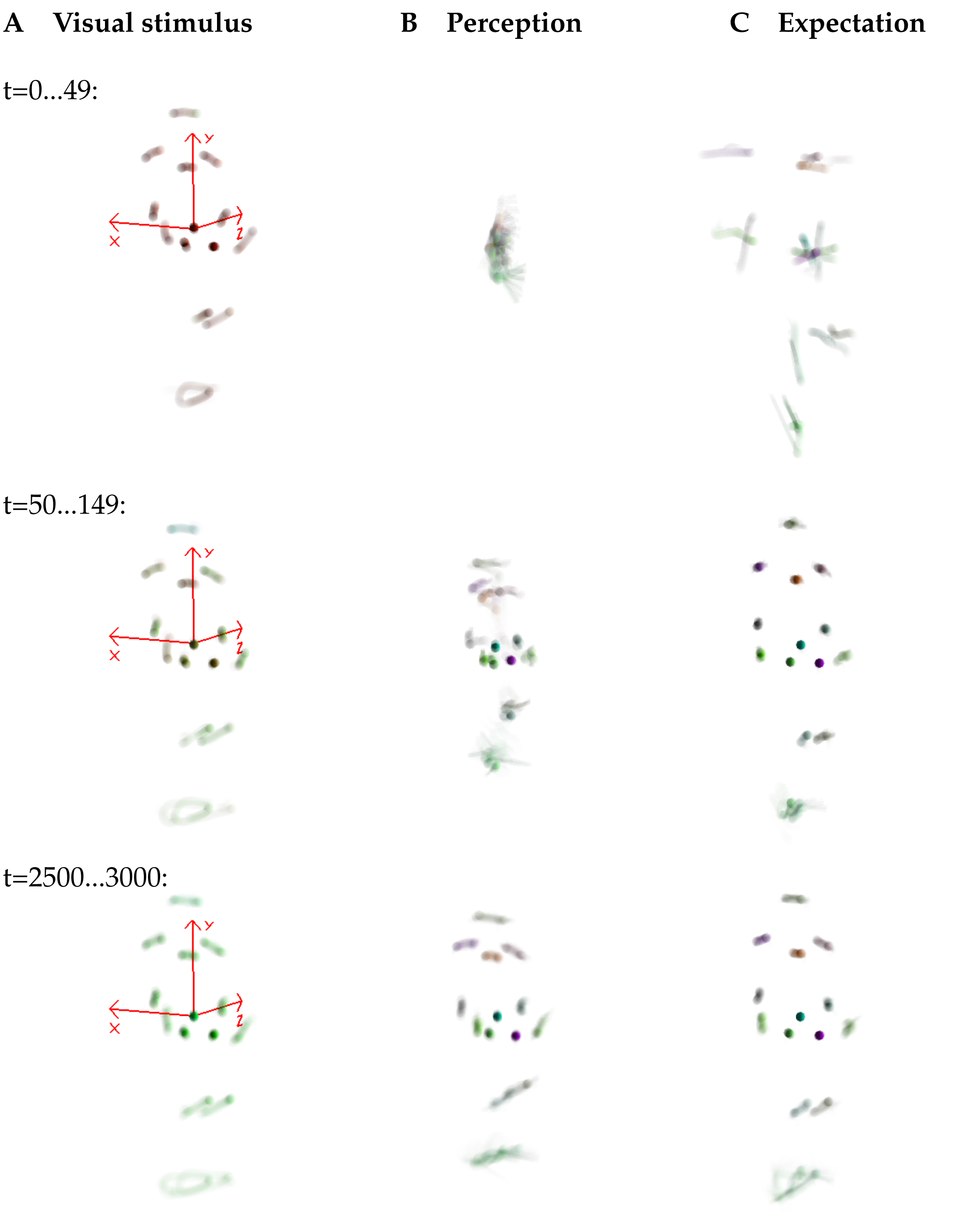


FIGURE 6.13: Comparison of a visual stimulus, the model's perception and its expectation over time during feature binding. Feature binding results both in a concretization of the submodal perceptions after perceptual inference (column **B**) as well as the expectations generated from the embodied codes (column **C**). The difference between the perception and the expectation is in turn used to drive perceptual inference. In column **A**, the feature binding error is color coded for each visual input coordinate (black for unassigned, red for incorrectly assigned, green for correctly assigned). However, color coding in column **B** and **C** represents the identity of the assigned, bodily features.

TABLE 6.11: The derived measures for feature binding performance after convergence in the isolated feature binding task. The results were obtained by averaging over four independently trained network instances.

Subject	Trial	Class	Mean FBE	Var. FBE	Mean IA	Var. IA
5	1	walking	6.93	0.0439	0.773	0.0331
6	1	walking	3.34	0.00153	0	0
10	4	walking	4.89	0.027	0.299	0.0332
12	1	walking	4.54	0.000997	0.658	0.0145
2	3	running	3.85	0.00056	0	0
16	46	running	4.37	0.00098	0	0
35	19	running	4.11	0.00199	1.1	0.0151
35	22	running	3.94	0.0016	0.36	0.0282
6	2	basketball (r)	4.79	0.0182	0.189	0.0223
6	3	basketball (r)	5.07	0.224	0.031	0.0068
6	4	basketball (l)	5.68	0.0124	0.185	0.0478
6	5	basketball (r)	3.82	0.00496	0	0
		average	4.61	0.0282	0.3	0.0168

always ahead, providing a suitable heuristic for optimization. The model generates expectations strongly biased towards its embodied training, and neglects the differences of the observed test trials for its expectations, which is an effect of the non-linear code compression.

Table 6.11 shows the results of feature binding for all test trials after convergence, averaged over different network instances. Again, 100% of the trials were successful within short time, as no more than two visual features were assigned incorrectly. Again, incorrect assignments were predominantly caused by differences in postural control of the tested subjects to the subjects that provided the training samples. The trials for which incorrect assignments were measured varied slightly across different, independently trained networks, and typically concerned the arm features, as they show the most variability across the subjects in the data set.

Although it would be possible to enforce that all features are assigned, in particular when only a single assignment is left, this would prevent the model from not assigning irrelevant features, such as distractors that do not represent bodily features. In this experiment, the model did sometimes consider features that were too far off the expectations as irrelevant. Up to a certain degree, however, missing assignments are recomplemented by the model when activating the whole-body Gestalt codes, and thus *not missing* in the generated Gestalt expectations, such that the model's performance was not compromised.

Taken together, the model was able to systematically, quickly, and robustly infer the identity of observed visual features that correspond to its embodied experience, and assign them to the correct processing paths, given that the perspective was already taken. In the following, possible interactions between feature binding and perspective-taking are investigated.

6.4.3 Interactions of Feature Binding and Perspective-Taking

TABLE 6.12: Parameters for perception and adaptation chosen for the simultaneous evaluation of perspective-taking and feature binding.

Parameter	**Description**
$\eta^s = 0.01$	Adaptation rate of the origin of the visual FOR
$\gamma^s = 0.85$	Momentum of the adaptation of the origin
$\eta^r = 0.01$	Adaptation rate of the orientation of the visual FOR
$\gamma^r = 0.85$	Momentum of the adaptation of the orientation
$\eta^w = 1$	Adaptation rate of the feature selection and assignment
$\gamma^w = 0.9$	Momentum of the feature selection and assignment
$\mathbf{b}^{\text{data}} =$ allocentric	Data translation offset
$A^{\text{data}} =$ allocentric	Data rotation offset
$w_{ij} =$ not provided	Feature binding weights

In this experiment, translation and rotation offsets are added to the tested trials as in Section 6.4.1, *and* the feature assignments are reset at the beginning of each test as in Section 6.4.2. Consequently, both adaptive components are enabled, as shown in the parameters in Table 6.12.

Table 6.13 shows the error remainders for *successful* trials in this combined experiment. The model was able to infer the orientation of the previously unseen test data with an average accuracy of $5.55 \pm 2.44°$, and it inferred their origin with an average accuracy of 3.87 ± 0.185 cm. Comparing these numbers with the individual evaluations of perspective-taking and feature binding reveals no significant differences.

TABLE 6.13: The derived measures for perceptual inference performance after convergence in the combined perspective-taking and feature binding task. The results were obtained by averaging over four independently trained network instances.

Subject	Trial	Class	Mean OD	Var. OD	Mean TD	Var. TD
			Mean FBE	**Var. FBE**	**Mean IA**	**Var. IA**
5	1	walking	10.6	5.01	4.6	0.0912
			6.46	0.0413	0.415	0.0188
6	1	walking	5.07	0.348	3.79	0.0735
			3.55	0.00145	0	0
10	4	walking	5.6	0.425	5.17	0.106
			4.31	0.0154	0.0415	0.00867
12	1	walking	5.79	0.976	5.52	0.0438
			4.43	0.00307	0.496	0.0224
2	3	running	2.91	0.313	2.5	0.573
			4.18	0.00108	0	0
16	46	running	4.27	0.377	3.05	0.0166
			4.47	0.000798	0.5	0
35	19	running	4.99	0.157	4.31	0.0152
			4.39	0.0005	1	0
35	22	running	3.67	0.517	3.31	0.134
			4.08	0.00262	0.414	0.0141
6	2	basketball (r)	7.3	13.3	2.35	0.284
			4.68	0.0142	0.211	0.031
6	3	basketball (r)	5.82	0.734	3.15	0.0683
			5.36	0.0103	0.152	0.0292
6	4	basketball (l)	6.32	4.97	5.71	0.33
			6.17	0.00585	0.681	0.18
6	5	basketball (r)	4.2	2.24	3.01	0.482
			4	0.0304	0	0
		average	5.55	2.44	3.87	0.185
			4.68	0.0106	0.326	0.0254

As indicated in the experiment example in Figure 6.14, the orientation systematically began being inferred *after* some features were correctly recognized. The typical convergence time was significantly longer, since the model sometimes temporarily adapted its internal orientation and translation biases into wrong directions, because it selected the wrong features for a short time, or because it deselected already correctly assigned features. Given that the initial OD was below 90°, these ambiguities in the perceptual inference processes nonetheless were overcome in all evaluated cases. Unlike earlier, however, there were also

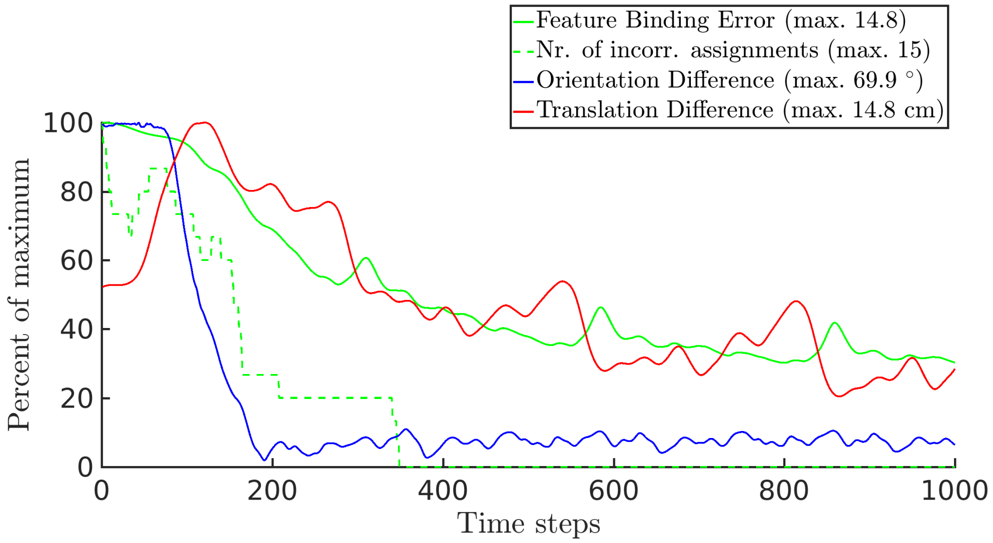


FIGURE 6.14: Simultaneous inference of feature binding and perspective. The inference is more volatile and takes more time. The orientation difference starts to decrease as soon as some of the features are derived. Note that the TD was initially relatively low in this trial, and thus is not decreased much. The graph is normalized with regard to the respective maximum defined in the legend.

unsuccessful trials when the initial OD was above 90°. The model sometimes converged to approximately inverted views of the movement (i.e. top-down an front-back inverted). Thus, under unfavorable circumstances, the model can converge towards locally optimal perceptual attractors.

Taken together, the combined adaptive perceptual processes did not obstruct each other, potentially leading to similarly precise convergence as when tested individually. However, the combination lead to longer convergence times as well as perceptual ambiguities, the systematics of which is evaluated in the next section.

6.4.4 Spatial Systematics in Perceptual Inference

When feature binding and perspective taking are simultaneously applied, thus simultaneously minimize embodied multimodal expectation errors, ambiguous perceptions may occur, which is systematically correlated with the view point

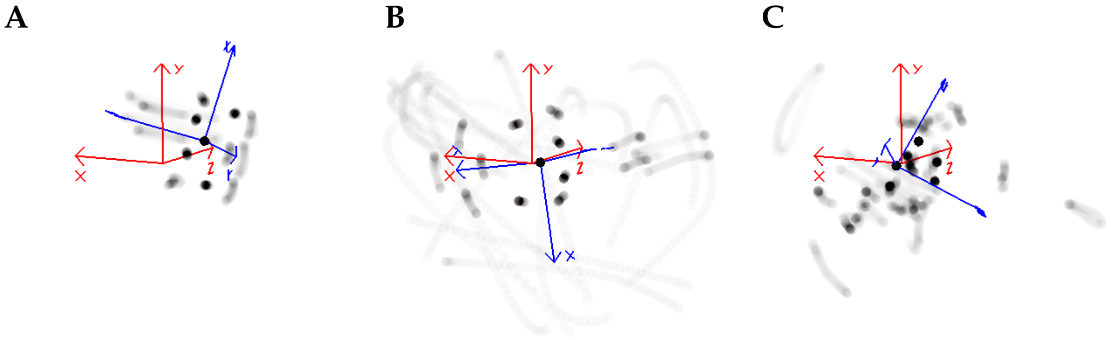


FIGURE 6.15: Segments of visual point light stimuli without distractor points (**A**), with artificial distractor points (**B**), and with biological distractor points (**C**).

the action is shown from. This systematics is evaluated in the following, using the same setup as described in the last section.

Furthermore, also the influence of distractor inputs is evaluated. Figure 6.15 shows an average stimulus display over 100 time steps while an arbitrary biological motion input is presented from an arbitrary view point. In Figures 6.15 **A** to 6.15 **C**, 15 of the point trajectories correspond to the visual bodily features of the subject, while in 6.15 **B** and 6.15 **C**, additional 15 points correspond to distractor features. The distractors follow either artificial or biological dynamics: In 6.15 **B**, each distractor moves into a random direction with random velocity in a specified 3D area (corresponding to the volume of the visual receptive field) around the walker. The distractors rebound at the borders of the area, and their trajectories are smoothed, such that also the resulting motion direction and motion magnitude perceptions cover the whole range of the modeled perceptual system. Furthermore, the motion directions and velocities of the distractors are reset randomly at specific time steps, inducing further complexity in the overall perception.

In 6.15 **C**, the motion dynamics of the distractors are correlated with the bodily features of the currently shown point light display. Each distractor replicates the motion trajectory of a distinct bodily feature, and its trajectory is in temporal synchronization with it. The distractors are independently and randomly rotated and translated in space, in addition to their intrinsic translation with respect to the origin of the frame of reference. Thus, unlike the artificial distractors

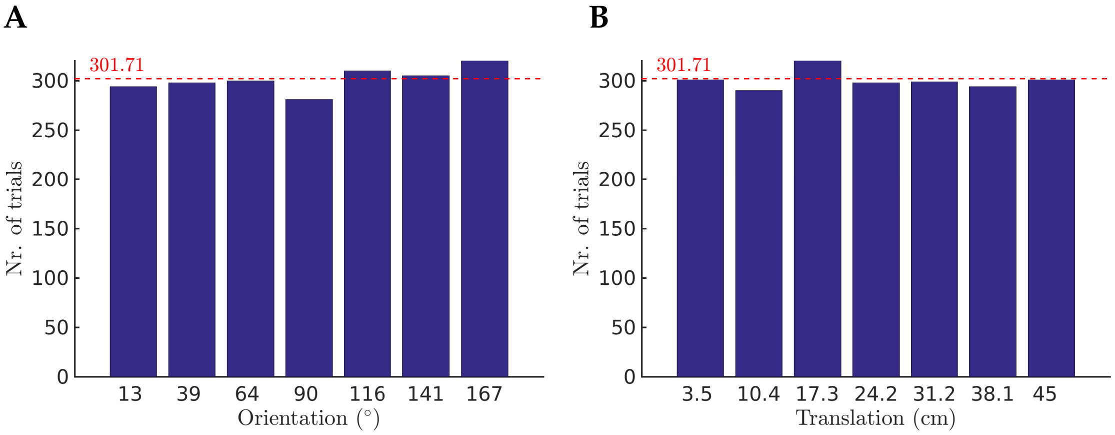


FIGURE 6.16: Number of combined perspective-taking and feature binding experiments. Each evaluated model test was assigned to the nearest of 49 defined initial orientation difference / translation difference configurations. The histograms are marginalized over different initial TDs (**A**) or ODs (**B**), respectively. On average, each OD or TD configuration was evaluated about 300 times per distractor type (red dashed lines).

in Figure 6.15 **B**, the biological distractors in 6.15 **C** typically only move locally and permanently close to other input features. Biological distractors thus generate perceptual ambiguities, mimicking motion dynamics as well as relative proximities of bodily features.

The influence of the distractor type (none, artificial, biological), initial orientation difference, and initial translation difference on the perceptual characteristics of the model was evaluated in 6336 experiments, performed by 88 separately trained networks, that is 2112 experiments per distractor type. Each experiment consisted of the presentation of 120 seconds (7200 time steps) of biological motion in addition to the respective distractor points. Again, the initial orientation and translation of the presented biological motion was selected randomly and uniformly distributed with respect to the orientation difference and translation difference measures, such that all constellations appeared approximately equally often, as validated in Figure 6.16. As also shown in the figure, each initial OD and initial TD was assigned to one of 7 equally distributed ranges, resulting in combined 49 evaluated subconfigurations.

The recognition of a point-light stimulus was considered *successful* when all of the following criteria were appropriate for at least 50 time steps after repeatedly presenting a single motion capture trial for 120 seconds:

- The low-pass filtered orientation difference was less than 15°.
- The low-pass filtered translation difference was less than 7 cm.
- The low-pass filtered number of incorrect assignments was less than 2.

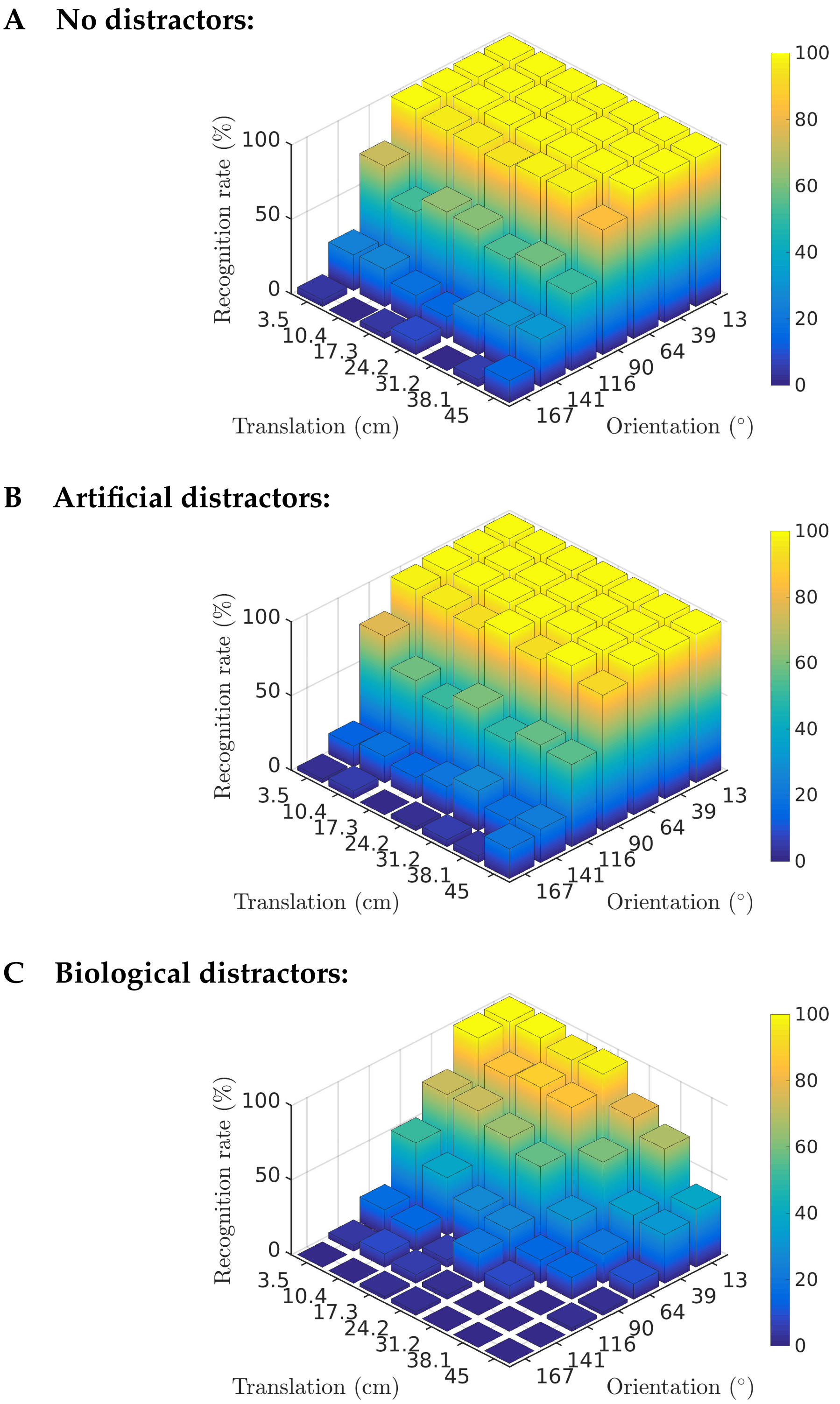
Furthermore, the time until the following criteria were met first in successful trials was determined to evaluate the *speed of recognition*:

- The low-pass filtered orientation difference was less than 15°.
- The low-pass filtered translation difference was less than 7 cm.
- The low-pass filtered number of incorrect assignments was less than 7.5.

Thus, the latter criteria define that an action is already recognized as soon as the perspective is derived, even when only half of the bodily features are identified correctly. Whereas success in the recognition is identified by stable convergence also of the incorrect assignments. All low-pass filtered measures were determined by calculating a moving average over 20 time steps. The recognition rate and time in dependency on perspective and distractors are evaluated in the following.

Recognition Rate

Figure 6.17 shows the median recognition success rate of the model for each initial OD / TD configuration as well as different distractor types. Without distractors, as shown in Figure 6.17 **A**, the percentage of correctly established visual correspondences depended on the shown orientation, while translation had no or no significant influence in the investigated range. Overall, over 69.1% of the trials were successful. For orientations below 74°, even 100% of the trials were successful. The performance began to decrease as the initial orientation exceeds 90° with respect to the egocentric frame of reference. Between 154 and 180° of initial orientation, only about 6% of the trials were successful. In the unsuccessful trials, the network typically converged to inverted perspectives



A No distractors:

B Artificial distractors:

C Biological distractors:

FIGURE 6.17: Influence of the distractor type, orientation and translation on the recognition rate in inferring the observed frame of reference and feature binding.

within the provided time span. This result clearly shows the orientation specificity of the implemented perspective-taking mechanism.

The addition of 15 distractors with artificial motion dynamics and trajectories did not significantly influence these results, as Figure 6.17 **B** indicates. The model was equally able to detect the bodily features and assign them correctly while taking the perspective. The orientation dependency characteristics remained the same, and the overall recognition rate was only slightly lower with 68.4%. This result strongly substantiates that the model learned encodings that particularly focus on features of biological motion, resulting in robustness to a number of non-biological distractors.

Consequently, given 15 biological distractors (Figure 6.17 **C**), the overall success rate decreased significantly to 33.9%. While the negative correlation between the recognition rate and the orientation mismatch increased, the initial translation now had a negative influence on the performance as well. This can be explained by the fact that, given the relatively stationary and systematic biological distractors, more potential candidates are available for the model that approximately match the multimodal expectations in the long run, resulting in additional perceptual ambiguities. On the contrary, none of the artificial distractors is constantly at an expected position and shows the expected motion dynamics.

Interestingly, the above findings with biological distractors are strongly reminiscent of the results of psychometric studies. For example, Pavlova and Sokolov, 2000, measured the rate of success of human subjects in reporting the presence of a point light walker – which was shown to the subjects beforehand – in a point light display with biological distractor points. The display was shown for 1 second, and the walker was either absent, or present and shown in one of seven fixed orientations on the picture plane. As far as apparent from the article and the referenced work, the walkers were *not* translated on the plane, and the distractor features were distributed rather uniformly.

The authors noted a non-linear dependency of the recognition rate on the orientation of the walker. Figure 6.18 shows the overlay of the results obtained by Pavlova and Sokolov, 2000 and the results obtained with the model. The model results used for this comparison are equivalent to the lowest TD-column

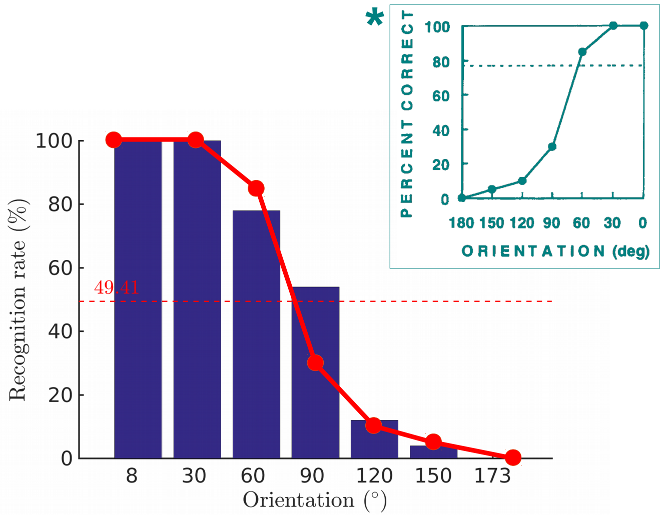


FIGURE 6.18: The orientation specificity of the perspective-taking and feature binding model in comparison to human subjects (* comparison figure from Pavlova and Sokolov, 2000). The model (blue bars) shows a non-linear slope in the successful recognition trials with respect to display orientation similar to human performance as reported by Pavlova and Sokolov, 2000 (thick red overlay). The dashed red line indicates the average number of successful trials.

in Figure 6.17 **C**, thus the comparison was obtained with an initial translation between 0 and 7 cm, random orientation in 3D, biological distractors, and for the purpose of comparability, slightly adapted configuration ranges for the OD histogram. For most initial orientations, the model qualitatively performs precisely as human subjects. It overperforms at around 90° of initial orientation, and underperforms slightly at about 60°. Although the respective task in the compared experiments are admittedly not exactly identical, the model results clearly replicate the non-linear slope of the orientation specificity measured for human subjects.

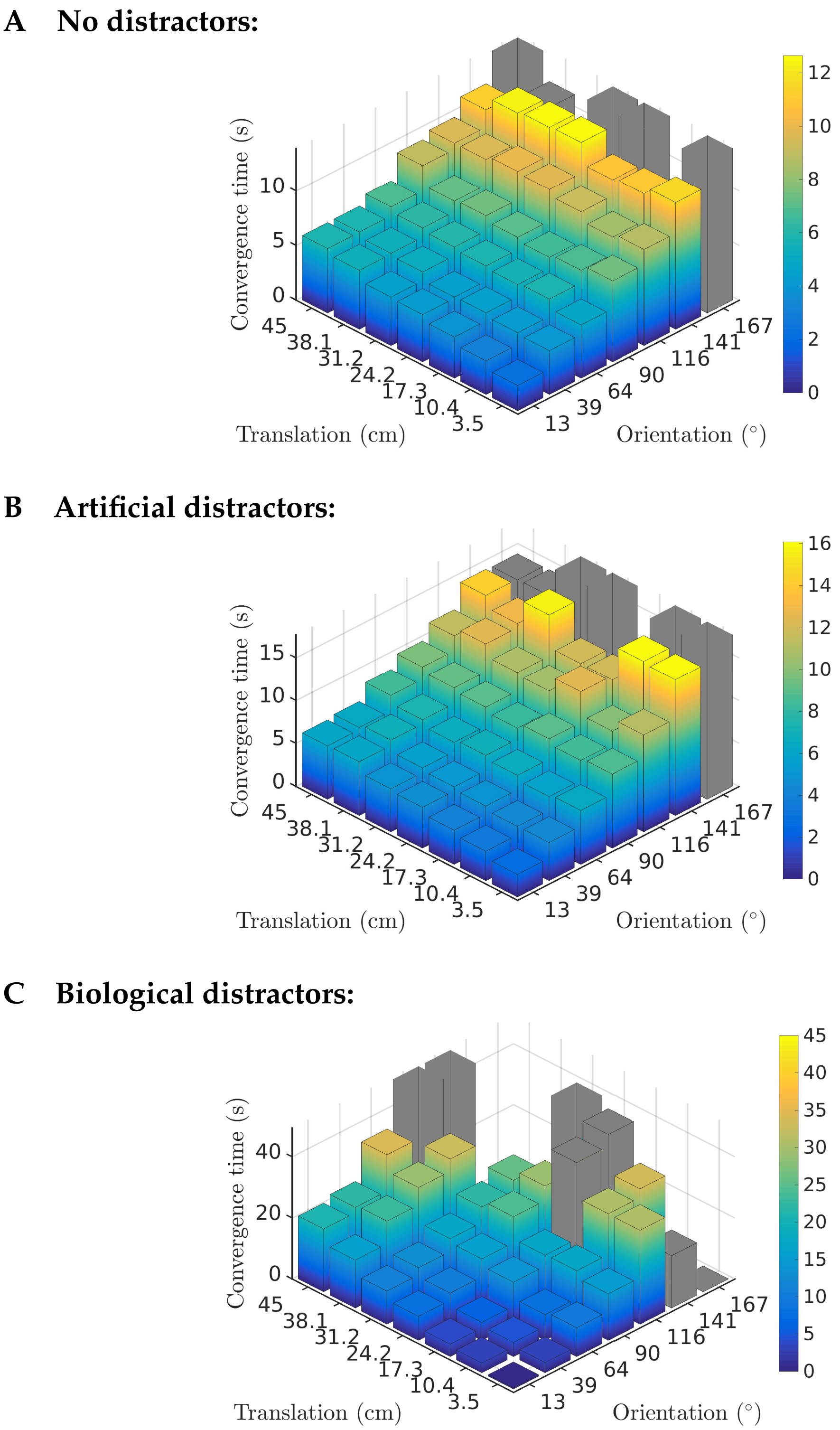
Taken together, the results show that the adaptive perceptual mechanisms were clearly tuned to biological motion perceptions and typically ignore other dynamics. The perceptual characteristics were realistic, which underlines the plausibility of the hypothesis and implementation.

Recognition Time

Without distractor features, the median time until the perspective was inferred and half of the features were identified and assigned correctly *linearly* depends *both* on the initial orientation and the initial translation, as shown in Figure 6.19 **A**. The minimum convergence time (for an initial OD around 13° and an initial TD around 3.5 cm, which already fulfills two of the three criteria for recognition) was about 2 seconds, which is determined mainly by the average time the model needs to bind half of the features, and partly by situations in which the model may have spuriously moved away from the already correct perspective in the meantime. When initial orientations or translations were tested separately, the median recognition time varied linearly between 2 to 11 seconds for different orientations up to 154°, and linearly between 2 to 6 seconds for different translations in the tested range. Given that both orientation and translation were applied simultaneously, the recognition time was below the sum of separate recognition times, substantiating that both inferences ran in parallel and did not obstruct each other.

When the initial orientation exceeded a certain level, as also shown in Figure 6.17, only a few trials were successful, and the variance in the recognition time of these few trials increased drastically. Some of the trials converged only in the long run, taking almost the whole time span provided for the recognition task (120 seconds). Because of this, and because of the lack of more data, the measured median recognition times for these configurations are unreliable, and these are indicated by grayed out bars in the figures.

The addition of artificial distractor points did not influence the rate of convergence in the last section. It did, however, have a negative influence on the speed of convergence with respect to the presented orientation, as Figure 6.19 **B** confirms: The average recognition time with artificial distractors again varied from about 2 to 6 seconds with translation offsets only, but it varied from 2 to 16 seconds for different orientations. Thus, although the rate of success in recognizing biological motion was not compromised by artificial distractor points, it very well influenced the recognition time when increasing the orientation mismatch.



A No distractors:

B Artificial distractors:

C Biological distractors:

FIGURE 6.19: Influence of the distractor type, orientation and translation on the model's recognition time when inferring the observed frame of reference and binding visual features.

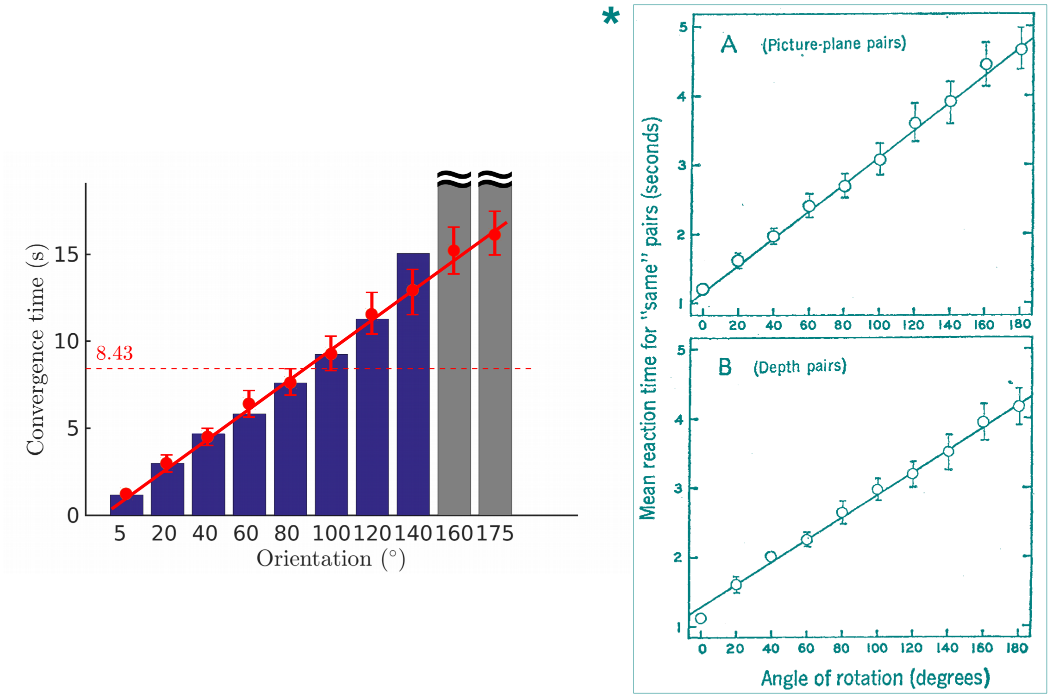


FIGURE 6.20: The recognition time of the perspective-taking and feature binding model in comparison to human subjects (* comparison figure from Shepard and Metzler, 1988). For orientations below 150° , the model (blue bars) shows a linear recognition time with respect to display orientation similar to human performance as determined by Shepard and Metzler, 1988 (thick red overlay). The dashed red line indicates the average recognition time. Statistically inaccurate values are grayed out.

The addition of biological distractors compromised the recognition time more severely, as shown in Figure 6.19 **C**. The figure suggests a rather exponential influence of initial orientations and translations on the recognition time. Again, because of the low number of successful trials for some of the higher initial OD and TD configurations, the respective unreliable values are grayed out in the figure.

As in the last chapter, the linear dependency of recognition time on display orientation observed in Figure 6.19 **A** and Figure 6.19 **B** is in accordance with psychometric studies on mental rotation. It is generally accepted that mental rotation is a continuous process in which the time for transforming one view into another linearly depends on the magnitude of rotation between the views. In

Shepard and Metzler, 1971, the authors investigated the reaction time for comparing two either identical or different static objects with different orientations in 3D space (which is assumed to involve mental rotation). In comparison to the task of recognizing biological motion from a point light display, in both cases, visual features have to be identified and compared (to embodied expectations in the model), and mentally transformed into each other by rotations if possible.

Figure 6.20 shows a direct comparison between data collected by Shepard and Metzler, 1971 and the model results. The model results correspond to the lowest TD-column in Figure 6.19 **B**, and the OD configuration ranges were adapted to match the comparison. The experiment with artificial distractors was chosen for this comparison since in the task evaluated by Shepard and Metzler, 1971, also disparate objects were presented to the subjects, resulting in visual features that could not be assigned. Again, the offset and slope of the real and the modeled results are virtually identical for orientations up to 140°. The recognition time of the model increased linearly from roughly above 1.2 seconds at around 5° (mostly due to feature binding) to 15 seconds at around 140°, while the human results chosen for comparison increased linearly from 1 second for 0° to about 4 seconds for 140° of display orientation.

Again, the comparison underlines the plausibility of the model. The scales in this comparison differ to a certain extent, which however does not harm its significance. Amongst other reasons, although both experiments investigate tasks that involve mental rotation, the respective task itself is again rather different. Computations in the model most likely work on a different time scale than in the human brain. The model can as well be parameterized to converge more or less quickly, depending on the task. Nonetheless, the notion of seconds is used here to reveal how much of the real world data the model has already seen before inference is complete. Moreover, the recognition criterion is a matter of definition. Instead of the chosen criterion, recognition could be determined by the time the model has successfully identified the class of the observed action, or when the proprioceptive codes begin to resonate, which typically happens about half-way of the perceptual inference process. However, this section only considers the dependency of the implemented visual perceptual inference methods on visuo-spatial properties, while inferences from visual stimuli are investigated in the following sections.

6.4.5 Conclusion

Taken together, the model was able to perform perspective-taking and feature binding given stimuli of the classes it was trained on. Individually, the two adaptive perceptual components are consistently robust. Applying both adaptive processes in parallel resulted in reasonably increased convergence times, as well as in perceptual ambiguities. A clear dependency of the model's performance on the stimulus orientation was determined, and the perceptual characteristics of the model replicated findings from psychological studies on biological motion recognition and mental rotation. It was provided empirically that the model learned specific features of biological motion, in that its performance was most significantly more robust to the presence of non-biological distractor stimuli, than to biological distractor stimuli.

6.5 Test Evaluations 2: Action Understanding by Embodied Predictive Codes

The last chapter evaluated the characteristics of perceptual inferences and their interactions in the model. As mentioned in the model hypothesis (see Chapter 3), I assume that similar inference processes – especially perspective-taking – take an important part in human action understanding. In the model, action understanding encompasses two synergistic aspects: First, the *action class* is inferred from the recently perceived visual submodal codes, as well as the recently simulated proprioceptive submodal codes. Second, *proprioceptive submodal simulations* are recurrently biased by the inferred class, previously simulated proprioceptive codes, as well as previously observed visual codes. Because of the recurrent, predictive, bidirectional connectivity between the modules and the continuous influence of observations, it can be said that the proprioceptive system begins to resonate to, and to synchronize with the visual observations in the sense that all activities become increasingly consistent, and converge to the learned, embodied states sequences over time.

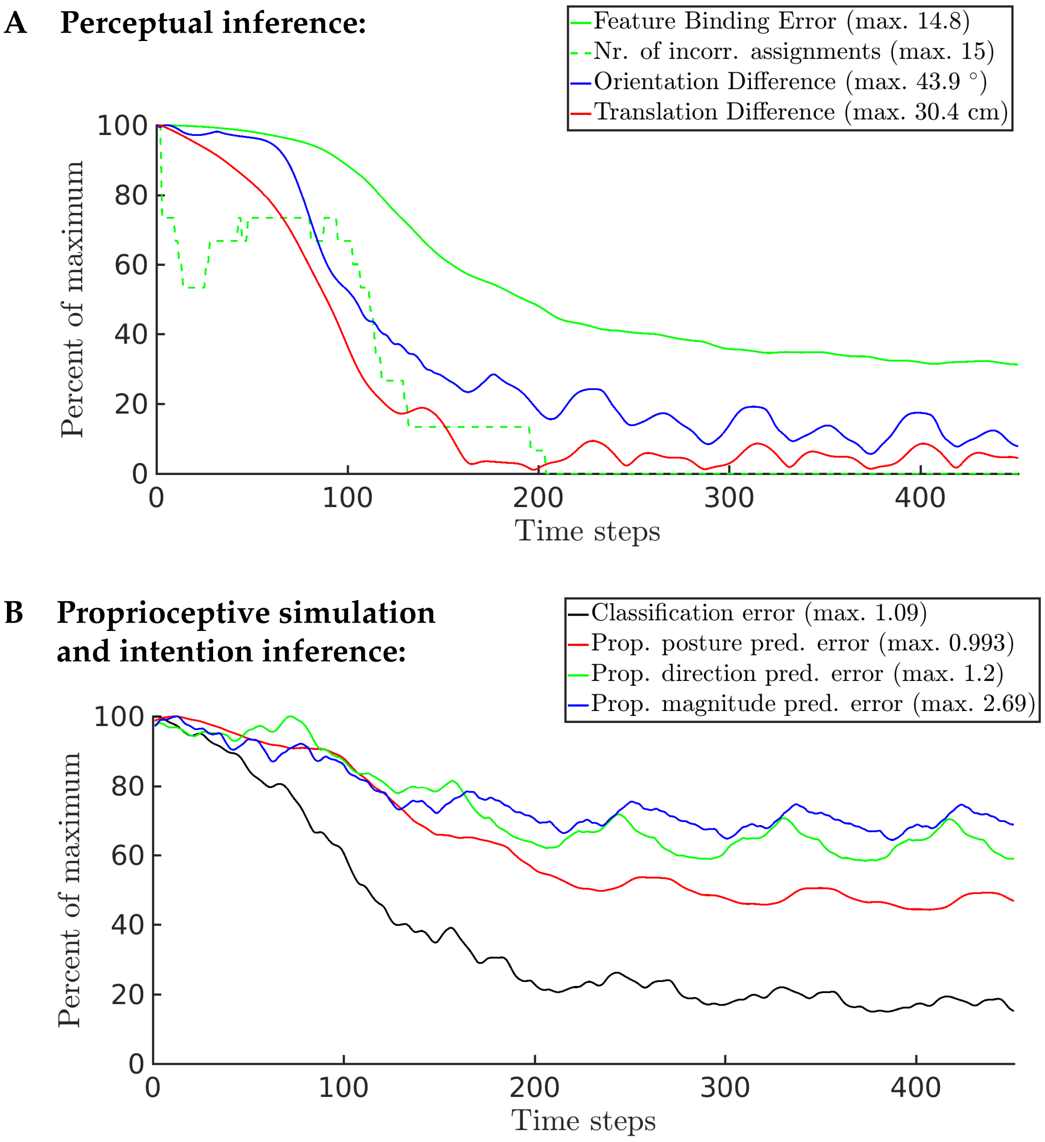


FIGURE 6.21: Perceptual inference simultaneously leads to resonance in the proprioceptive module and inference of the type of observed action. The model is presented with a running movement, while it is internally in a proprioception and intention state that corresponds to walking. While the model infers the perspective and binds the visual features (**A**), also the classification error and proprioceptive code prediction errors with respect to the ground truth decrease (**B**). Note that none of the errors and measures shown here are known to the model. In fact, they are reduced implicitly by predictive coding and by minimizing embodied expectation errors. The graphs are normalized with regard to the respective maximum defined in the legend.

The consistency of the visual modality with the overall action understanding circuit is predominantly determined by the validity of the made perceptual inferences, that is, perspective-taking and feature binding. Thus, in the model, action understanding particularly relies on perceptual inferences, which is illustrated in the following example. The model observes a running trial from the test data set and is, again, not provided with knowledge about the feature identity or frame of reference, nor proprioceptions, just as in Section 6.4.3. Also as before, the Feature Binding Error decreases in the process and the orientation is inferred as soon as some of the features are rudimentarily assigned, while the models translation is continuously dragged into a valid direction, as Figure 6.21 **A** shows. When concurrently taking a look at the classification error Δ^c in the intention module, and the code prediction errors Δ^{ppp}, Δ^{pdp} and Δ^{pmp} in the proprioceptive module, as shown in Figure 6.21 **B**, the dependency of action understanding on the perceptual modulations becomes explicit. Note that here, because of the lack of a proprioceptive stimulus and knowledge about the observed class, all error signals are artificial measures with respect to the *ground truth*, that is, they compare the code that the model predicts with the code that *would have been* activated by the proprioceptive sensation of the test trial, as they compare the inferred class with the actual target class. With that said, all submodal proprioceptive prediction errors decline at about the same time the perspective and features are inferred in the visual pathway, indicating that the recurrent, proprioceptive simulations begin to synchronize with the observation and approximately represent the correct sensations. In its simulation, the model is for the most part restricted to it embodied encodings, thus the remaining error is majorly caused by differences between the trained (embodied) trial and the (observed) test trial (besides the code prediction error remainder from training).

Analogously to the code prediction errors, the classification error (with respect to running in this case) declines, indicating that the classification of the visual and proprioceptive codes becomes possible after perceptual inference, which in turn improves the proprioceptive simulation via top-down biasing. Here, the slightly earlier decrease in the classification error in comparison to the code prediction errors (i.e. before the orientation is beginning to be inferred), can possibly be attributed to the transformation invariant motion magnitude features that rely on feature binding only, and do not require perspective-taking.

Taken together, the error measures derived here for action understanding coincide with the error measures derived for perceptual inference in Section 6.4. The model 'understands' the observed action in terms of its type and corresponding proprioceptive simulations. Again, none of the shown error measures is known to the model. They are implicitly and synergistically minimized through principles of predictive coding, perceptual inference, and embodied simulations. Typically, the resonance of codes in the proprioceptive and intention modules works already part-way of the perceptual inference just as in this example. The resonance is flexible to changes in the observed class, and simulations are typically in temporal synchronization with the observation (i.e. the same gait cycle).

The actual resonance of the proprioceptive system (comparable to motor resonance) the and intention module (comparable to kinematic intention inference) is qualitatively illustrated in Figure 6.22: It compares the average display of (i) the input stimuli, showing the inferred frame of reference and the assignment errors of individual features, (ii) the model's currently expected visual stimulus, as well as (iii) the simulated proprioceptive body model, showing also the inferred class in color coding. The model initially infers the walking class (represented by the red body), which is mostly learned as a default class by the model, and simulates a constant posture, motion direction, and (negligible) magnitude. No relevant changes are visible in the first 50 time steps. The classification then slowly changes from walking to running (green) as the perspective and features are derived and the visual expectations that influence the classification and simulation are enhanced. The simulation then transits to an average running posture, and eventually gets into synchronized resonance with the visual observation.

This is possible because for each action classes the model was trained on, an attractor state sequence was established in the distributed submodal and crossmodal predictive structures. All three attractors reside in the same structures, and thus, there are implicit transitions between the attractors, which are however typically instable. They can be activated by incomplete or jammed visual perceptions, but they are pushed onto the nearest attractor afterwards. Thus, once again, perceptual inference is the decisive component for action understanding in the model, while action understanding itself goes beyond a static mapping from visual to proprioceptive sensations, and can rather be seen as a

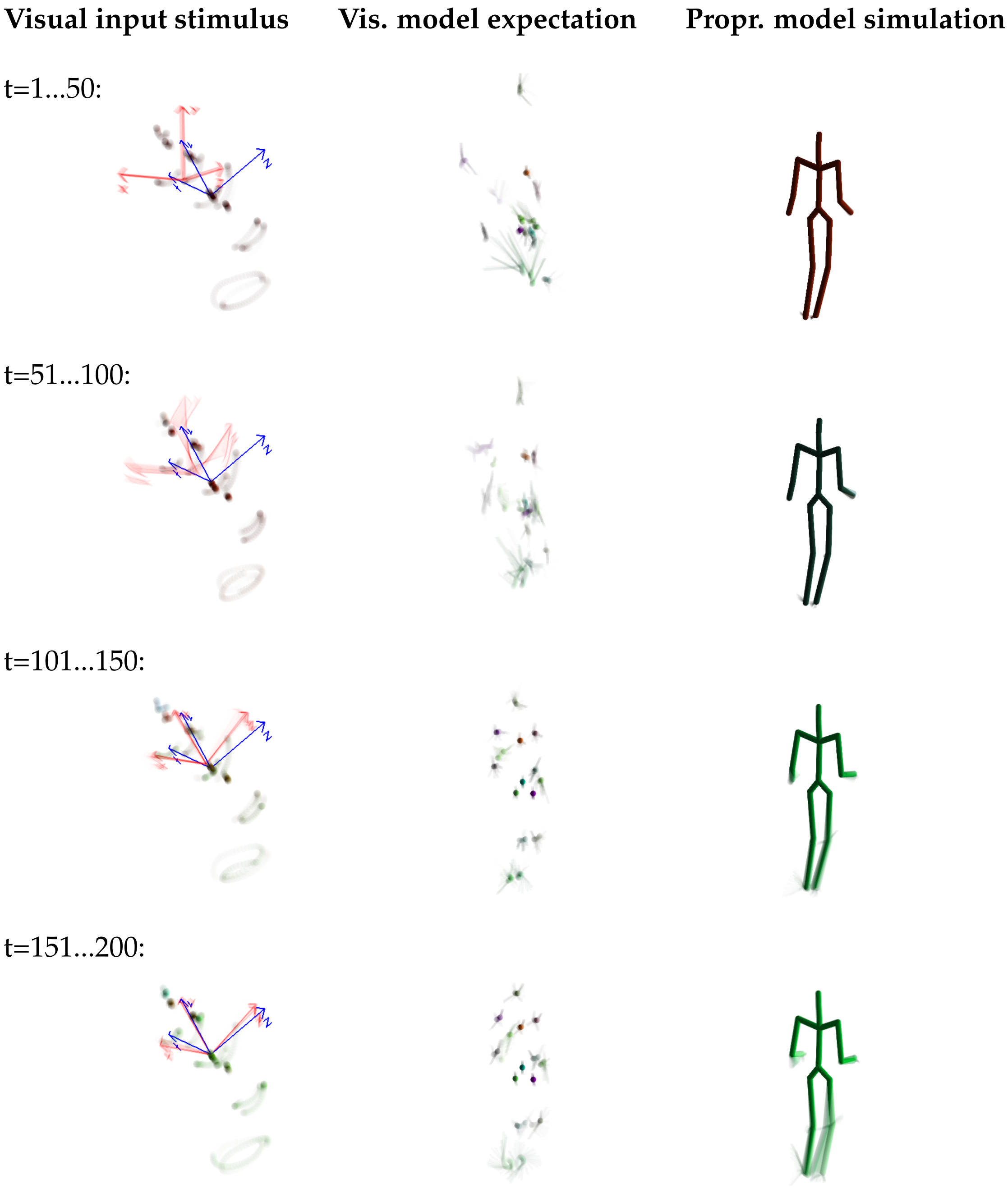


FIGURE 6.22: Illustration of how perceptual inference concretizes model expectations and results in proprioceptive resonance over time. Shown is a visual input stimulus (a running trial) and perceptual inferences on it (left column), the development of the model's visual expectations during perceptual inference (center column), and the beginning resonance in the proprioceptive module (right column) over time (from top to bottom). The expectations begin to concretize while the frame of reference and feature binding is inferred, and at the same time, the proprioceptive module is caused to resonate, and the action class is inferred (motion and color coding in right column).

dynamical system with multiple stable attractors distributed over multiple state spaces. The systematics of these action understanding attractors is evaluated in the following.

6.5.1 Systematics in Action Understanding

TABLE 6.14: Parameters for perception and adaptation chosen for the evaluation of action understanding.

Parameter	Description
$\eta^s = \eta^r = \eta^w = 0$	Adaptation rates for perceptual inferences
$\gamma^s = \gamma^r = \gamma^w = 0$	Momentum for perceptual inferences
$\mathbf{b}^{\text{data}} =$ egocentric	Data translation offset
$A^{\text{data}} =$ egocentric	Data rotation offset
$w_{ij} =$ provided	Feature binding weights
$\bar{\mathbf{i}} =$ inferred by model	Class bias in autoencoders
$\tilde{\mathbf{i}} =$ not provided	Classification target
$q^{vp/d/m} = 1$	Visual stimulus reliabilities
$q^{pp/d/m} = 0$	Proprioceptive stimulus reliabilities

Here, the model's action understanding capabilities are evaluated using all motion capture trials of the first test set. Although action understanding does work in the model also when feature binding and perspective-taking are enabled, perceptual inference is disabled in these experiments to exclude mutual interferences. Thus, visual inputs as well as their feature assignment and frame of reference are provided, while the classification and proprioceptive correspondences are to be inferred by the model. The respective parameters for this evaluation are shown in table 6.14.

Table 6.15 shows an quantitative evaluation of this experiment. For each trial of the test set, statistics were calculated over $t_{\max} - t_0 = 950$ time steps after the overall system was allowed to converge to an attractor for $t_0 = 50$ time steps, which typically happens within 20 time steps given that visual perceptual inference is not needed. First, the statistical evaluation shows the mean and variance of the classification error Δ^c (Mean Classification Error, M. CE; Variance of Classification Error, Var. CE), as well as the percentage of time steps the model

TABLE 6.15: The classifier results and mean code prediction errors in the visual and proprioceptive autoencoders when visual stimuli are provided. Shown are the mean of the classification error (M. CE) and its variance (Var. CE) per trial, as well as the mean of the sum of code prediction errors in the visual (M. VPE) and proprioceptive (M. PPE) modalities. The results were obtained by averaging over four independently trained network instances.

Subj.	**Trial**	**Class**	**M. CE**	**Var. CE**	**% Corr.**	**M. VPE**	**M. PPE**
5	1	walking	0.0406	0.0253	97.9	0.184	0.197
6	1	walking	0.00946	5.37e-06	100	0.169	0.154
10	4	walking	0.00933	8.84e-06	100	0.178	0.172
12	1	walking	0.0152	0.000133	100	0.189	0.171
2	3	running	0.0797	0.00157	100	0.2	0.231
16	46	running	0.0581	0.00188	100	0.228	0.245
35	19	running	0.0688	0.00301	100	0.197	0.207
35	22	running	0.0955	0.00887	99.4	0.199	0.212
6	2	basket. (r)	0.933	0.279	36.6	0.214	0.221
6	3	basket. (r)	0.842	0.246	45.5	0.197	0.216
6	4	basket. (l)	1.31	0.0329	3.54	0.173	0.199
6	5	basket. (r)	0.745	0.118	48.8	0.165	0.193
		average	0.35	0.0598	77.6	0.191	0.201

predicted the correct class with the maximum of its classifier output (% Corr.). These measures are provided by

$$\text{M. CE} = \sum_{t=t_0}^{t_\text{max}} \frac{\Delta^c(t)}{t_\text{max} - t_0} \tag{6.7}$$

$$\text{Var. CE} = \text{var}(\Delta^c(t_0), ..., \Delta^c(t_\text{max})) \tag{6.8}$$

$$\text{\% Corr.} = 100 \cdot \sum_{t=t_0}^{t_\text{max}} \frac{\max(1 - |\text{argmax}_{n=1}^{3}(i_n(t)) - \text{argmax}_{n=1}^{3}(\tilde{i}_n(t))|, 0)}{t_\text{max} - t_0} \tag{6.9}$$

where i_n is the n-th element of the model classification $\mathbf{i}$, and $\tilde{i}_n$ is the respective target. Furthermore, Table 6.15 shows the *normalized average* of all three code prediction errors in the visual (Mean Visual Prediction Error, M. VPE) and proprioceptive (Mean Proprioceptive Prediction Error, M. PPE) modules for each of the test trials:

$$\text{M. VPE} = \sum_{t=t_0}^{t_\text{max}} \frac{3.23 \cdot \Delta^{vpp}(t) + 1.43 \cdot \Delta^{vdp}(t) + 0.91 \cdot \Delta^{vmp}(t)}{16.71 \cdot (t_\text{max} - t_0)} \tag{6.10}$$

$$\text{M. PPE} = \sum_{t=t_0}^{t_{\max}} \frac{3.03 \cdot \Delta^{ppp}(t) + 1.64 \cdot \Delta^{pdp}(t) + 0.81 \cdot \Delta^{pmp}(t)}{16.44 \cdot (t_{\max} - t_0)} \tag{6.11}$$

The normalization is based on the different training set validation errors for the submodal code predictions shown in Section 6.3.2 and accounts for the differences in the magnitude of the code prediction errors across different submodalities. Again, all modal prediction errors were obtained by comparing the ground truth codes (either actually observed or hypothetically observed) to the codes predicted by the model.

The predicted visual codes were not used for future predictions (stimulus reliability $q^{vp/d/m} = 1$). Seeing that all autoencoders predict codes using the same logarithmic code history, the prediction of *both* visual and proprioceptive codes was therefore partially based on observed and thus correct codes, and partially based on predicted and thus typically imperfect codes. Nonetheless, the results in Table 6.15 show that the visual prediction error M. VPE was slightly lower than the proprioceptive prediction error, suggesting that the visual intramodal recurrences are more involved in, and seemingly more suitable for predictions inside the modality than the crossmodal recurrences.

Furthermore, it can be noticed in the results is that the model consistently classified all walking and running trials correctly with high certainty and low variance. The basketball trials, however, are only partially classified correctly, and partially classified incorrectly. This is mainly due to the nature of the training data and the features that the networks learned from it: The basketball trials in the test set contained some segments were the subject stands relatively still, or walks without dribbling while simply holding the basketball (which is not visible to the model) in the right hand. Furthermore, dribblings were partially not synchronously with the gait. In contrast, the training trial included only continuous dribbling, consistent with the gait (i.e. the ball is pushed towards the ground with the right hand at the same time the left foot touches the ground). As a result, the basketball test trials were alternatingly classified as either walking or basketball dribbling: The subject standing still or walking without dribbling was consistently classified as walking. The transitions to the basketball class were specifically triggered by an upwards movement, followed by quick

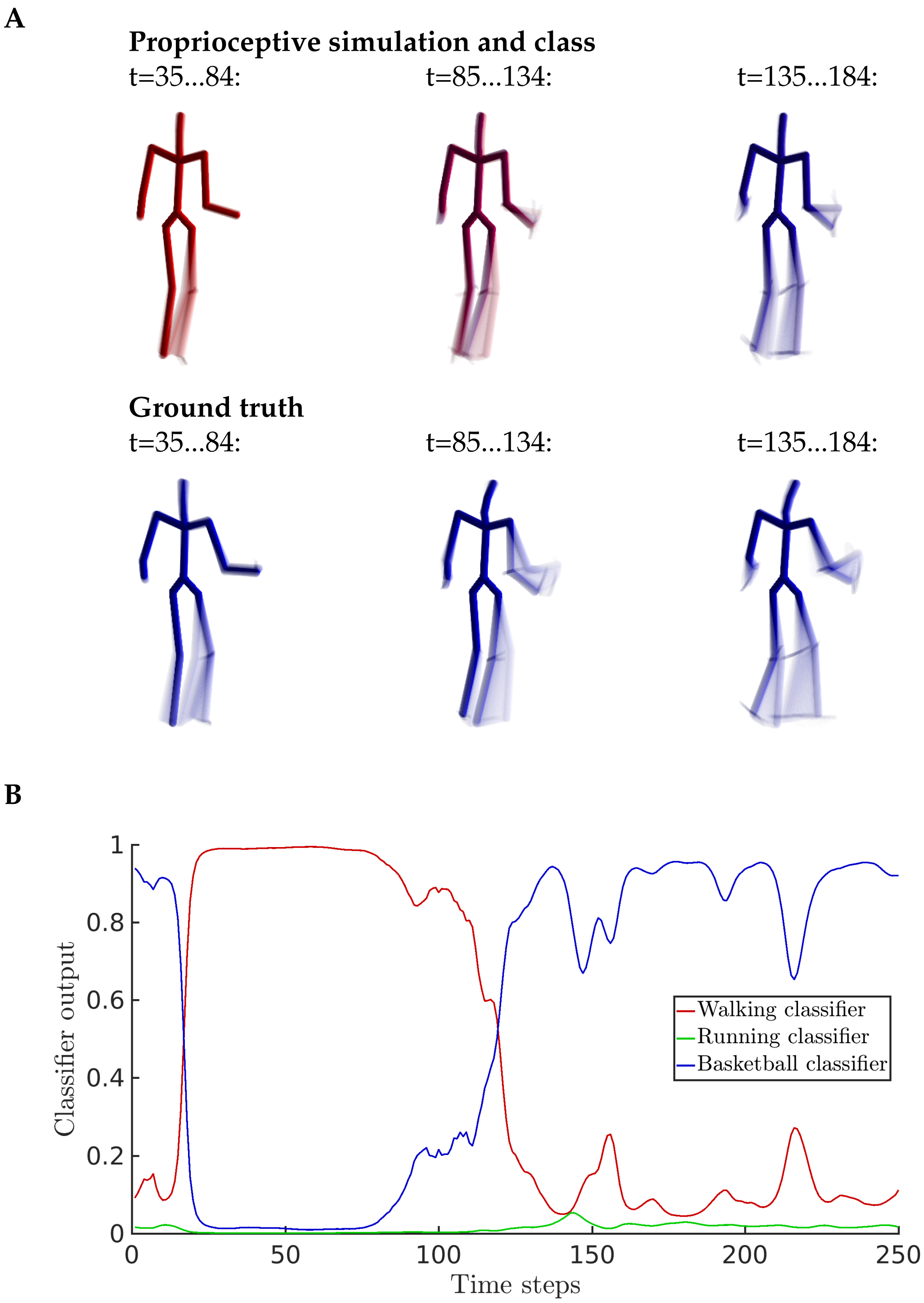


FIGURE 6.23: Classification ambiguity of the basketball and walking trials. **A**: The basketball actions are classified as walking while the subject does not dribble (left column). A downwards hand movement (center column) followed by an upwards hand movement (right column) is recognized by the model as basketball-specific gesture, such that the correct classification is triggered. **B**: The classifier output in this example.

downwards movement of the right arm, as shown in Figure 6.23. This finding shows that the model learned a discriminative right-hand gesture from the multimodal logarithmic code history to identify the basketball trial. The third basketball test trial with left-handed dribbling was rather consistently classified as walking, since the model ignored left-hand gestures. After this gesture, the classifier typically remained in the basketball class for a short time also without dribbling or walking, or with asynchronous dribbling, because of the partially self-preserving / recurrent submodal simulations.

As also indicated in Figure 6.23, the models' proprioceptive simulation does not accurately match the ground truth of proprioceptive stimuli from the test set. This is due to the fact that the learned code manifolds are strongly biased towards the sensory contingencies experienced during the embodied training phase, such that observations rather activate the nearest of these submodal codes. Thus, visual expectations and simulated proprioceptions are always close to the training trials, which is shown for all movement classes in Figure 6.24, Figure 6.25, and Figure 6.26, respectively. As illustrated, the model does not learn general movement patterns, but rather learns self-specific, individual action contingencies, which are activated akin to nearest-neighbour matching, and which ultimately improves the consistency of the proprioceptive simulations and the performance of the kinematic intention classifier.

Because of the bi-directional, predictive connectivity between the modules, and because of the selective usage of either predictions or observations for the model's inferences, it is possible to cross-validate the above results in scenarios where only proprioceptive stimuli are provided, or both visual and proprioceptive stimuli are provided, as shown in the following.

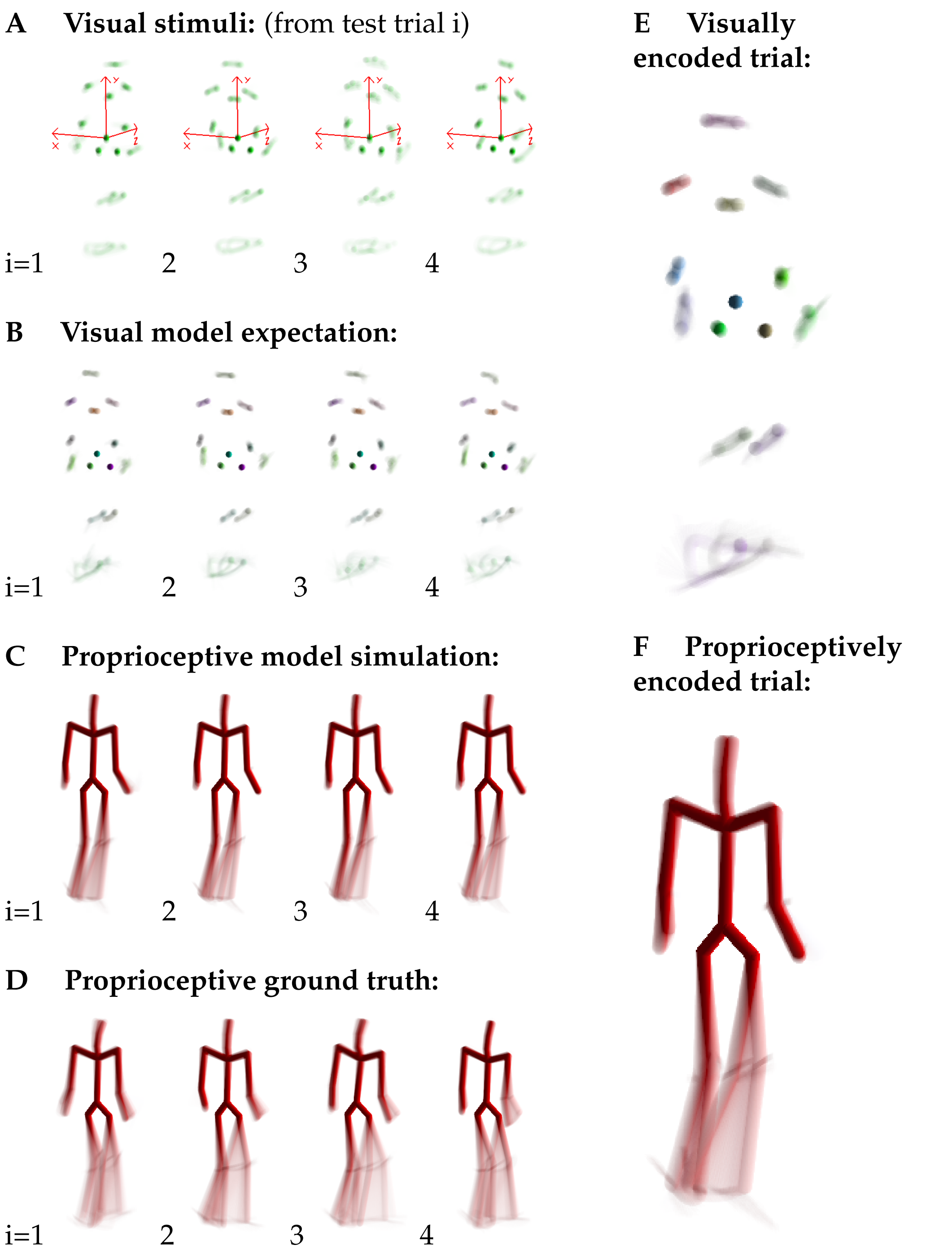

A Visual stimuli: (from test trial i)
y
x
z
i=1
2
3
4
B Visual model expectation:
i=1
2
3
4
C Proprioceptive model simulation:
i=1
2
3
4
D Proprioceptive ground truth:
i=1
2
3
4
E Visually encoded trial:
F Proprioceptively encoded trial:

FIGURE 6.24: Comparison of the model's expectations and simulations for all tested walking trials. Although the visual input stimuli (**A**) and also their theoretical proprioceptive counterparts (**D**) vary in this class, the model's visual embodied expectations (**B**) and proprioceptive simulations (**C**) are strongly biased towards the learned trials (**E** and **F**). This shows that action understanding is possible by means of embodied encodings that can be triggered by perceptual inferences and predictive encodings.

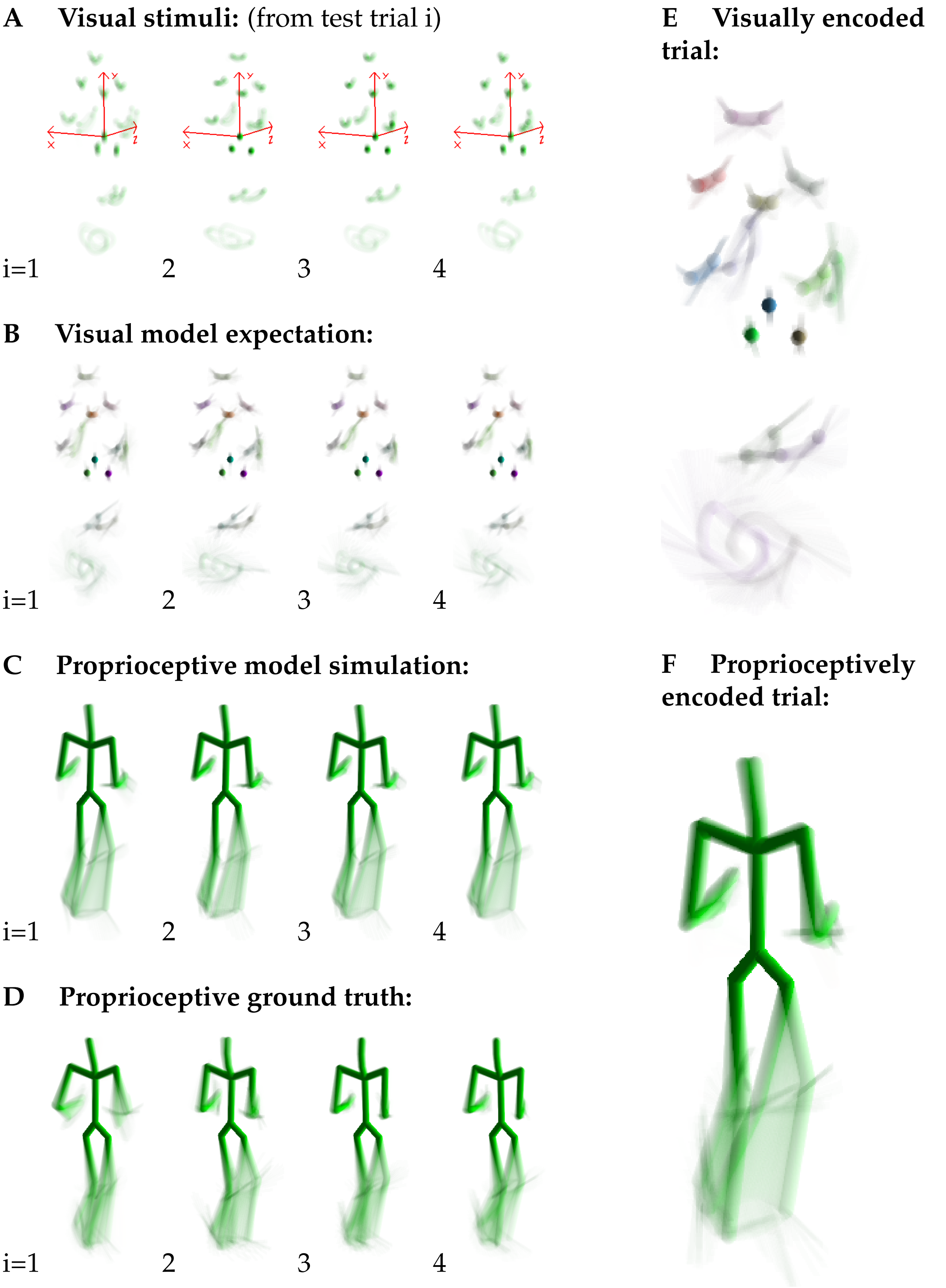

A Visual stimuli: (from test trial i)
y
x
z
i=1
2
3
4
B Visual model expectation:
i=1
2
3
4
C Proprioceptive model simulation:
i=1
2
3
4
D Proprioceptive ground truth:
i=1
2
3
4
E Visually encoded trial:
F Proprioceptively encoded trial:

FIGURE 6.25: Comparison of the model's expectations and simulations for all tested running trials (annotation as in Figure 6.24).

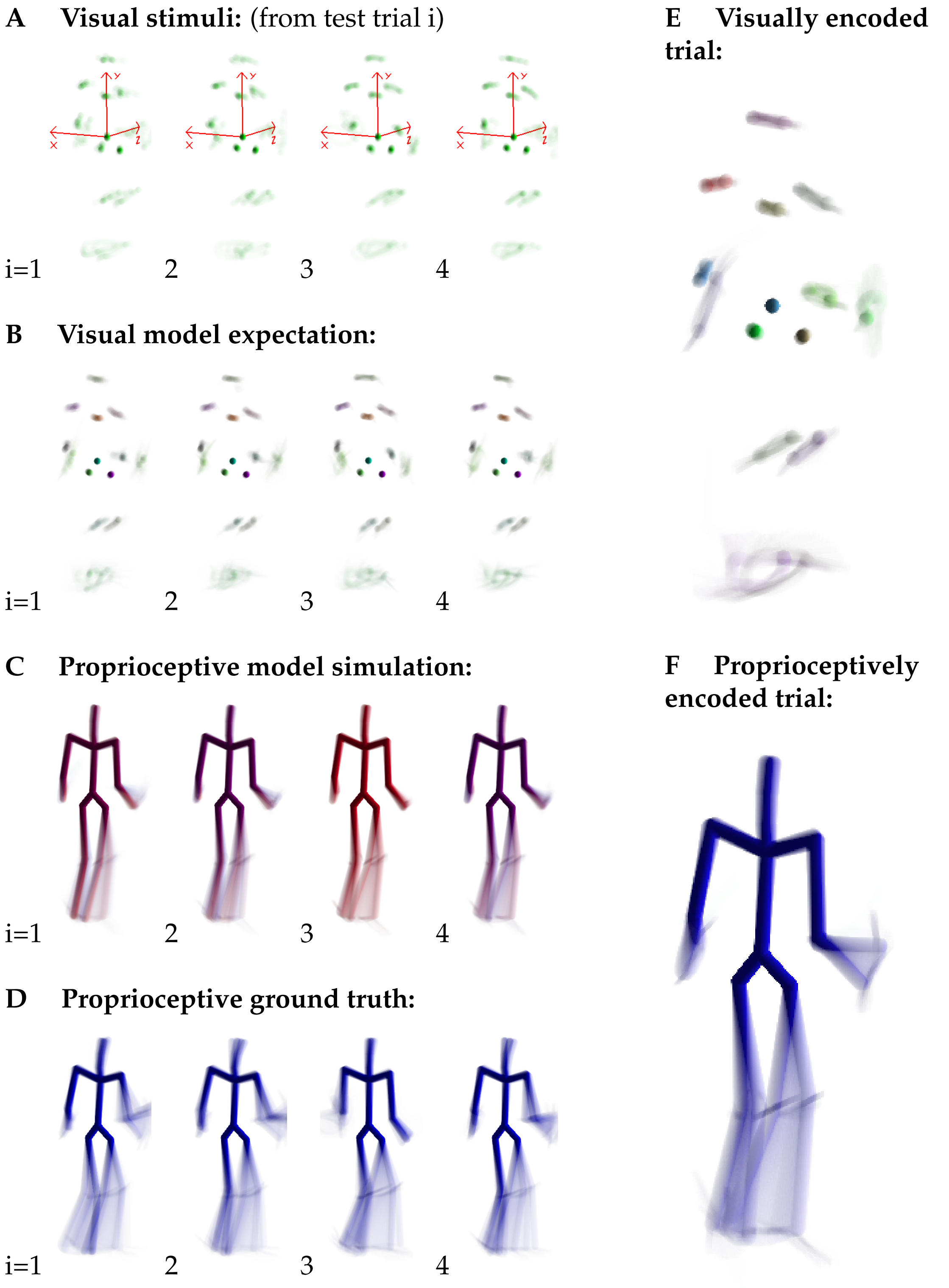

A Visual stimuli: (from test trial i)
y
x
z
i=1
2
3
4
B Visual model expectation:
i=1
2
3
4
C Proprioceptive model simulation:
i=1
2
3
4
D Proprioceptive ground truth:
i=1
2
3
4
E Visually encoded trial:
F Proprioceptively encoded trial:

FIGURE 6.26: Comparison of the model's expectations and simulations for all tested basketball trials (annotation as in Figure 6.24).

6.5.2 Cross-validation

First, the results obtained in the last chapter are cross-validated with the results obtained when proprioceptions are provided, and visual codes are to be simulated by the model, the results of which are shown in Table 6.16. The model performance in both evaluations is basically comparable, here, however, the model was able to achieve slightly better classification results for the basketball trials. This might indicate that the model simulates visual codes more alike to the training set using the observed proprioceptive codes, than vice versa, which would explain the relatively higher visual prediction error with respect to the ground truth. The proprioceptive code prediction error is rather consistently below the visual code prediction error. This furthermore suggests that also the proprioceptive intramodal recurrences are more involved in the predictions than the crossmodal recurrences.

TABLE 6.16: The classifier results and mean code prediction errors in the visual and proprioceptive autoencoders when proprioceptive stimuli are provided ($q^{vp/d/m} = 0$ and $q^{pp/d/m} = 1$). The results were obtained by averaging over four independently trained network instances.

Subj.	**Trial**	**Class**	**M. CE**	**Var. CE**	**% Corr.**	**M. VPE**	**M. PPE**
5	1	walking	0.0106	5.38e-06	100	0.214	0.188
6	1	walking	0.00909	2.45e-06	100	0.195	0.153
10	4	walking	0.00917	4.01e-06	100	0.198	0.164
12	1	walking	0.0155	0.000108	100	0.233	0.179
2	3	running	0.06	0.000449	100	0.227	0.22
16	46	running	0.0374	3.75e-05	100	0.259	0.231
35	19	running	0.0449	0.00054	100	0.237	0.205
35	22	running	0.0426	0.000107	100	0.228	0.205
6	2	basket. (r)	0.875	0.316	40.4	0.248	0.219
6	3	basket. (r)	0.736	0.288	52	0.218	0.206
6	4	basket. (l)	1.28	0.0343	2.98	0.19	0.188
6	5	basket. (r)	0.376	0.0703	80.6	0.17	0.174
		average	0.291	0.0591	81.3	0.218	0.194

When *both* visual and proprioceptive stimuli are provided, the results of which are shown in Table 6.17, the average classification performance improves only slightly, substantiating that visual or proprioceptive simulations were indeed suitable for classifications. As pointed out above, modal predictions seem to be established mostly by activity in the same modality. However, if no crossmodal activity was used at all, the two modalities were dependent on each other only via the top-down class bias, such that crossmodal influences weren't possible, and the two modalities could run asynchronously. However, in the results shown here – where predictions and classifications are purely driven by sensations – both the visual and proprioceptive prediction errors decrease below the errors observed when only one modality was provided. This suggests that modal predictions and simulations are not solely established by the respective modal information, but also apply crossmodal information, which is also confirmed by the fact that actions are typically simulated synchronously with the other modal stimulus.

TABLE 6.17: The classifier results and mean code prediction errors in the visual and proprioceptive autoencoders when both visual and proprioceptive stimuli are provided ($q^{vp/d/m} = 1$ and $q^{pp/d/m} = 1$). The results were obtained by averaging over four independently trained network instances.

Subj.	Trial	Class	M. CE	Var. CE	% Corr.	M. VPE	M. PPE
5	1	walking	0.0218	0.0059	99.8	0.177	0.176
6	1	walking	0.00923	7.94e-06	100	0.161	0.139
10	4	walking	0.00902	5.88e-06	100	0.164	0.151
12	1	walking	0.0236	0.00114	100	0.182	0.156
2	3	running	0.0826	0.00114	100	0.19	0.202
16	46	running	0.051	0.00076	100	0.201	0.2
35	19	running	0.0721	0.00465	100	0.182	0.18
35	22	running	0.0802	0.00571	100	0.189	0.187
6	2	basket. (r)	0.783	0.347	46.4	0.198	0.19
6	3	basket. (r)	0.678	0.352	58.1	0.181	0.184
6	4	basket. (l)	1.22	0.0786	6.88	0.164	0.178
6	5	basket. (r)	0.446	0.137	77.1	0.147	0.16
		average	0.289	0.0778	82.4	0.178	0.175

6.5.3 Conclusion

In this chapter, it was shown that the model's proprioceptive system resonates to visual point-light stimuli, and that this resonance is contingent upon perceptual inferences by means of perspective-taking and feature binding. Further evaluations revealed that the model identifies general, predictive key features for action understanding inside and across visual and proprioceptive domains, as well as an abstract, slow-dynamical intention state space that drives distributed simulations. For classification, these key features can be as specific as right-hand gestures that span several time steps. The embodied simulation clearly relies on embodied, predictive encodings, and nonetheless the model is able to activate the nearest, best-matching action patterns for a respective observation, resulting in synchronized, multimodal simulations and inferences. The problems in classifying basketball actions were ascribed to the specificity of the training trial, which contained considerably less variance in postural control than the test trials.

6.6 Test Evaluations 3: From Action Understanding to Imagery

This section investigates the linkage between action understanding by means of embodied simulation, and mental imagery of actions according to the model and its hypothesis. Because of the bidirectional predictive connectivity, the model can preserve internal activity sequences and submodal simulations also without external stimuli, and thus *imagine* encoded actions both in visual and proprioceptive domains. Two experiments that show the model's imagination abilities are performed in the following: In the first experiment, simulations are *primed* by driving the model by a visual stimulus for a short time, after which the internal predictions and classifications run freely. In the second experiment, no stimulus is provided. Instead, the classification is provided, such that the modal simulations are constantly *biased* and pushed towards converge to the respective, class-specific attractor.

6.6.1 Primed Simulation

In this experiment, the model is driven by a visual stimulus from the test set for 300 time steps, followed by 2700 time steps without stimulus in which the model simulates actions without any driving signal. The respective parameters for the experiment are shown in Table 6.18. While the stimulus is present, the

TABLE 6.18: Parameters for perception chosen for the evaluation of simulation priming.

Parameter	**Description**
$\bar{\mathbf{i}}$ = inferred by model	Class bias in autoencoders
$q^{vp/d/m} = 1$ for 300 time steps, then 0	Visual stimulus reliabilities
$q^{pp/d/m} = 0$	Proprioceptive stimulus reliabilities

model infers the class of the action it observes, and the proprioceptive module resonates to the visual perception, as shown in the last section. After the stimulus offset, the model continues to simulate activity in the modules, or rather, imagine the progress of the action it observed. The overall, recurrent, predictive system is then pushed towards a self-sustaining activity attractor formed during embodied training. Here, we investigate if the learned action attractors are stable, generate continuous, cyclic actions, and in particular, if a presented stimulus primes the imagination to the effect that the network converges to a corresponding action attractor.

Figure 6.27 shows the classifier output of a network that was primed by all trials of the test set successively. The figure shows that the classification of the imagined visual and proprioceptive codes indeed converges to a specific class in each case, and that it remains there relatively stable. This confirms that the network successfully and robustly generates visual and proprioceptive, class-specific imaginations when running freely. The speed of convergence depends on the class (quickly for walking, and rather slowly for running and basketball), and after convergence, there are different magnitudes of variance for each class (low variance for walking, comparatively high variance for running and basketball). A visual display of the class specific, visual and proprioceptive imaginations of the network is shown in Figure 6.28. In all cases, the model simulates prototypical, cyclic actions in multiple modalities and sustains the simulation for an arbitrary time span.

The above observations are quantified in Table 6.19 on average over multiple, independently trained networks. The calculation of prediction errors was skipped here, since there is no driving signal to compare with, and since the simulation may not run synchronously with, or not on the exact same time scale as a hypothetical driving stimulus, which to compensate is nontrivial.

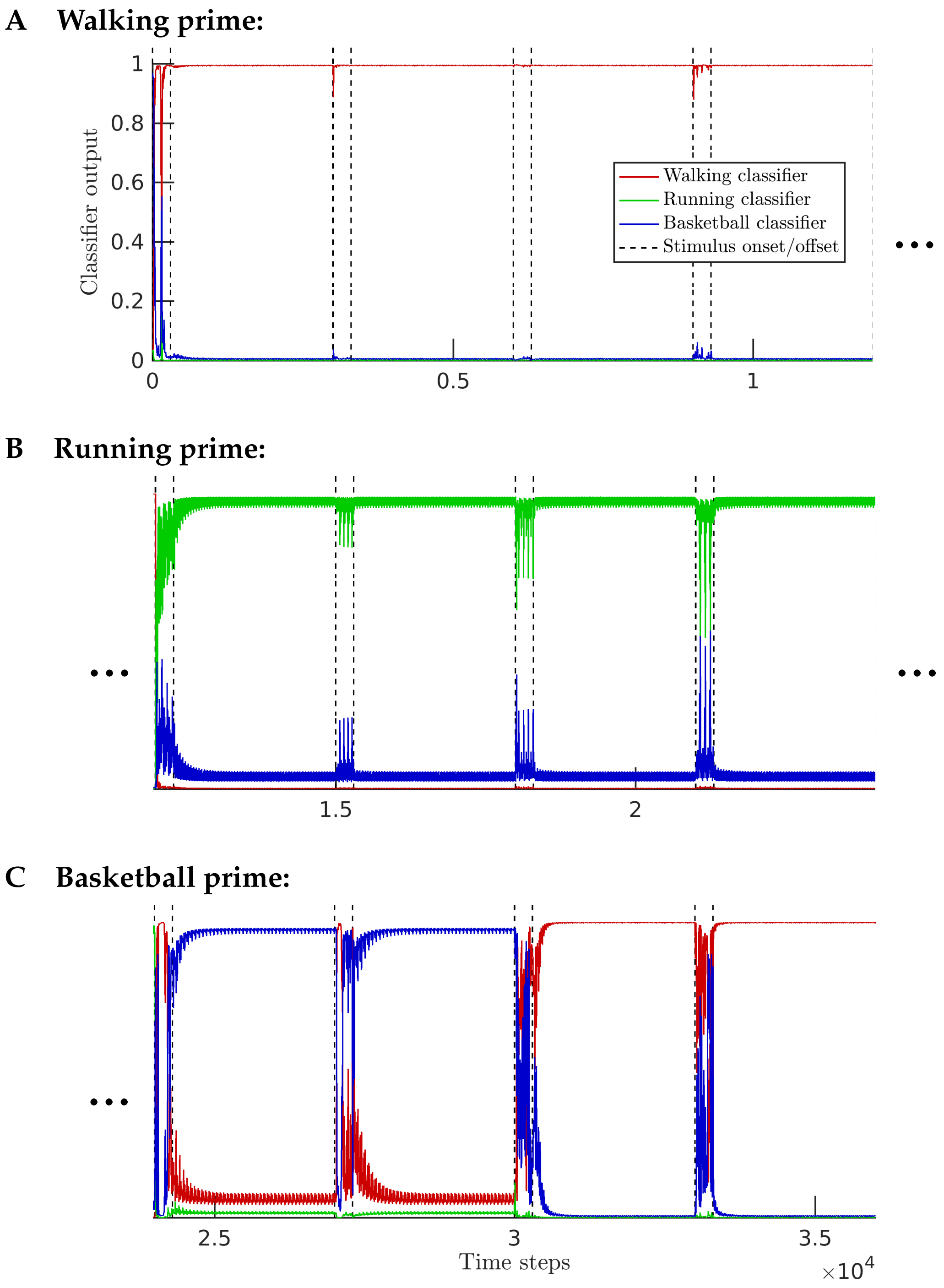


FIGURE 6.27: Classifier output during the simulation priming experiment. For each of the test trials (4 of each type in **A**, **B**, and **C**), the model is driven by a random segment of the trial for 300 time steps, and is then configured to simulate activity in all three modules without stimulus. The classifier output converges to the shown class, respectively, except for the last two basketball segments because of the ambiguity of the basketball classification.

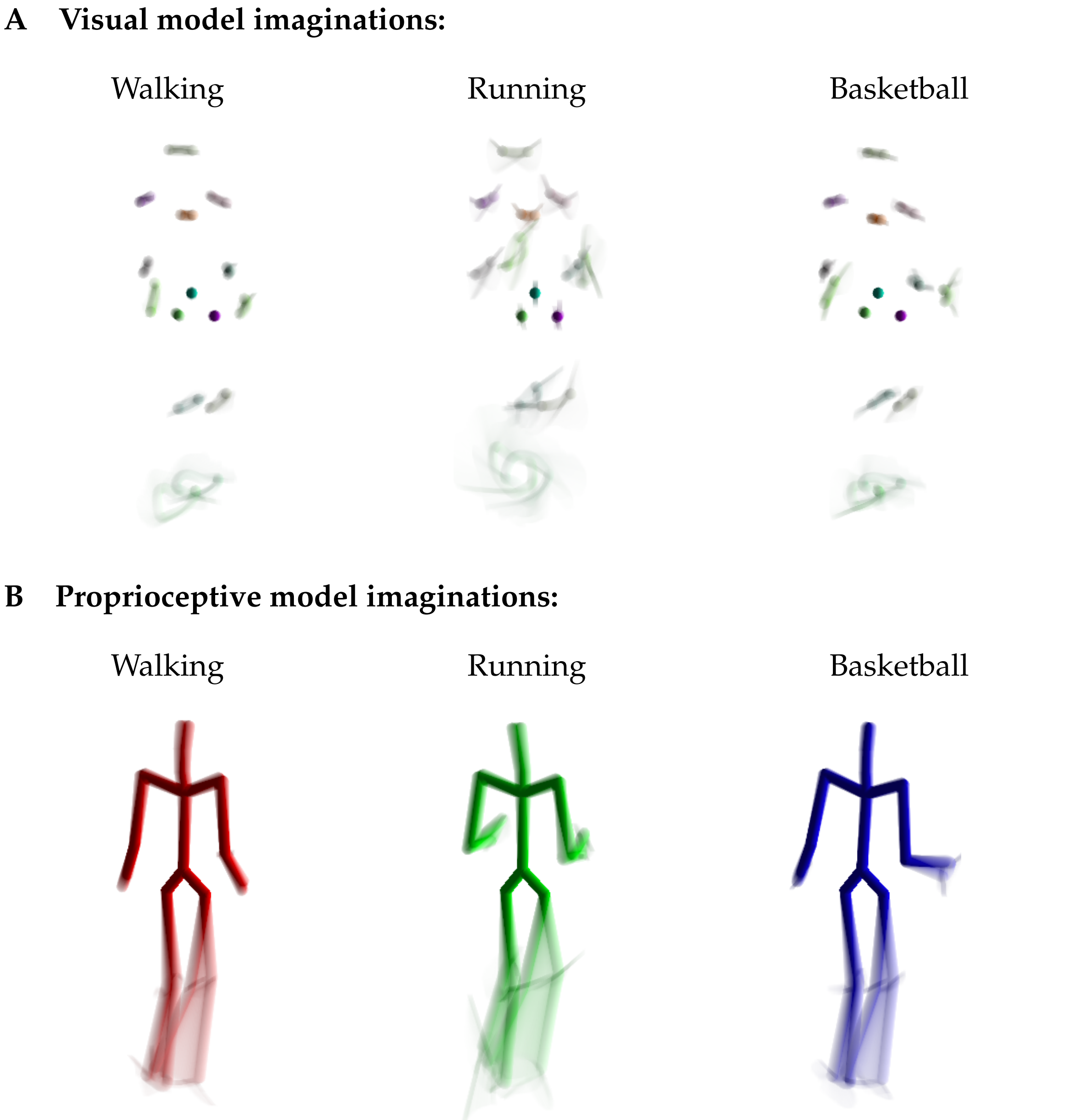
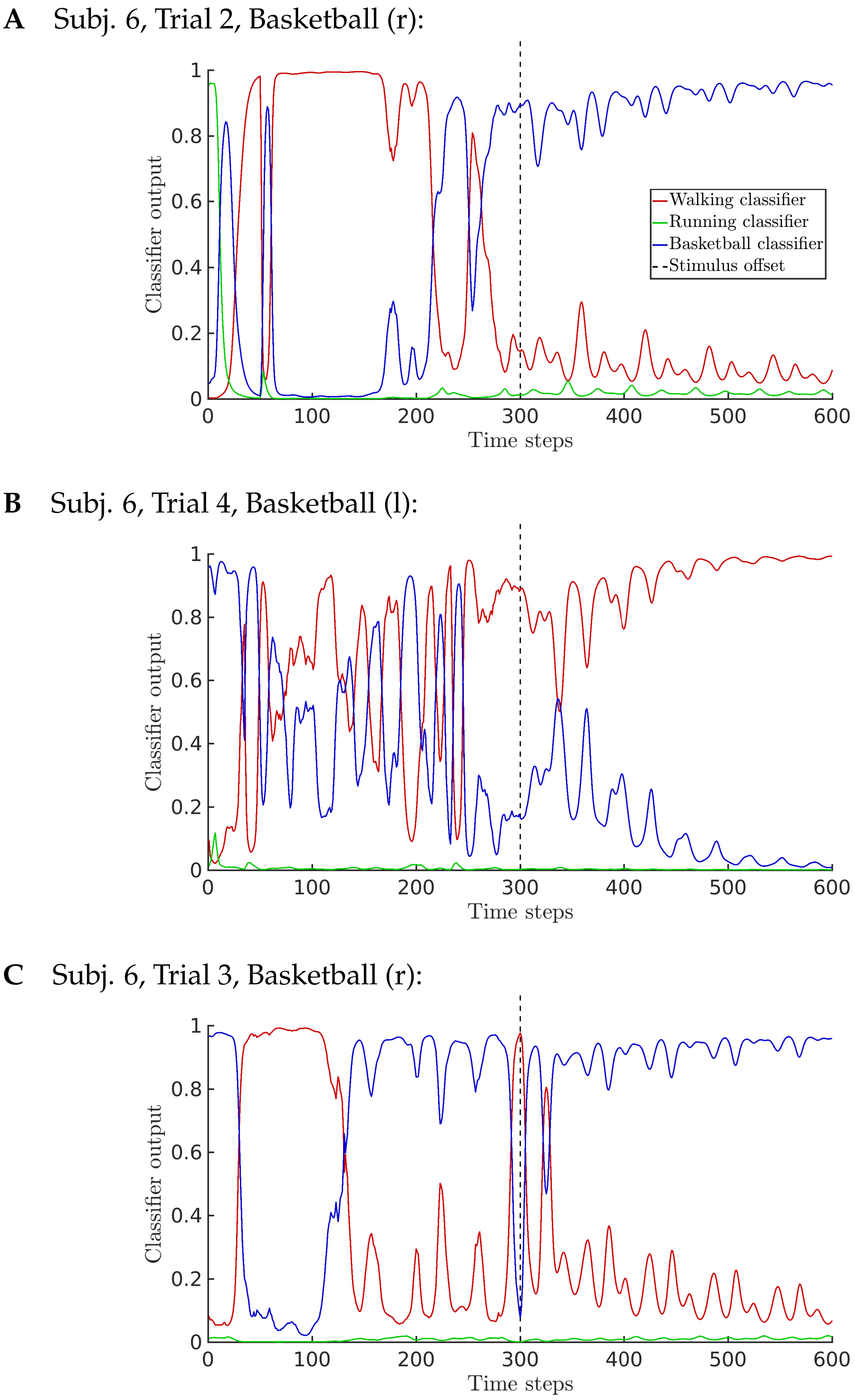

A Visual model imaginations:
Walking
Running
Basketball
B Proprioceptive model imaginations:
Walking
Running
Basketball

FIGURE 6.28: Multimodal imaginations after simulation priming. Depending on the internal attractor the model converged to, the model expectations form three different cycles that represent walking, running, or basketball dribbling. Converged attractors are generally stable and self-preserving, and the visual (**A**) and proprioceptive (**B**) imaginations are temporally synchronous and consistent. Similar results appear without priming when only a constant class bias is provided to the model.



A Subj. 6, Trial 2, Basketball (r):

B Subj. 6, Trial 4, Basketball (l):

C Subj. 6, Trial 3, Basketball (r):

FIGURE 6.29: Case differentiation for the convergence of the model's simulation after priming by a basketball stimulus. The model typically settles to the attractor that corresponds to the classifier that was recently most active (**A** and **B**). Intermediate classifications do not determine the convergence to the respective class (**C**).

TABLE 6.19: The results of classification after convergence in the simulation priming experiment. As shown, the variance is very low and the convergence is stable. The results were obtained by averaging over four independently trained network instances.

Subj.	**Trial**	**Class**	**M. CE**	**Var. CE**	**% Corr.**
5	1	walking	0.337	0.00703	75.7
6	1	walking	0.331	0.0127	76.2
10	4	walking	0.31	0.0312	78.1
12	1	walking	0.339	0.00599	75.5
2	3	running	0.0429	0.000188	100
16	46	running	0.0411	0.000109	100
35	19	running	0.0411	0.00011	100
35	22	running	0.0412	0.000112	100
6	2	basketball (r)	0.735	0.000512	50
6	3	basket. (r)	0.738	0.000377	50
6	4	basket. (l)	1.07	0.000575	25
6	5	basket. (r)	1.07	3.09e-06	25
		average	0.424	0.0049	71.3

Despite the successful formation of consistent, distributed, basically predeterminable simulation attractors in the network, the percentage of trials that converged to the class of the respective prime (% Corr.) in Table 6.19 reveals two more phenomena: Firstly, in three out of four tested networks, some of the basketball primes did not lead to convergence to the basketball attractor in the long run. In these cases, the network converged to the walking class after the stimulus offset. Seeing that the classification of the basketball test trials was already ambiguous in the previous experiments, it is coherent that also the convergence to an internal attractor, primed by these basketball trials, is ambiguous. This is exemplified in Figure 6.29, showing multiple, possible scenarios: In Figure 6.29 **A**, the network classifies the observed stimulus as basketball at the stimulus offset. The classification (which is driven solely by the internal modal simulations from then on) then converges to the basketball class. Figure 6.29 **B** shows the opposite case, when the basketball stimulus is currently classified as walking at the stimulus offset, such that the network converges to the walking attractor. Thus, the ambiguity of the basketball test trials results in ambiguities in the imagination of the network after priming. In Figure 6.29 **C**, the network converges to the basketball attractor although the network momentarily classified the observation as walking at the stimulus offset. This shows that not only the current classification, but rather the recent average of classifications determines the convergence, since both the classification and the autoencoders' predictions follow from the history of activated submodal codes. Taken together, the classifier performance is typically the most crucial factor for simulation priming in the model. When the stimulus is a randomly selected snipped out of the whole test trial, then the success rate for simulating ahead in the correct action class is typically about equal to the classifier correctness when showing the whole trial (which corresponds to the performance shown in Table 6.15 in Section 6.5.1).

Secondly, however, the model was not trained in particular to converge to the nearest attractor in terms of the classification error during imagery tasks. In fact, there is no objective reason for the network not to converge to a different, stable attractor. Based on the random initialization of weights, and based the random application of error gradients during training, also network instances with a more or less distinct simulation bias towards specific classes resulted. One of the tested networks was biased towards simulating the basketball action, such that it converged to the basketball attractor whenever a walking trial was shown, which condenses in the lower overall percentage of class correctness for walking in Table 6.19. To avoid networks with biased imagination characteristics, it might be helpful to introduce mechanisms that stabilize the classification after priming, which in turn stabilizes the modal simulations. This can be thought of as deciding on a class to provide a constant, top-down bias for the subsequently imagined action, which is evaluated in the following.

6.6.2 Top-down Biased Simulation

TABLE 6.20: Parameters for perception chosen for the evaluation of simulation biasing.

Parameter	Description
$\bar{\mathbf{i}}$ = teacher forcing	Class bias in autoencoders
$q^{vp/d/m} = 0$	Visual stimulus reliabilities
$q^{pp/d/m} = 0$	Proprioceptive stimulus reliabilities

This evaluation shows that the self-preserving imaginations of the network can be determined solely by providing a constant, top-down class signal. The class

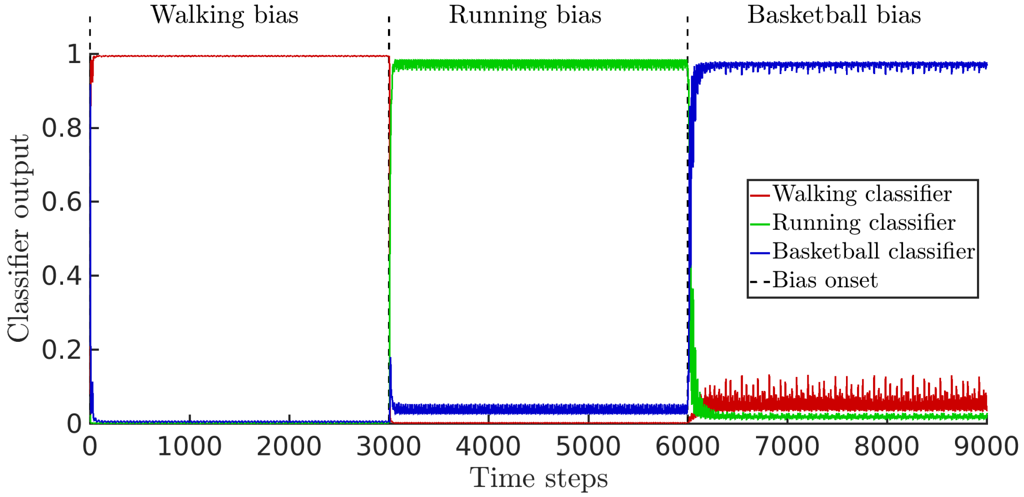


FIGURE 6.30: Classifier output during the simulation biasing experiment. The same (model driven) classifications and simulation attractors are reached when the top-down class biasing in the submodal autoencoders is provided.

signal biases the distributed predictions, such that they will step out of their current attractor and converge to the provided class attractor. For this experiment, each of the three classes is tested for 3000 time steps, in which no sensory stimulus is provided, while the intention biases in the autoencoders are overridden with the desired class respectively (teacher forcing). The parameters for this evaluation are shown in Table 6.20.

TABLE 6.21: The results of classification after convergence in the simulation biasing experiment. The results were obtained by averaging over four independently trained network instances.

Class	**M. CE**	**Var. CE**	**% Corr.**
walking	0.0177	0.000164	100
running	0.0387	7.84e-05	100
basketball	0.0905	0.000328	100
average	0.049	0.00019	100

Again, the stability of the converged (model generated) classification (see Figure 6.30 and Var. CE in Table 6.21) was similar to results of the simulation priming experiment. This further substantiates that the model uses the top-down classification biases for all modal predictions consistently, and that changing only

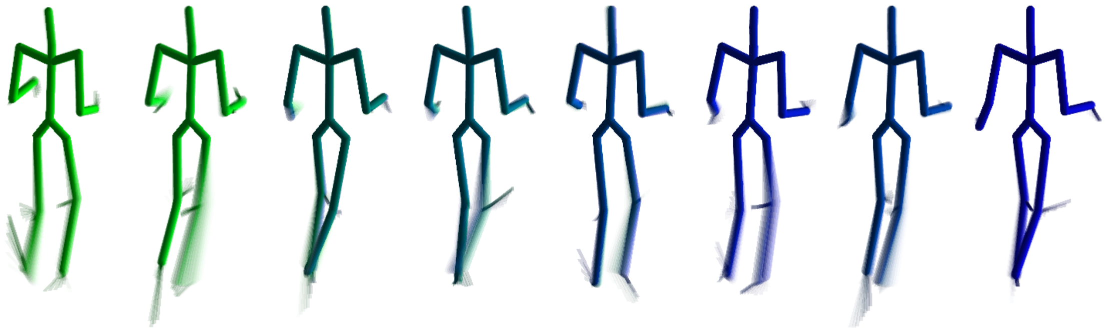


FIGURE 6.31: An imagined transition from a running to a basketball movement. When the top-down bias changes (color coded), the overall network transits to the respective class. Implicit inter-classes are encoded in the learned predictive encodings.

a single, binary, constant input is sufficient to enforce transitions between the learned attractors. The tested network instances were identical to the instances used in the last experiment where one of the networks showed a convergence tendency towards the basketball attractor. Here, however, all desired attractors were reached correctly given the constant top-down bias, confirming that a simulation bias can be prevented by stabilizing the intention inference.

The imagination of inter-class actions can be observed when transitioning from one top-down bias to another, as shown in Figure 6.31. Coming from the running class, the network transits to the basketball class. In the meantime, while the imagined locomotion and step cycle remain intact and slow down a little, the arm postures smoothly transit from typical poses for running to typical poses for basketball dribbling.

6.6.3 Conclusion

In the above sections, it was confirmed that the model is able to generate self-sustaining, continuous, multimodal and multi-submodal simulations also without sensory stimulation. In separate modules, predictive encodings and temporal representations enabled consistent imaginations that corresponded to kinematic intentions, and all of the evaluated simulation attractors were stable in themselves. When the simulation was primed by the presentation of a biological motion stimulus, the network typically converged to the nearest simulation attractor in terms of the classified action class afterwards. It was found that the model can have a simulation tendency towards a specific class after priming, which can however be compensated by stabilizing the top-down intention biases. Taken together, the results provide further support for the theories of embodied simulation via predictive coding, by showing how it is possible to prime embodied simulations via bottom-up stimuli, or bias them via top-down, higher level representations. The results furthermore suggest a functional link between action understanding and mental imagery.

6.7 Test Evaluations 4: Tracking Perspectives and Understanding Novel Movements

TABLE 6.22: Parameters for perception and adaptation chosen for the evaluation of tracking and understanding novel movements.

Parameter	Description
$\eta^s = 0.01$	Adaptation rate of the origin of the visual FOR
$\gamma^s = 0.85$	Momentum of the adaptation of the origin
$\eta^r = 0.05$	Adaptation rate of the orientation of the visual FOR
$\gamma^r = 0.85$	Momentum of the adaptation of the orientation
$\eta^w = 0$	Adaptation rate of the feature selection and assignment
$\gamma^w = 0$	Momentum of the feature selection and assignment
$\mathbf{b}^{\text{data}} =$ allocentric	Data translation offset
$A^{\text{data}} =$ allocentric	Data rotation offset
$w_{ij} =$ not provided	Feature binding weights
$\bar{\mathbf{i}} =$ inferred by model	Class bias in autoencoders
$q^{vp/d/m} = 1$	Visual stimulus reliabilities
$q^{pp/d/m} = 0$	Proprioceptive stimulus reliabilities

This experiment evaluates if and how the model can understand actions that do not belong to the learned action repertoire. To do so, the model's performance is analyzed while being stimulated with three movements of the *second* test set as described in Section 6.1. All adaptive and predictive model components are activated in this action understanding task. That is, the model is to infer the perspective, feature binding, the type of the observed action with respect to the learned action repertoire, and to simulate proprioceptions that correspond to the observations.

Furthermore, in contrast to the previous experiments where the model inferred a *constant offset* of rotation and translation on the data, here, additionally a *dynamic, intrinsic orientation* of the data has to be compensated by the model on the fly. In the selected test trials, the body orientation of the subjects is partially not constantly aligned to a specified orientation in the visual frame of reference as is was approximately the case during training and in the previous experiments. Thus, the experiment investigates whether the model is able to also *track* perspectives dynamically, although the observed action is unknown to the model. To this end, the adaptation rate η^r for the model's rotation matrix was slightly increased. The respective parameters for this experiment are shown in Table 6.22.

The tested trials of the second test set involve three different types of actions: A lambada dance trial, a jumping trial, and a waiting-for-the-bus trial. In the *lambada dance* trial[3], a single subject (without dancing partner in the data) performs short forward steps, alternating, unilateral sidesteps, and is swinging the hips while the arms occupy a partner dance pose (left arm holding the dancing partner at the torso, right arm holding the dancing partner's hand up high). The global upper body orientation is relatively constant in this trial, though relative to the hips, the orientation varies greatly. In the *jumping* trial[4], the subject stands on the ground in the beginning, then jumps up on both feet four times. They proceed jumping on the left foot four times, then on the right foot three times. Subsequently, the subject jumps on both feet again two times and crosses or touches the feet in mid air. The upper body orientation remains relatively constant. In the *waiting* trial[5], the subject rests their hands on the hips, briefly looks at the clock, leans forward, and scratches their head while (seemingly impatiently) stepping forwards and backwards, constantly changing the movement direction. In contrast to the other trials, here, the orientation of the whole body changes quickly and multiple times, covering a range of about 180° around the vertical axis.

Table 6.23 summarizes the results of perceptual inference given these stimuli after the model was trained on walking, running, and basketball dribbling. The

[3] see video at http://mocap.cs.cmu.edu/subjects/55/55_02.avi as of 08.01.2018

[4] see video at http://mocap.cs.cmu.edu/subjects/49/49_03.avi as of 08.01.2018

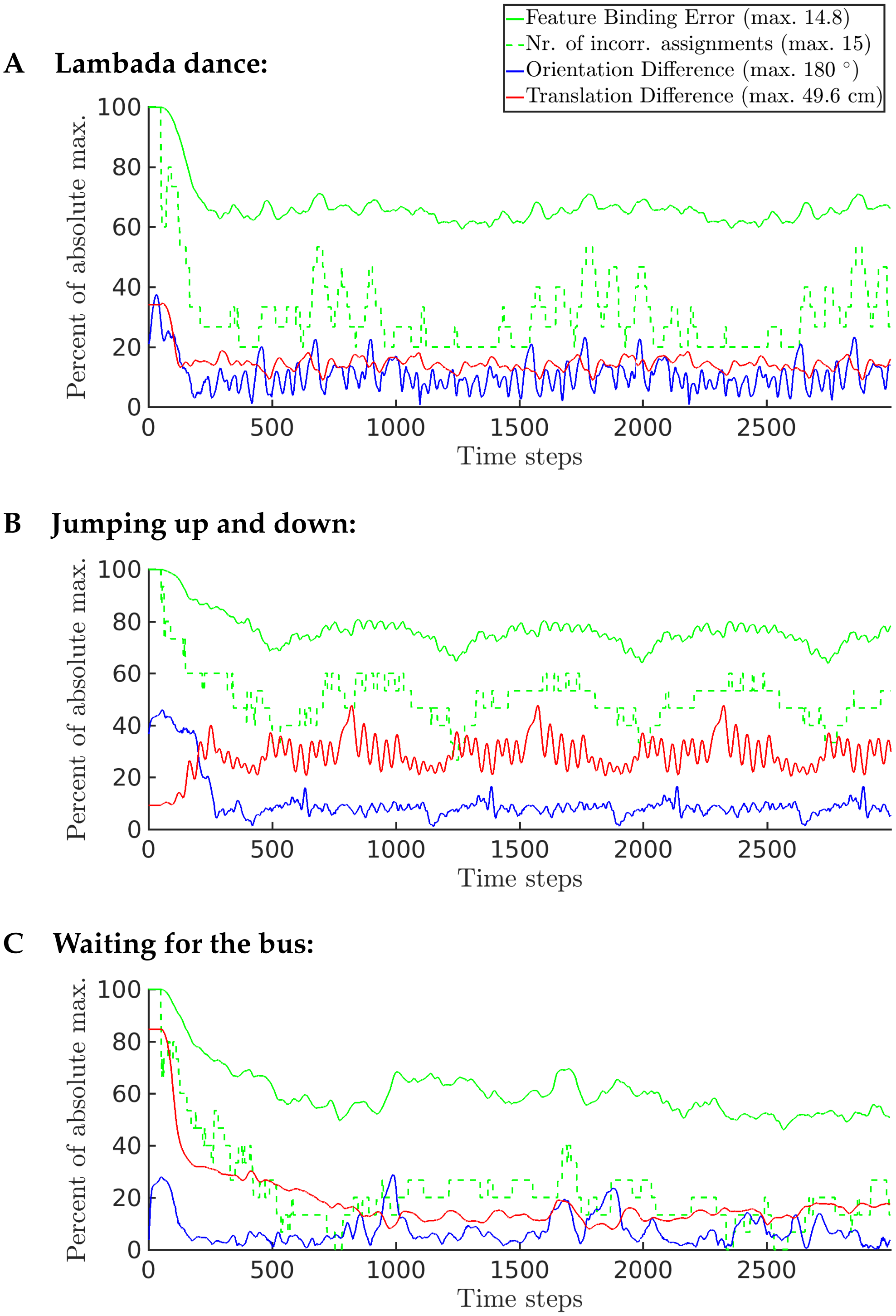
[5] see video at http://mocap.cs.cmu.edu/subjects/40/40_11.avi as of 08.01.2018

TABLE 6.23: The derived perceptual inference measures after convergence in the combined perspective-taking and feature binding task with novel movements.

Subject	Trial	Class	Mean OD	Var. OD	Mean TD	Var. TD
			Mean FBE	**Var. FBE**	**Mean IA**	**Var. IA**
55	2	Lambada	18.0	58.0	8.524	0.0602
			9.53	0.148	4.23	1.94
49	3	Jumping	15.0	17.12	15.05	0.518
			10.9	0.298	7.14	0.934
40	11	Waiting	10.7	45.9	7.182	0.0882
			7.82	0.202	2.15	1.07

average orientation difference (here, with respect to the actual upper body direction) was relatively low for all trials after perceptual inference (about 10 to 18°). The average translation difference was between 7 and 15 cm, which is considerable more than for the known trials, but still not critical. As well, most of the features were identified and grouped correctly: About 2 to 7 out of the 15 visual features were assigned incorrectly. The feature assignment error was nonetheless considerably higher than observed in previous experiments, speaking for the increased uncertainty in the assignments. Looking at the numbers, perceptual inference can be considered successful, despite the fact that the network had never seen similar actions before. Thus, the introduced information coding and learning schemas provide general templates for perspective-taking, and seem to be suitable for bootstrapping general action recognition.

To that end, the network infers an error-optimal interpretation of its sensations with respect to its embodied encodings. As shown in Figure 6.32, this leads to a minimization of the derived error measures towards a relatively stable local minimum also for novel movement stimuli. For the lambada trial, as illustrated in Figure 6.32 **A**, all derived measures are relatively constant after the minimization of expectation errors, except for the assignments. The model identifies most features correctly, but loses some of the assignments for a short time while observing the sidesteps, and recovers while observing the forward steps. In comparison to the forward steps, whose motion dynamics are somewhat known from walking, the sidesteps result in unknown motion dynamics, especially due to the simultaneous hip swing. Thus, the hip joint assignments are partially lost. The sidesteps also result in a high variance in the body orientation,



A Lambada dance:

B Jumping up and down:

C Waiting for the bus:

FIGURE 6.32: The performance of the perceptual inference model when novel movements are observed. All tested trials have stable perceptual attractors, and thus the perspectives can be inferred. Because of the wide postural differences the learned trials, not all of the features are correctly recognized, depending on the specific type of action. The graphs are normalized with regard to the respective absolute maximum defined in the legend.

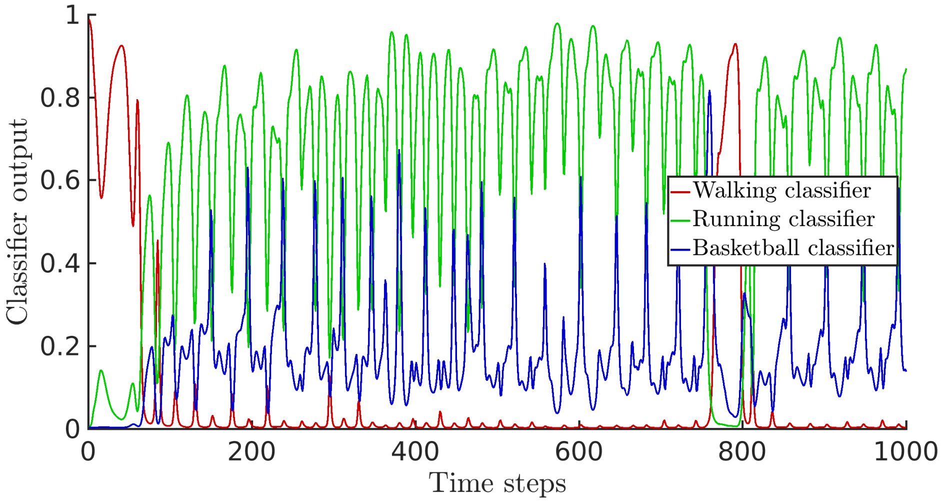


FIGURE 6.33: The classifier output while observing a jumping trial. As observed before, the subject standing around is classified as walking in the beginning. When they start jumping, this is classified as running because of the whole body upwards/down dynamics. At the transition from falling downwards to jumping upwards (every second green local maximum), the model briefly classifies basketball, triggered by the learned, basketball specific right hand gesture.

which is compensated by the model nonetheless. Other features that are not assigned correctly or lose the assignment temporarily are the right arm and left hand because of the unknown posture. The lambada dance trial is classified as walking by the network, which is the typical default class learned by the network. The proprioceptive system is set into a less articulated walking resonance every time the subject steps forwards or sidewards.

In the jumping trial, as seen in Figure 6.32 **B**, the orientation is constantly inferred with relatively high precision, since the subjects body orientation does not drastically change while jumping. Only about half of the features are assigned correctly, since all of the perceived motion features represent strict upwards or downwards movements, which are primarily known to the model from the shoulders, thorax, neck, head, and hips while running. The other features are partially and temporarily inferred by the model. It obtains the correct assignments of the legs while jumping up on either of the legs likely because of the similarity to the running posture, and loses the assignment again while jumping on both legs and and crossing the feet in the air, which, again, implies

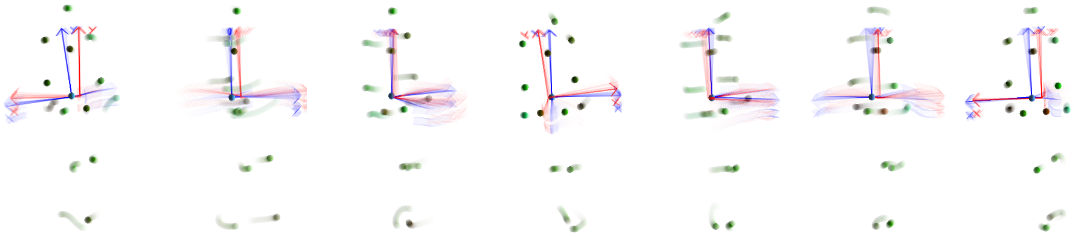


FIGURE 6.34: Tracking the perspective and binding the features in the waiting-for-the-bus trial. Shown are the consecutive linear moving averages of screenshots, each averaged over 50 time steps.

a sideways leg dynamics unknown to the model. Figure 6.33 confirms that the model uses motion features known from running for its perceptual inferences: The classifier outputs walking in the beginning while the subject stands and does not move. While the subject is moving upwards or downwards, the model classifies this as running, because of the similarity of the motion features. Interestingly, in the moment the subject jumps up *again*, this is classified as basketball for a short time, since the discriminative feature for the basketball classifier is triggered by the right hand downwards-upwards movement. As a result, in Figure 6.33, the basketball classifier is activated every second time the running classifier is at a local maximum. At every other second time, the network does not interpret the opposite case, the upwards-downwards dynamic, such that the classifier output is unnormalized for a short time. Consequently, also the proprioceptive system is set into resonance related to the running movement: The simulated posture occupies an average walking posture, the arms move up and down, and the model predicts motion of the legs, while however, no postural changes of the legs can be observed.

Interestingly, the waiting-for-the-bus trial shows the lowest of all average orientation differences in this experiment, although it shows the most variance in the body orientation. Figure 6.34 shows that the model is able to track the orientations of the subject with relatively high precision even while they change their orientation. As shown in Figure 6.32 **C**, even all features are identified correctly for short periods of time. This is due to the similarity to the postures and motion dynamics known to the model from walking. Consequently, the waiting trial is classified as walking by the network (just as the lambada dance).

The proprioceptive module starts to resonate when the subject is sporadically stepping forwards, mimicking a minorly pronounced walking movement.

6.7.1 Conclusion

The foregoing results show that the model is able to infer and track the perspective of an actor robustly, and bind features to a degree that depends on several factors, when the observed action is rather dissimilar to the actions encoded during training. Thus, the embodied action priors are suitable for generalized perceptual inferences. Resonance in the proprioceptive system can be determined when the observed action contains similar action patterns to those seen during training – akin to neuroscientific findings that substantiate that mirror neurons only respond to observed actions that belong to the motor repertoire of the observer. Furthermore, the results clarify that the network learned sensible (e.g. gestural) features from the training data for the inference of kinematic intentions.

Chapter 7

Conclusion and Discussion

In this thesis, I reviewed neuroscientific and psychological studies on action understanding that led to dilemmas and open questions which are still highly disputed today. Subsequently, I raised and combined fundamental theories and findings to establish a general hypothesis on action understanding. After relating to own previous work and the work of others, a neurocomputational model was implemented that aimed at investigating the tenability of the hypothesis and its partial aspects, while identifying functional principles and candidate mechanisms for action understanding. The model shows how different types of self-experiences can be learned and linked in distributed, generative and predictive encodings. This essentially led to the ability to consistently imagine and simulate the experienced self-perceptions. Moreover, the model was able to reenact observed actions in the learned, embodied terms, activating the same simulations, and inferring the kinematic intentions of the observed actions. The correspondence from observations to self-representations was actively established by top-down expectation driven visuo-spatial adaptations that refer to perspective-taking and feature binding. The model thus provides support for the direct matching and embodied simulation hypotheses within a framework of predictive coding and multiple types of structural biases and perceptual inferences. Based on the results and evaluations, several insights are gained, and statements can be made, which I conclude in Section 7.1.

Despite these successes, the neural network model is by far not a complete model of action understanding and its cognitive development. Several simplifications were applied, and problems were sidestepped, such that not all of the open questions on this topic have been tackled to the last detail. Both – new insights as well as simplifications – however offer potential and opportunities for further investigations, which I put forward in Section 7.2.

7.1 Insights and Statements

The proposed neural network model was able to learn compressed action patterns from submodal self-perceptions. The developing spatial codes covered the contingencies in their respective domain independently. To link the encodings, the model learned bidirectional, temporally predictive encodings within and across the modalities. As well, the modal encodings were successfully linked to abstract, kinematic intention classes, which then were utilized to learn biases for the submodal code predictions. For learning spatial codes, their temporal correlations, and classifications robustly, several aspects turned out to be crucial. Submodal contingencies by means of spatial Gestalt patterns had to be learned first, followed by the biased temporal predictions and classifications of the Gestalt codes, to avoid that the target mapping for the latter is constantly changing. Furthermore, autoencoders are prune to recoding of already learned codes during online training on continuous inputs, particularly when the inputs are very smooth such as the biological motion stimuli utilized here. Representing the inputs by specifically parameterized Gaussian neural populations reduced the problem of recoding significantly. As well, a clear distinction between learning and adaptation was raised: While adaptations gradually alter the model's momentary perceptual processing to match top-down expectations, learning alters the encoded expectations and their correlations in longer terms. During training, a direct influence of expectation errors on the model's temporary perception was explicitly avoided by applying error gradients at random time steps, while in contrast it was explicitly urged during adaptation. With the use of these methods, codes were learned to optimize expectations permanently, effectively reducing recoding and catastrophic forgetting to a minimum, while adapting the perception to match expectations momentarily.

In evaluating the trained model, I have shown that the correspondence and binding problems can be solved by top-down driven perceptual online adaptations. The learned expectations focused on specific features of biological motion. Moreover, the evaluation reveals a functional dependency between the orientation a movement is observed from, as well as recognition rate and time. These results qualitatively match the results of psychometric studies on biological motion perception and mental rotation. Thus, mental transformations of the own body scheme onto an observed person (as also assumed by Kessler and Thomson, 2010) can indeed be considered a *"functional bridge between first-and third-person perspectives"* (Decety and Meltzoff, 2011) that explains the activation of mirror neurons. For solving the binding problem, the introduced population coding was yet again substantial for resolving perceptual ambiguities and conflicts, which was not possible using Cartesian coordinates in exploratory studies. Another vitally important component was the segregation of input stimuli into their submodal components, and the weighting of the respective submodal expectation error signals. The submodal components provided specific spatio-temporal invariances and inference properties such that they established constructive information priors for binding observed features into Gestalt percepts and establishing the correspondence between reference frames.

The second evaluation of the trained model revealed how action observations can activate simulations of learned embodied states. It was cross-validated that the modules of the network can predict each other's modal activity selectively, and that the predictive components utilize local as well as distal information in an appropriate manner. Thus, the concept of motor cognition was verified in the model, in that it does not only visually recognize an observed action after perspective-taking and feature binding, but also that action observations expose attractors to the overall, dynamical and recurrent neural system to the effect that unobservable action encodings are brought into resonance, simulating and reenacting the observed actions in embodied terms. To obtain the results, it was again important not only to avoid recoding, but also to fine-tune the learning of spatial codes, and predictions and classifications thereof. Determining separate learning rates for spatial and temporal training of each autoencoder, for the top-down biasing of predictions, and for the intention classifier was a critical factor. Although the proposed population coding already accounted for comparable levels of activity between the submodalities, the different variabilities and configuration spaces of the submodal inputs still had an influence on the respective learning performance. Given the proposed parameterization of the model, neither of the developing submodal encodings had an excessive influence on the predictions, the biasing thereof, nor on the inference of kinematic intentions. In fact, the model seemed to learn descriptive short-term action patterns for its inferences, as specific as a hand gesture to detect a basketball action.

Furthermore, I have shown that the same principles of distributed predictive coding are suitable for preserving consistent simulations of multiple action classes also without any sensory stimulation, while the simulation can indeed be primed by a preceded observation. Thus, multiple, stable, and self-preserving attractor state sequences were formed by the model, while the top-down kinematic intention bias was appropriate for modulating the simulation and generating transitions between the simulated action classes. The results confirm that embodied simulations can be linked to visual and motor imagery, and that both might indeed be core mechanisms of action understanding. To obtain stable attractors when activity in the model was completely self-preserving, it was most important to predict the absolute codes (not the difference to the last prediction) from a time window of preceded codes, and to use non-linear activation functions for the code layers. The absolute prediction and non-linearities provided robustness to error accumulations in the predictions, effectively avoiding divergence from the normal codomain of the codes. Similarly important, the utilization of a logarithmic time window of recently activated codes for the predictions was beneficial for noise robustness and necessary for detecting more complex temporal dependencies.

At last, it was shown that the model learned general perceptual biases for biological motion recognition and action understanding, in that it was able to determine the perspective and, to a certain degree, also the feature identity of completely novel actions. The model also maintained the feature inference and tracked the frame of reference over time when it changed. Moreover, the model simulated corresponding proprioceptions when the observation corresponded to a known self-perception, and it interpreted the observed kinematic intention in terms of its learned action repertoire. Again, this result is reminiscent of mirror neuron properties, in that they are as well assumed to interpret observed actions in terms of the own motor repertoire, and in that they are nonetheless assumed to be involved in imitation learning. Thus, the results obtained in this experiment set the stage for speculations about which role the proposed action understanding mechanisms could possibly play in social learning.

Based on the successful verification of the hypothesis suggested in Chapter 3, and based on the applied processing structures and methodologies that were successful in obtaining the above results, I can provide an *opinion* on some of the big open questions that result from the dilemmas of action understanding as described in Section 2.1:

- Are mirror neurons a byproduct of mental development, or is there a genetic predisposition? What is their purpose?

  The actual *role* of mirror neurons – whether and how they take part in mental development – can not be identified directly using this model. Taking into account the apparent structural biases of cognition that were considered for the proposed model, it seems likely that there is a genetic predisposition for the development of mirror neurons which *unfolds during and contributes to mental development*, rather than being a byproduct of it without particular purpose. As shown in the results, the contribution to mental development could indeed be to recognize already learned actions, and 'understand' visually unobservable characteristics of the actions (via motor preparations and intention inferences) from a self-centered perspective on the basis of the own embodied experiences. Although the model can be considered to be particularly build to support this claim, it shows that it is indeed *possible* that mirror neurons facilitate action understanding by this means. Embodied simulations, of course, should encompass a greater variety of action-related encodings than shown in the model, which might after all lead to the development of empathy and emotional intelligence, as suggested by Rizzolatti and Craighero, 2004; Rizzolatti and Craighero, 2005; Rizzolatti and Sinigaglia, 2007.

  It could be taken as counterevidence against this claim that infants are able to attribute goals to observed actions that they are not yet able to perform themselves (Kamewari et al., 2005). Not necessarily all forms of action understanding may emerge from visuomotor couplings. Similarly, understanding goal states of observed actions may be possible also without simulating motor activity, and without taking the perspective of an observed actor. There are certainly several ways and forms of action understanding, but it is likely that one of the forms is motor preparation, facilitated by encodings related to mirror neurons, which is a view also shared by Rizzolatti and Craighero, 2004 (see also Hickok, 2009). Hence, the primary role of mirror neurons might be action understanding, but action understanding is not facilitated solely by mirror neurons, as they represent merely an aspect of it.

- Do mirror neurons facilitate imitation learning? Or is there an innate mechanism?

  The model did not directly investigate imitation, nor learning by imitation. In the training procedure, the executed actions were provided, and shown from an egocentric perspective, offering full visual, proprioceptive, and intention information for developing mirror-neuron-like tunings. Thus, the approach followed the assumption that mirror neurons only encode goal-directed actions that belong to the own motor repertoire already, and thus develop *after or while* learning to execute the respective actions.

  Nonetheless, it can be assumed that cognitive processes involved in the formation of mirror neurons help in learning by imitation. While mirror neurons are apparently involved with action understanding rather than learning by imitation, cognitive processes that activate mirror neurons seem to be strongly entangled with visuo-spatial abilities, as extensively supported in this thesis. The results in Section 6.7 have shown that visuospatial perspective-taking is also possible if the observed model performs an *unknown* movement. The observer may thus take the perspective of an (either observed or memorized) actor who performs a novel movement, and then minimize the difference between the observed (and mentally adopted) posture of the actor and the own posture by applying motor commands. This results in imitation, after which the observer is able to establish embodied visuomotor correspondences and also visuomotor goal encodings for the observed action, resulting in *learning* by imitation. After this, they will be able recognize the action when re-encountering it and to prepare motor programs, such that the developing neural codes ultimately form mirror neurons for that action and its understanding by means of embodied simulations. Thus, in a sense, imitation might precede the development of mirror neurons, but similar perceptual mechanisms might be involved.

Once again, this may be recognized as a bold assumption by some who argue that perspective-taking develops later in life than imitative capabilities. For example, studies indicate that children make mistakes in the 'three mountains problem' up to an age of 9 or 10 years (Piaget, 2013), where they are to select pictures of a model of mountains that correspond to a specified vantage point. However, this task requires complex spatial visualization abilities, and the results may also be influenced by context and language understanding to a certain degree. More basic visuo-spatial abilities develop indeed very early in life. For example, Moore and Johnson, 2008 show that (male) infants are capable of mental rotation as early as 5 months of age, long before they seem to imitate simple actions (Jones, 2009) (other than facial expressions which is disputed but might be innate).

Although with these words, I can neither endorse nor refuse the existence of *innate* mechanisms for imitation, I argue that visuo-spatial abilities and active inference may bootstrap learning by imitation, and that mirror neurons may be the developmental *result*. Thus, multimodal action encodings might be learned nearly in the same way from ideomotor and exploratory learning, as they are learned from imitative behavior, the difference being that visuo-spatial perspective-taking is involved in the latter as described. Note that visual representations of actions (i.e. biological motion in STS) are still not necessarily limited to egocentric reference frames. Canonical views of actions (for example a view of the left side of a walking person) may develop as a result of observation *and* learning by imitation, which can nevertheless activate self-centered representations by means of spatial transformations.

- What comes first, learning by imitation, or learning how to imitate?

  Under the assumptions just made, and given that the hypothesis of the model is consistent, the answer is *fairly both*: Provided that visuo-spatial abilities facilitate imitation to form embodied visuomotor experiences, and provided that visuo-spatial abilities are driven by the very same developing action experiences, then clearly both processes are intertwined and synergistic. Children, however, first learn about their own body already in the womb, and then proceed to imitating others. Thus, it can be assumed that self-observations and embodied encodings provide the necessary perceptual biases that later drive the ability to imitate others, as also suggested by the model results.

- Which mental processes equate action and perception? What establishes the correspondence between observations of others, and self-representations?

  The suggested principles of information processing provide a solution to the correspondence problem: Given that an observed action activates some embodied expectation in an egocentric reference frame – by means of a broad and even fuzzy perceptual bias – the model can infer the perspective of the observed actor by minimizing the expectation error via mental rotations and translations. If and to which degree such an embodied, driving signal is activated might indeed be connected to social cues like racial bias (Lamm, Batson, and Decety, 2007; Avenanti, Sirigu, and Aglioti, 2010). The progress of perspective-taking then concretizes the activated embodied expectations and thus synergistically improves its own driving error signal until convergence of the perspective, as shown in the experiments. Similarly, the model provides a solution to the binding problem: Submodal top-down expectations with multiple perceptual invariances are utilized to establish the correspondence between observed visual features and own bodily features. This, again, improves the expectation by itself, such that the relevant features can eventually be selected, ordered, and grouped into stable, embodied Gestalt percepts. Both processes are closely related to spatial abilities and attention.

- Is there even a mirror neuron system in humans?

  The study of Lingnau, Gesierich, and Caramazza, 2009, which doubted that there is a mirror neuron system in humans, resulted in thorough debates. Although neglecting the existence of a mirror neuron system in humans by not being able to replicate another study that finds direct evidence for it (Chong et al., 2008) might be a false syllogism, the study and the subsequent debates discussed several other possibilities how humans can be able to understand actions, such as context-sensitive inferential processes.

  Relating to the model, there is as well no particular technical reason to infer motor encodings solely from visual observations, and intentions solely from motor encodings – *unless* the observed actions are going to be imitated. Hence, although mirror neurons apparently do not directly enable *learning* by imitation, they might very well be involved in imitation and the processing of social cues. Nonetheless, in the model, intentions could also be inferred solely from visual observations, and the proprioceptive module is not necessary to do so. Thus, it can be said that motor preparation is not the only way how action understanding could be explained in partial aspects. It has to be noted, however, that visual mirror neuron areas do not seem to directly communicate with the frontal areas (Rizzolatti and Luppino, 2001).

  The answer to the above question in my personal opinion is thus again *yes, but* the human mirror neuron system is probably not the exclusive cognitive mechanism for action understanding. Given that the activation of mirror neurons is contingent upon perspective-taking, this becomes obvious from the fact that in many cases, it is not necessary to fully adopt a momentary perspective of someone else to understand their position and feelings – abilities that have been mentioned in connection with the mirror neuron system – but rather to adopt their situational context. Whilst the term *action understanding* is often vaguely defined in the literature, also vague definitions of the mirror neuron system's role in it are immanent.

  Still, it is beyond dispute that humans are capable of some form of action understanding. Looking at the studies reviewed in this thesis, and the support by the model, motor simulation most likely offers one path to it, and it is most likely linked to imitation of already known actions. This thesis furthermore suggests that the involvement of perspective-taking may decide whether motor simulation takes part in action understanding, or whether context-sensitive inferential processes are involved (or both).

- What is possible from a computational perspective? What functionality can explain the observations and verify the made assumptions?

  The computational model developed in this thesis provides support for the addressed and unified hypotheses that center around action understanding and the mirror neuron system, as it substantiates their computational feasibility. The foundation for this feasibility was specifically found in spatial perceptual inferences by means of perspective-taking and feature binding.

Taken together, the combined hypothesis in this work contributes to answering the above questions, but certainly cannot answer them completely. Further investigations are needed to verify the model as well as its assumptions, some of which I suggest in the following.

7.2 Prospects and Opportunities

First of all, the introduced model apparently is no model of low-level visual processing, and thus basically abstracts most aspects of it: The model obtains its data in the form of preprocessed coordinates that represent salient visual cues. It does not approach the perception of raw visual, retinotopic, binocular sensory inputs and the extraction of information from them, e.g. of contrast and motion signal as observed in visual cortex (cf. Grossberg, 2007). Although the proposed population encodings in the model are somewhat comparable to retinotopic representations in the brain, in fact, they represent each visual feature independently, the neural plausibility of which can be questioned. Depth perception is assumed to be provided instead of being inferred from binocular visual streams. While the binding of features into meaningful Gestalt perceptions is basically solved by the model, the *Gestalt law of good continuation* (Jäkel et al., 2016) is not considered: The model does neither resolve occluded visual features, nor does it actively track the visual features over time. However, it might be feasible to transfer the proposed processing structures into a model of visual processing of image sequences. For example, convolutional neural networks provide a suitable basis for extracting higher-level features from high-dimensional topological inputs (e.g. LeCun et al., 1990). Visual features could be detected and tracked based on bottom-up and top-down salience cues, akin to the laws of Gestalt perception. Depth information can be detected from binocular vision (e.g. Hayashi et al., 2004), temporal filters can be added to extract motion information (e.g. Karpathy et al., 2014), and spatial transformations can be implemented to transform the retinotopic inputs by means of gated connectivity (e.g. Memisevic, 2013). Validating the suggested theoretical framework on low-level visual data could possibly reveal further insights about the involved bottom-up as well a top-down processes and their functional dependencies. Analogously, the format and processing of proprioceptive inputs is consistent with the visual pathway, but rather abstracted, although the application and processing of realistic sensory inputs would more problematic in this case.

In the experiments, it was shown that the model is able to internally visualize observations from different perspectives, and that it was able to imagine actions also when no sensory input was present. However, it lacks the ability to imagine actions from different perspectives, since the top-down, generative processes do not involve the complete visual processing cascade: The model can generate Gestalt expectations, but it cannot transform them internally via its perspective-taking mechanisms. Future models should thus consider to transform top-down generated expectations by means of Gestalt templates, instead of adapting bottom-up perception. This may also be reasonable from a computational perspective, given that the model processes massive amounts of realistic visual inputs, since the Gestalt expectations to be transformed should come only with a fraction of the complexity in comparison to the whole visual percept.

Along similar lines, although the perceptual inference methods introduced in this thesis are related to spatial abilities and visual attention, they lack a realistic mechanism for fixating and tracking *overall* observations, besides the individual features. Biological motion stimuli were adjusted for their locomotion on the ground, and the visual receptive field was just as large as the observed actors. Thus, the adaptive visual translation bias in the model can be considered the center of rotation for perspective-taking, or a mechanism for object centering in foveal vision, rather than an information driven eye fixation mechanism. Limiting the receptive field in this way was primarily done for reasons of computational complexity. Further investigations are needed that also include the processing of peripheral vision and object tracking.

Other opportunities concern the information propagation principles. After the preprocessing of abstracted visual and proprioceptive stimuli as described in Section 5.3, submodal spatial codes are learned with respect to their suitability for reconstructing the Gestalt stimuli momentarily, that is, without a temporal component. The spatial codes were not specifically optimized to foster temporal predictions of the codes within and across modalities. Analogously, the temporal code predictions were not particularly trained to optimize the Gestalt reconstructions. Rather, the error signal relied on the reconstruction of the respective codes. Although this separated learning approach worked well and is technically sound and robust, additional mechanisms may further improve the unsupervised learning of spatial expectations and temporal predictions. On the one hand, the temporal code predictions in the autoencoders could backpropagate the code prediction errors (via the logarithmic history, inside and across the modules) to the causative Gestalt codes in order to optimize them for better predictions, which could however worsen the quality of the spatial reconstructions. On the other hand, stimulus reconstructions could be obtained given the predicted instead of the observed codes to also optimize the code prediction for better spatial reconstructions, which could however, worsen the quality of the predicted codes. Thus, mixing both spatial and temporal predictions in this way might result in complex interferences. Finding suitable and robust parameterizations, learning and fusion procedures remains open for further investigations.

Also with respect to the distinction between spatial and temporal network components, the mixture of predicted and perceived codes was set manually during training and testing. The perception was considered completely reliable given that sensory information was provided. Allowing a more fine-grained, adaptive fusion of prediction and perception might also allow the development of more robust codes: The network might simultaneously filter noise in the perceptions given its already learned predictions. Mechanisms are available to avoid that the system falls into illusory loops (Kneissler et al., 2015). As well, adaptive information fusion might result in better recognition performance: Given that observed features are not sufficiently bound and the perspective is not sufficiently derived, then the expected submodal patterns – which currently stem from bottom-up activity solely – are typically distorted. Thus, allowing an influence of predicted codes will instantly provide a better expectation for the stimulus, such that the perceptual inference is sped up. However, it might indeed not be trivial to avoid self-delusions, particularly when a high number of different stimuli have been trained and are potentially available for interpreting a current stimulus. Considering that humans learn a variety of stimuli that do not solely refer to bodily actions, the perceptual bias should also be very broad and not strongly pre-determined by class-specific simulations.

Another possibility to improve the segmentation of stimuli into meaningful codes is to detect significant stimulus non-linearities, to construct a quasi-discretized representation from the continuous inputs. A similar approach was pursued in preliminary studies (e.g. Schrodt et al., 2015), where *motion patterns* of sufficiently linear parts of the data were formed, which predicted subsequent motion patterns. The event segmentation theory suggests that streams of perceptions are segmented into event codes in a similar manner (Zacks and Tversky, 2001; Zacks et al., 2007). Here, non-linearities in the Gestalt perceptions could be compared to event boundaries, while sufficiently linear parts between boundaries could be compared to events. By limiting the number of transition points between well-defined events, the robustness of the developing temporal attractors could be further improved. It remains open, however, how multiple, interacting autoencoders and event coding can be technically unified.

The presented model is able to consistently and robustly simulate different types of actions by temporal predictions that form stable attractors when the model is not driven by inputs. The model considers multiple time steps for its predictions, and is thus able to learn and unfold high-dimensional dynamics. Following the universal approximation theorem (Cybenko, 1989), it can approximate any continuous function on a compact subset of $\mathbb{R}^n$. Nonetheless, the model is deterministic, and thus cannot predict multimodal distributions of possible successor codes based on the same subset of recently activated codes and classified intentions. Thus, given that two actions that represent the same intention are *identical* for a sufficient amount of time, the network is unable to learn distinct predictive attractors for them. Accordingly, in case they are similar, the distinction of attractors is problematic. It might thus be beneficial to use probabilistic models (for example conditional variational autoencoders, Kingma and Welling, 2013) to represent uncertainties in the predicted codes.

In a similar manner, although the model can generate high-dimensional imaginations, it cannot learn very prolonged time series of data, in that it is limited to the length of its temporal horizon. For prolonged cyclic time series, that is, superimposed oscillations, it could be advantageous to integrate reservoirs of recurrently connected neurons into the model's prediction mechanisms, akin to Liquid State Machines (Maass, Natschläger, and Markram, 2002) or Echo State Networks (Jaeger, 2007), which are, however, typically not trained or trainable using backpropagation. For arbitrary, also non-cyclic time series, the usage of Long-Short-Term-Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), which are in theory able to preserve both activity and backpropagation signals internally for arbitrary time, could improve the prediction performance. LSTMs unfold their recurrent connections through time, but as well only up to a technically limited number of time steps. Thus, training algorithms for LSTMs truncate the backpropagated error signals at some point, and they are thus effectively similarly limited in the temporal horizon in which they can detect functional dependencies. Further studies are encouraged to clarify possible advantages of these methodologies and combinations thereof for simulations of biological motion.

Other prospects for advancements affect the evaluated data. The model was trained only on three short, cyclic actions, which served well as proof of the introduced concept and hypothesis. Even so, it can be criticized that the model was not able to classify the basketball action consistently. Note that it would have been a simple matter to achieve optimal results also in this case just by selecting a more extensive training set. However, this thesis did not aim at setting up a benchmark for the classification performance on the selected data set, but purposely focused on showing perceptual effects that result from the proposed neurocomputational principles, and for this task, the choice of data was productive. Nonetheless, in follow-up studies, the model should be trained and evaluated on a greater variety of actions, and possibly other appearances as well.

In preliminary studies, the model was able to learn twelve canonical views of actions: Three different movements each from four different vantage points (Schrodt et al., 2015). The network developed motion pattern neurons that *individually* encoded a snapshot-like representations of a submodal Gestalt input from the respective view-point. These patterns were activated by a winner-takes-all mechanism (cf. Grossberg, 1976), such that only one neuron was actively predicting its input at a time. In the current autoencoder approach, the Gestalt representations are distributed over the whole code layer, where neurons *jointly* predict Gestalt inputs, and are thus able to re-combine already learned parts of observations. Consequently, exploratory studies have shown that additional training stimuli result in a less than linear increase in training time. Thus, considering the advantages of distributed encodings of this character, the current model should potentially be able to encode a notably larger variety of actions, which remains to be verified and tested.

Other databases are available for evaluating further aspects of cognition typically attributed to mirror neurons. For example, the network could be trained to classify a subject's identity, or emotional properties of actions. The Emotional Body Motion Database[1] could be used, which provides motion captures of actions performed in different, intended emotional states, as well as information about the actor and classifications by human observers.

If it can be achieved to learn submodal encodings sufficiently abstracted from low-level sensory inputs, which could for example be promoted by hierarchical or deep autoencoders, the actions learned by the model could potentially cover the complete sensorimotor contingencies of a human body. However, the bigger the action manifold, the more important are components that select and distinguish subsets of it to produce particular trajectories that correspond to particular kinematic intentions, emotional states, or lead to particular goals states. Thus, the top-down biasing of modal simulations is an aspect that becomes increasingly important, and has to be increasingly elaborate for encoding more

[1] see http://ebmdb.tuebingen.mpg.de/index.php as of 08.01.2018

actions, and also action goals. Further investigations are needed to evaluate the suitability of the proposed top-down biasing, or find alternatives for it.

As a result of the cyclic nature of the considered actions, and the biasing by constant, kinematic intentions, the model formed predictive state attractors that were distributed evenly over several *submodalities and time*. In a similar manner, to model also final goal states of actions, mechanisms should be integrated that activate attractors selectively *distant in time*. Then, the submodal predictions would be directed along overall consistent state trajectories towards a final goal state attractor, akin to dynamic motion primitives (Schaal, 2006; Ijspeert et al., 2013).

Action understanding is eventually about social intelligence, and social intelligence is about interaction. In respect thereof, the model lacks several properties. First of all, the model is a purely perceptual modal, and it does not act. Ideally, the perceptual adaptations can be considered mental actions. To make the model more of a complete model of action understanding, it should learn the effects of force and motor-based dynamics in addition to proprioceptions and kinematics, also taking account of kinematic and physical constraints like gravity. As a first step towards a model of mental development of social intelligence, sensorimotor encodings should develop simultaneously by exploration, observation, and imitation as suggested before, which is however an enormous task for future studies. As it is now, the model gets idealized actions and the respective kinematic codes provided during training. As a second step, motivational aspects could be implemented as driving force of action and interaction, and in this context, considerations about when and if perspective-taking is initiated could be integrated.

Despite possible technical improvements and investigations, the model also opens up opportunities for behavioral and neuroimaging studies. Above all, studies should investigate the involvement of perspective-taking on the activation of mirror neurons. This could substantiate the claim of this thesis and generate new insights with further implications.

Closing Remarks

The only certainty about human cognition probably is that it is a peculiar and complex process that has evolved over millions of years. Not merely the several leftovers of evolution found in our genome, which do not follow a practical use anymore, indicate that also our brains do not apply only a single, homogeneous principle for all of their work. Moreover, the topic of action understanding and mirror neurons in particular is one of the most controversial today, as it is one of the most meaningful for human culture, leaving plenty of scope for interpretation and speculation. Although the endeavour to formulate universal, unifying principles of cognition is understandable and may indeed lead to a more general view on how the brain works, one should thus take them with a grain of salt. All of the theories and studies that provided the guideline for this thesis, just as the thesis itself and its framework, probably take their part in the truth, but none of them can be complete and undisputed, universal and unifying at the same time. All the more so, I personally hope that this thesis and its approach will contribute to the debates on mirror neurons and action understanding reasonably and constructively, the achievement of which I hereby leave to the interested readers to assess.

Bibliography

[1] Bessie Alivisatos and Michael Petrides (1997). "Functional activation of the human brain during mental rotation". In: *Neuropsychologia* 35.2, pp. 111–118.

[2] Adrian J.T. Alsmith and Matthew R. Longo (2014). "Where exactly am I? Self-location judgements distribute between head and torso". In: *Consciousness and cognition* 24, pp. 70–74.

[3] Richard A. Andersen (2011). "Inferior parietal lobule function in spatial perception and visuomotor integration". In: *Comprehensive Physiology*. John Wiley & Sons, Inc. ISBN: 9780470650714.

[4] Richard A. Andersen, Greg K. Essick, and Ralph M. Siegel (1985). "Encoding of spatial location by posterior parietal neurons". In: *Science* 230.4724, pp. 456–458.

[5] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele (2009). "Pictorial structures revisited: People detection and articulated pose estimation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1014–1021.

[6] — (2010). "Monocular 3D pose estimation and tracking by detection". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 623–630.

[7] Michael A. Arbib (2010). "Mirror system activity for action and language is embedded in the integration of dorsal and ventral pathways". In: *Brain and language* 112.1, pp. 12–24.

[8] William A. Ashton and Ann Fuehrer (1993). "Effects of gender and gender role identification of participant and type of social support resource on support seeking". In: *Sex Roles* 28.7, pp. 461–476.

[9] Alessio Avenanti, Angela Sirigu, and Salvatore M. Aglioti (2010). "Racial bias reduces empathic sensorimotor resonance with other-race pain". In: *Current Biology* 20.11, pp. 1018–1022.

[10] Chris Baker, Rebecca Saxe, and Joshua B. Tenenbaum (2006). "Bayesian models of human action understanding". In: *Proceedings of the Advances in Neural Information Processing Systems Conference*, pp. 99–106.

[11] Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum (2009). "Action understanding as inverse planning". In: *Cognition* 113.3, pp. 329–349.

[12] Lawrence W. Barsalou (1999). "Perceptual symbol systems". In: *Behavioral and Brain Sciences* 22.4, pp. 577–600.

[13] — (2008). "Grounded cognition". In: *Annual Review of Psychology* 59, pp. 617–645.

[14] T. Beardsworth and T. Buckner (1981). "The ability to recognize oneself from a video recording of one's movements without seeing one's body". In: *Bulletin of the Psychonomic Society* 18.1, pp. 19–22.

[15] Aude Billard and Maja J. Matarić (2001). "Learning human arm movements by imitation: Evaluation of a biologically inspired connectionist architecture". In: *Robotics and Autonomous Systems* 37.2, pp. 145–160.

[16] Stephen Blomfield (1974). "Arithmetical operations performed by nerve cells". In: *Brain research* 69.1, pp. 115–124.

[17] Marcel Brass and Cecilia Heyes (2005). "Imitation: Is cognitive neuroscience solving the correspondence problem?" In: *Trends in cognitive sciences* 9.10, pp. 489–495.

[18] Marcel Brass, Ruth M. Schmitt, Stephanie Spengler, and György Gergely (2007). "Investigating action understanding: Inferential processes versus action simulation". In: *Current Biology* 17.24, pp. 2117–2121.

[19] Cynthia Breazeal, Matt Berlin, Andrew Brooks, Jesse Gray, and Andrea L. Thomaz (2006). "Using perspective taking to learn from ambiguous demonstrations". In: *Robotics and Autonomous Systems* 54.5, pp. 385–393.

[20] Charles Bruce, Robert Desimone, and Charles G. Gross (1981). "Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque." In: *Journal of neurophysiology* 46.2, pp. 369–384.

[21] Giovanni Buccino, Stefan Vogt, Afra Ritzl, Gereon R. Fink, Karl Zilles, Hans-Joachim Freund, and Giacomo Rizzolatti (2004). "Neural circuits underlying imitation learning of hand actions: An event-related fMRI study". In: *Neuron* 42.2, pp. 323–334.

[22] Timothy J. Buschman and Earl K. Miller (2007). "Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices". In: *Science* 315.5820, pp. 1860–1862.

[23] Judith Bütepage, Hedvig Kjellström, and Danica Kragic (2017). "Anticipating many futures: Online human motion prediction and synthesis for human-robot collaboration". In: *arXiv preprint arXiv:1702.08212*.

[24] Stephen A. Butterfill and Corrado Sinigaglia (2014). "Intention and motor representation in purposive action". In: *Philosophy and Phenomenological Research* 88.1, pp. 119–145.

[25] Martin V. Butz (2016). "Toward a unified sub-symbolic computational theory of cognition". In: *Frontiers in Psychology* 7.925.

[26] Martin V. Butz and Esther F. Kutter (2016). *How the mind comes into being: An introduction to cognitive science from a functional and computational perspective*. Oxford University Press.

[27] Antonia F. de C. Hamilton and Scott T. Grafton (2007). "Action outcomes are represented in human inferior frontoparietal cortex". In: *Cerebral Cortex* 18.5, pp. 1160–1168.

[28] Manuel Cabido-Lopes and José Santos-Victor (2003). "Visual transformations in gesture imitation: What you see is what you do". In: *Proceedings of the IEEE International Conference on Robotics and Automation*. IEEE, pp. 2375–2381.

[29] Roberto Caminiti, Stefano Ferraina, and Paul B. Johnson (1996). "The sources of visual information to the primate frontal lobe: A novel role for the superior parietal lobule". In: *Cerebral Cortex* 6.3, pp. 319–328.

[30] Roberto Caminiti, Paul B. Johnson, Cesare Galli, Stefano Ferraina, and Yves Burnod (1991). "Making arm movements within different parts of space: The premotor and motor cortical representation of a coordinate system for reaching to visual targets". In: *Journal of Neuroscience* 11.5, pp. 1182–1197.

[31] Roberto Caminiti, Elena Borra, Federica Visco-Comandini, Alexandra Battaglia-Mayer, Bruno B. Averbeck, and Giuseppe Luppino (2017). "Computational architecture of the parieto-frontal network underlying cognitive-motor control in monkeys". In: *eNeuro* 4.1, ENEURO–0306.

[32] Malinda Carpenter, Katherine Nagell, Michael Tomasello, George Butterworth, and Chris Moore (1998). "Social cognition, joint attention, and

communicative competence from 9 to 15 months of age". In: *Monographs of the society for research in child development* 63.4, pp. 1–174.

[33] William Benjamin Carpenter (1852). "On the influence of suggestion in modifying and directing muscular movement, independently of volition". In: Royal Institution of Great Britain.

[34] John Bissell Carroll (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge: Cambridge University Press.

[35] Umberto Castiello, D. Lusher, M. Mari, Martin Edwards, and Glyn Humphreys (2002). "Observing a human or a robotic hand grasping an object: Differential motor priming effects". In: *Common Mechanisms in Perception and Action: Attention and Performance* XIX, pp. 315–333.

[36] Caroline Catmur, Vincent Walsh, and Cecilia Heyes (2007). "Sensorimotor learning configures the human mirror system". In: *Current Biology* 17.17, pp. 1527–1531.

[37] — (2009). "Associative sequence learning: The role of experience in the development of imitation and the mirror system". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1528, pp. 2369–2380.

[38] Thierry Chaminade, Andrew N. Meltzoff, and Jean Decety (2005). "An fMRI study of imitation: Action representation and body schema". In: *Neuropsychologia* 43.1, pp. 115–127.

[39] Trevor T.J. Chong, Ross Cunnington, Mark A. Williams, Nancy Kanwisher, and Jason B. Mattingley (2008). "fMRI adaptation reveals mirror neurons in human inferior parietal cortex". In: *Current biology* 18.20, pp. 1576–1580.

[40] Andy Clark (2013). "Whatever next? Predictive brains, situated agents, and the future of cognitive science". In: *Behavioral and Brain Sciences* 36.3, pp. 181–204.

[41] Yale E. Cohen and Richard A. Andersen (2002). "A common reference frame for movement plans in the posterior parietal cortex". In: *Nature Reviews. Neuroscience* 3.7, pp. 553–562.

[42] Richard Cook, Geoffrey Bird, Caroline Catmur, Clare Press, and Cecilia Heyes (2014). "Mirror neurons: From origin to function". In: *Behavioral and Brain Sciences* 37.02, pp. 177–192.

[43] James E. Cutting and Lynn T. Kozlowski (1977). "Recognizing friends by their walk: Gait perception without familiarity cues". In: *Bulletin of the psychonomic society* 9.5, pp. 353–356.

[44] George Cybenko (1989). "Approximation by superpositions of a sigmoidal function". In: *Mathematics of Control, Signals, and Systems (MCSS)* 2.4, pp. 303–314.

[45] Jean Decety and Andrew N. Meltzoff (2011). "Empathy, imitation, and the social brain". In: *Empathy: Philosophical and Psychological Perspectives*, pp. 58–81.

[46] Yiannis Demiris (2007). "Prediction of intent in robotics and multi-agent systems". In: *Cognitive Processing* 8.3, pp. 151–158.

[47] Giuseppe Di Pellegrino, Luciano Fadiga, Leonardo Fogassi, Vittorio Gallese, and Giacomo Rizzolatti (1992). "Understanding motor events: A neurophysiological study". In: *Experimental brain research* 91.1, pp. 176–180.

[48] Ilan Dinstein, Uri Hasson, Nava Rubin, and David J. Heeger (2007). "Brain areas selective for both observed and executed movements". In: *Journal of neurophysiology* 98.3, pp. 1415–1427.

[49] Ilan Dinstein, Cibu Thomas, Marlene Behrmann, and David J. Heeger (2008). "A mirror up to nature". In: *Current Biology* 18.1, pp. 13–18.

[50] Paul E. Downing, Yuhong Jiang, Miles Shuman, and Nancy Kanwisher (2001). "A cortical area selective for visual processing of the human body". In: *Science* 293.5539, pp. 2470–2473.

[51] Jean-René Duhamel, Carol L. Colby, and Michael E. Goldberg (1992). "The updating of the representation of visual space in parietal cortex by intended eye movements". In: *Science* 255.5040, pp. 90–92.

[52] — (1998). "Ventral intraparietal area of the macaque: Congruent visual and somatic response properties". In: *Journal of Neurophysiology* 79.1, pp. 126–136.

[53] John Carew Eccles (2013). *The physiology of synapses*. Academic Press.

[54] Martin G. Edwards, Glyn W. Humphreys, and Umberto Castiello (2003). "Motor facilitation following action observation: A behavioural study in prehensile action". In: *Brain and Cognition* 53.3, pp. 495–502.

[55] Stephan Ehrenfeld and Martin V. Butz (2014). "An embodied kinematic model for perspective-taking". In: *Cognitive Processing*. Vol. 15. 1. Springer, pp. 97–100.

[56] Stephan Ehrenfeld, Oliver Herbort, and Martin V. Butz (2013). "Modular neuron-based body estimation: Maintaining consistency over different limbs, modalities, and frames of reference". In: *Frontiers in Computational Neuroscience* 7.148.

[57] Stephan Ehrenfeld, Fabian Schrodt, and Martin V. Butz (2015). "Mario lives! An adaptive learning AI approach for generating a living and conversing Mario agent". In: *Video Proceedings of the 29th Conference of the Association for the Advancement of Artificial Intelligence (AAAI 2015)*.

[58] John Eliot and Ian Macfarlane Smith (1983). *An international directory of spatial tests*. Andover: Cengage Learning EMEA.

[59] Jeffrey L. Elman (1998). *Rethinking innateness: A connectionist perspective on development*. Vol. 10. MIT Press.

[60] Birgit Elsner (2007). "Infants' imitation of goal-directed actions: The role of movements and action effects". In: *Acta psychologica* 124.1, pp. 44–59.

[61] Andreas K. Engel, Alexander Maye, Martin Kurthen, and Peter König (2013). "Where's the action? The pragmatic turn in cognitive science". In: *Trends in Cognitive Sciences* 17.5, pp. 202 –209.

[62] Luciano Fadiga, Leonardo Fogassi, Giovanni Pavesi, and Giacomo Rizzolatti (1995). "Motor facilitation during action observation: A magnetic stimulation study". In: *Journal of Neurophysiology* 73.6, pp. 2608–2611.

[63] Scott E. Fahlman (1988). *An empirical study of learning speed in back-propagation networks*. Tech. rep. CMU-CS-88-262, Carnegie Mellon University.

[64] Brad M. Farrant, Tara A.J. Devine, Murray T. Maybery, and Janet Fletcher (2012). "Empathy, perspective taking and prosocial behaviour: The importance of parenting practices". In: *Infant and Child Development* 21.2, pp. 175–188.

[65] Pedro F. Felzenszwalb and Daniel P. Huttenlocher (2005). "Pictorial structures for object recognition". In: *International Journal of Computer Vision* 61.1, pp. 55–79.

[66] Falk Fleischer, Vittorio Caggiano, Peter Thier, and Martin A. Giese (2013). "Physiologically inspired model for the visual recognition of

transitive hand actions". In: *The Journal of Neuroscience* 33.15, pp. 6563–6580.

[67] Karl Friston, James Kilner, and Lee Harrison (2006). "A free energy principle for the brain". In: *Journal of Physiology-Paris* 100.1, pp. 70–87.

[68] Karl Friston, J. Mattout, and James Kilner (2011). "Action understanding and active inference". In: *Biological Cybernetics* 104.1-2, pp. 137–160.

[69] Tom Froese, Charles Lenay, and Takashi Ikegami (2012). "Imitation by social interaction? Analysis of a minimal agent-based model of the correspondence problem". In: *Frontiers in human neuroscience* 6.202.

[70] Adam D. Galinsky and Gordon B. Moskowitz (2000). "Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism". In: *Journal of Personality and Social Psychology* 78.4, pp. 708–724.

[71] Shaun Gallagher (2006). *How the body shapes the mind*. Clarendon Press.

[72] Vittorio Gallese (2001). "The 'shared manifold' hypothesis. From mirror neurons to empathy". In: *Journal of consciousness studies* 8.5-7, pp. 33–50.

[73] — (2007). "Before and below 'theory of mind': Embodied simulation and the neural correlates of social cognition". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 362.1480, pp. 659–669.

[74] Vittorio Gallese and Alvin Goldman (1998). "Mirror neurons and the simulation theory of mind-reading". In: *Trends in cognitive sciences* 2.12, pp. 493–501.

[75] Vittorio Gallese, Christian Keysers, and Giacomo Rizzolatti (2004). "A unifying view of the basis of social cognition". In: *Trends in cognitive sciences* 8.9, pp. 396–403.

[76] Vittorio Gallese, Luciano Fadiga, Leonardo Fogassi, and Giacomo Rizzolatti (1996). "Action recognition in the premotor cortex". In: *Brain* 119.2, pp. 593–609.

[77] Vittorio Gallese, Magali Rochat, Giuseppe Cossu, and Corrado Sinigaglia (2009). "Motor cognition and its role in the phylogeny and ontogeny of action understanding". In: *Developmental psychology* 45.1, pp. 103–113.

[78] Javier O. Garcia and Emily D. Grossman (2008). "Necessary but not sufficient: Motion perception is required for perceiving biological motion". In: *Vision research* 48.9, pp. 1144–1149.

[79] Ricardo Gattass and Charles G. Gross (1981). "Visual topography of striate projection zone (MT) in posterior superior temporal sulcus of the macaque". In: *Journal of Neurophysiology* 46.3, pp. 621–638.

[80] Apostolos P. Georgopoulos, John F. Kalaska, Roberto Caminiti, and Joe T. Massey (1982). "On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex". In: *Journal of Neuroscience* 2.11, pp. 1527–1537.

[81] Martin A. Giese (2013). "Biological and body motion perception". In: *The Oxford Handbook of Perceptual Organization*, pp. 575–596.

[82] Martin A. Giese and Tomaso Poggio (2003). "Neural mechanisms for the recognition of biological movements". In: *Nature Reviews Neuroscience* 4.3, pp. 179–192.

[83] Daniel Goleman and Friedrich Griese (1996). *Emotionale intelligenz*. Hanser München.

[84] Melvyn A. Goodale and A. David Milner (1992). "Separate visual pathways for perception and action". In: *Trends in Neurosciences* 15.1, pp. 20–25.

[85] Stephen Grossberg (1973). "Contour enhancement, short-term memory, and constancies in reverberating neural networks". In: *Studies in Applied Mathematics* 52.3, pp. 213–257.

[86] — (1976). "Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors". In: *Biological cybernetics* 23.3, pp. 121–134.

[87] — (2007). *Form perception*. Tech. rep. CAS/CNS-TR-07-020, Boston University Center for Adaptive Systems, Department of Cognitive, and Neural Systems.

[88] Emily Grossman, M. Donnelly, R. Price, D. Pickens, V. Morgan, G. Neighbor, and R. Blake (2000). "Brain areas involved in perception of biological motion". In: *Journal of cognitive neuroscience* 12.5, pp. 711–720.

[89] George W. Hartmann (1935). *Gestalt psychology: A survey of facts and principles*. Ronald Press Company.

[90] Ryusuke Hayashi, Taro Maeda, Shinsuke Shimojo, and Susumu Tachi (2004). "An integrative model of binocular vision: A stereo model utilizing interocularly unpaired points produces both depth and binocular rivalry". In: *Vision Research* 44.20, pp. 2367–2380.

[91] Mary Hegarty and David Waller (2004). "A dissociation between mental rotation and perspective-taking spatial abilities". In: *Intelligence* 32.2, pp. 175–191.

[92] Marc Heiser, Marco Iacoboni, Fumiko Maeda, Jake Marcus, and John C. Mazziotta (2003). "The essential role of Broca's area in imitation". In: *European Journal of Neuroscience* 17.5, pp. 1123–1128.

[93] Cecil G. Helman (2007). *Culture, health and illness*. CRC Press.

[94] Cecilia Heyes (2001). "Causes and consequences of imitation". In: *Trends in cognitive sciences* 5.6, pp. 253–261.

[95] — (2010). "Where do mirror neurons come from?" In: *Neuroscience & Biobehavioral Reviews* 34.4, pp. 575–583.

[96] — (2016). "Homo imitans? Seven reasons why imitation couldn't possibly be associative". In: *Biological Sciences* 371.1686.

[97] Cecilia M. Heyes and Elizabeth D. Ray (2000). "What is the significance of imitation in animals?" In: *Advances in the Study of Behavior* 29, pp. 215–245.

[98] Gregory Hickok (2009). "Eight problems for the mirror neuron theory of action understanding in monkeys and humans". In: *Journal of Cognitive Neuroscience* 21.7, pp. 1229–1243.

[99] Sepp Hochreiter and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural Computation* 9.8, pp. 1735–1780.

[100] Alexander C. Huk, Robert F. Dougherty, and David J. Heeger (2002). "Retinotopy and functional subdivision of human areas MT and MST". In: *Journal of Neuroscience* 22.16, pp. 7195–7205.

[101] Marco Iacoboni (2005). "Neural mechanisms of imitation". In: *Current opinion in neurobiology* 15.6, pp. 632–637.

[102] Marco Iacoboni and Mirella Dapretto (2006). "The mirror neuron system and the consequences of its dysfunction". In: *Nature Reviews Neuroscience* 7.12, pp. 942–951.

[103] Marco Iacoboni, Roger P. Woods, Marcel Brass, Harold Bekkering, John C. Mazziotta, and Giacomo Rizzolatti (1999). "Cortical mechanisms of human imitation". In: *Science* 286.5449, pp. 2526–2528.

[104] Auke Jan Ijspeert, Jun Nakanishi, Heiko Hoffmann, Peter Pastor, and Stefan Schaal (2013). “Dynamical movement primitives: Learning attractor models for motor behaviors”. In: *Neural Computation* 25.2, pp. 328–373.

[105] Philip L. Jackson, Andrew N. Meltzoff, and Jean Decety (2006). “Neural circuits involved in imitation and perspective-taking”. In: *Neuroimage* 31.1, pp. 429–439.

[106] Pierre Jacob and Marc Jeannerod (2005). “The motor theory of social cognition: A critique”. In: *Trends in Cognitive Sciences* 9.1, pp. 21–25.

[107] Herbert Jaeger (2007). “Echo state network”. In: *Scholarpedia* 2.9. Revision #183563, p. 2330.

[108] Frank Jäkel, Manish Singh, Felix A. Wichmann, and Michael H. Herzog (2016). “An overview of quantitative approaches in Gestalt perception”. In: *Vision Research* 126, pp. 3–8.

[109] Marc Jeannerod (1994). “The representing brain: Neural correlates of motor intention and imagery”. In: *Behavioral and Brain sciences* 17.2, pp. 187–202.

[110] Gunnar Johansson (1973). “Visual perception of biological motion and a model for its analysis”. In: *Perception & psychophysics* 14.2, pp. 201–211.

[111] Matthew Johnson and Yiannis Demiris (2005). “Perceptual perspective taking and action recognition”. In: *International Journal of Advanced Robotic Systems* 2.4, pp. 301–308.

[112] Ian T. Jolliffe (1986). “Principal component analysis and factor analysis”. In: *Principal Component Analysis*. Springer, pp. 115–128.

[113] Susan S. Jones (2009). “The development of imitation in infancy”. In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 364.1528, pp. 2325–2335.

[114] J.F. Kalaska, R. Caminiti, and A.P. Georgopoulos (1983). “Cortical mechanisms related to the direction of two-dimensional arm movements: Relations in parietal area 5 and comparison with motor cortex”. In: *Experimental Brain Research* 51.2, pp. 247–260.

[115] Kazunori Kamewari, Masaharu Kato, Takayuki Kanda, Hiroshi Ishiguro, and Kazuo Hiraki (2005). “Six-and-a-half-month-old children positively attribute goals to human action and to humanoid-robot motion”. In: *Cognitive Development* 20.2, pp. 303–320.

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

[117] Klaus Kessler and Lindsey Anne Thomson (2010). "The embodied nature of spatial perspective taking: Embodied transformation versus sensorimotor interference". In: *Cognition* 114.1, pp. 72–88.

[118] Christian Keysers (2011). *The empathic brain: How the discovery of mirror neurons changes our understanding of human nature*. Social Brain Press.

[119] James M. Kilner (2011). "More than one pathway to action understanding". In: *Trends in Cognitive Sciences* 15.8, pp. 352 –357.

[120] James M. Kilner and Chris D. Frith (2008). "Action observation: Inferring intentions without mirror neurons". In: *Current Biology* 18.1, pp. 32–33.

[121] Diederik P. Kingma and Max Welling (2013). "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114*.

[122] Mary D. Klinnert (1984). "The regulation of infant behavior by maternal facial expression". In: *Infant Behavior and Development* 7.4, pp. 447–465.

[123] Jan Kneissler, Jan Drugowitsch, Karl Friston, and Martin V. Butz (2015). "Simultaneous learning and filtering without delusions: A Bayes-optimal combination of Predictive Inference and Adaptive Filtering". In: *Frontiers in Computational Neuroscience* 9.47.

[124] Kurt Koffka (2013). *Principles of Gestalt psychology*. Vol. 44. Routledge.

[125] Stephane Lallee and Peter Ford Dominey (2013). "Multi-modal convergence maps: From body schema and self-representation to mental imagery". In: *Adaptive Behavior* 21.4, pp. 274–285.

[126] Claus Lamm, C. Daniel Batson, and Jean Decety (2007). "The neural substrate of human empathy: Effects of perspective-taking and cognitive appraisal". In: *Journal of Cognitive Neuroscience* 19.1, pp. 42–58.

[127] Xiangyang Lan and Daniel P. Huttenlocher (2005). "Beyond trees: Common-factor models for 2D human pose recovery". In: *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*. Vol. 1. IEEE, pp. 470–477.

[128] Joachim Lange, Karsten Georg, and Markus Lappe (2006). "Visual perception of biological motion by form: A template-matching analysis". In: *Journal of Vision* 6.8, pp. 836–849.

[129] Joachim Lange and Markus Lappe (2006). “A model of biological motion perception from configural form cues”. In: *The Journal of Neuroscience* 26.11, pp. 2894–2906.

[130] Georg Layher, Fabian Schrodt, Martin V. Butz, and Heiko Neumann (2014). “Adaptive learning in a compartmental model of visual cortex – how feedback enables stable category learning and refinement”. In: *Frontiers in Psychology* 5.1287.

[131] Yann LeCun (1989). “Generalization and network design strategies”. In: *Connectionism in Perspective*, pp. 143–155.

[132] Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel (1990). “Handwritten digit recognition with a back-propagation network”. In: *Proceedings of the Advances in Neural Information Processing Systems Conference*, pp. 396–404.

[133] Angelika Lingnau, Benno Gesierich, and Alfonso Caramazza (2009). “Asymmetric fMRI adaptation reveals no evidence for mirror neurons in humans”. In: *Proceedings of the National Academy of Sciences* 106.24, pp. 9925–9930.

[134] David F. Lohman (1979). *Spatial Ability: A Review and Reanalysis of the Correlational Literature.* Tech. rep. DTIC Document.

[135] David G. Lowe (1999). “Object recognition from local scale-invariant features”. In: *Proceedings of the International IEEE Conference on Computer Vision*. Vol. 2. Ieee, pp. 1150–1157.

[136] Wolfgang Maass (1997). “Networks of spiking neurons: The third generation of neural network models”. In: *Neural Networks* 10.9, pp. 1659–1671.

[137] Wolfgang Maass, Thomas Natschläger, and Henry Markram (2002). “Real-time computing without stable states: A new framework for neural computation based on perturbations”. In: *Neural Computation* 14.11, pp. 2531–2560.

[138] Bradford Z. Mahon and Alfonso Caramazza (2005). “The orchestration of the sensory-motor systems: Clues from neuropsychology”. In: *Cognitive Neuropsychology* 22.3, pp. 480–494.

[139] Abraham H. Maslow (1943). “A theory of human motivation”. In: *Psychological Review* 50.4, pp. 370–396.

[140] Michael McCloskey and Neal J. Cohen (1989). "Catastrophic interference in connectionist networks: The sequential learning problem". In: *Psychology of learning and motivation* 24, pp. 109–165.

[141] Mark G. McGee (1979). "Human spatial abilities: psychometric studies and environmental, genetic, hormonal, and neurological influences". In: *Psychological bulletin* 86.5, pp. 889–918.

[142] Andrew N. Meltzoff (2007). "'Like me': A foundation for social cognition". In: *Developmental science* 10.1, pp. 126–134.

[143] Andrew N. Meltzoff and M. Keith Moore (1977). "Imitation of facial and manual gestures by human neonates". In: *Science* 198.4312, pp. 75–78.

[144] — (1983). "Newborn infants imitate adult facial gestures". In: *Child development*, pp. 702–709.

[145] Roland Memisevic (2013). "Learning to relate images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8, pp. 1829–1846.

[146] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller (2013). "Playing Atari with deep reinforcement learning". In: *arXiv preprint arXiv:1312.5602*.

[147] Pascal Molenberghs, Ross Cunnington, and Jason B. Mattingley (2009). "Is the mirror neuron system involved in imitation? A short review and meta-analysis". In: *Neuroscience & Biobehavioral Reviews* 33.7, pp. 975–980.

[148] Pascal Molenberghs, Christopher Brander, Jason B. Mattingley, and Ross Cunnington (2010). "The role of the superior temporal sulcus and the mirror neuron system in imitation". In: *Human brain mapping* 31.9, pp. 1316–1326.

[149] David S. Moore and Scott P. Johnson (2008). "Mental rotation in human infants: A sex difference". In: *Psychological Science* 19.11, pp. 1063–1066.

[150] Michael C. Mozer (1989). "A focused back-propagation algorithm for temporal pattern recognition". In: *Complex Systems* 3.4, pp. 349–381.

[151] Kevin R. Murphy (2014). *A critique of emotional intelligence: What are the problems and how can they be fixed?* Psychology Press.

[152] Yukie Nagai, Yuji Kawai, and Minoru Asada (2011). "Emergence of mirror neuron system: Immature vision leads to self-other correspondence". In: *Proceedings of the 1st International Conference on Development and Learning*. Vol. 2. IEEE, pp. 1–6.

[153] Chrystopher L. Nehaniv and Kerstin Dautenhahn (2002). “The correspondence problem”. In: *Imitation in animals and artifacts* 41, pp. 41–61.

[154] Nora Newcombe (1989). “The development of spatial perspective taking”. In: *Advances in child development and behavior* 22, pp. 203–247.

[155] Aude Oliva, Antonio Torralba, Monica S. Castelhano, and John M. Henderson (2003). “Top-down control of visual attention in object detection”. In: *Proceedings of the International Conference on Image Processing*. Vol. 1. IEEE, pp. 253–256.

[156] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu (2016). “Wavenet: A generative model for raw audio”. In: *arXiv preprint arXiv:1609.03499*.

[157] M.W. Oram and D.I. Perrett (1994). “Responses of anterior superior temporal polysensory (STPa) neurons to ’biological motion’ stimuli”. In: *Cognitive Neuroscience, Journal of* 6.2, pp. 99–116.

[158] — (1996). “Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey”. In: *Journal of neurophysiology* 76.1, pp. 109–129.

[159] Marina Pavlova and Alexander Sokolov (2000). “Orientation specificity in biological motion perception”. In: *Perception & Psychophysics* 62.5, pp. 889–899.

[160] Marina A. Pavlova (2012). “Biological motion processing as a hallmark of social cognition”. In: *Cerebral Cortex* 22.5, pp. 981–995.

[161] Tabitha C. Peck, Sofia Seinfeld, Salvatore M. Aglioti, and Mel Slater (2013). “Putting yourself in the skin of a black avatar reduces implicit racial bias”. In: *Consciousness and Cognition* 22.3, pp. 779–787.

[162] David I. Perrett, Mark H. Harries, Ruth Bevan, S. Thomas, P.J. Benson, A.J. Mistlin, A.J. Chitty, J.K. Hietanen, and J.E. Ortega (1989). “Frameworks of analysis for the neural representation of animate objects and actions”. In: *Journal of Experimental Biology* 146.1, pp. 87–113.

[163] D.I. Perrett, P.A.J. Smith, A.J. Mistlin, A.J. Chitty, A.S. Head, D.D. Potter, R. Broennimann, A.D. Milner, and M.A. Jeeves (1985). “Visual analysis of body movements by neurons in the temporal cortex of the macaque monkey: A preliminary report”. In: *Behavioural brain research* 16.2, pp. 153–170.

[164] D.I. Perrett, M.W. Oram, M.H. Harries, R. Bevan, J.K. Hietanen, P.J. Benson, and S. Thomas (1991). "Viewer-centred and object-centred coding of heads in the macaque temporal cortex". In: *Experimental Brain Research* 86.1, pp. 159–173.

[165] Jean Piaget (2013). *Child's conception of space: Selected works*. Vol. 4. Routledge.

[166] Alexandre Pouget, Peter Dayan, and Richard Zemel (2000). "Information processing with population codes". In: *Nature Reviews Neuroscience* 1.2, pp. 125–132.

[167] Wolfgang Prinz (1984). "Modes of linkage between perception and action". In: *Cognition and Motor Processes*, pp. 185–93.

[168] Marisa Przyrembel, Jonathan Smallwood, Michael Pauen, and Tania Singer (2012). "Illuminating the dark matter of social neuroscience: Considering the problem of social interaction from philosophical, psychological, and neuroscientific perspectives". In: *Frontiers in Human Neuroscience* 6.190.

[169] Aina Puce and David Perrett (2003). "Electrophysiology and brain imaging of biological motion". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 358.1431, pp. 435–445.

[170] Varun Ramakrishna, Daniel Munoz, Martial Hebert, James Andrew Bagnell, and Yaser Sheikh (2014). "Pose machines: Articulated pose estimation via inference machines". In: *Proceedings of the European Conference on Computer Vision*. Springer, pp. 33–47.

[171] Elizabeth Ray and Cecilia Heyes (2011). "Imitation in infancy: The wealth of the stimulus". In: *Developmental science* 14.1, pp. 92–105.

[172] Michael Rescorla (2017). "The computational theory of mind". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2017. Metaphysics Research Lab, Stanford University.

[173] Giacomo Rizzolatti and Michael A. Arbib (1998). "Language within our grasp". In: *Trends in neurosciences* 21.5, pp. 188–194.

[174] Giacomo Rizzolatti and Laila Craighero (2004). "The mirror-neuron system". In: *Annual Review of Neuroscience* 27, pp. 169–192.

[175] — (2005). "Mirror neuron: A neurological approach to empathy". In: *Neurobiology of human values*. Heidelberg: Springer, pp. 107–123.

[176] Giacomo Rizzolatti, Leonardo Fogassi, and Vittorio Gallese (2001). “Neurophysiological mechanisms underlying the understanding and imitation of action”. In: *Nature Reviews Neuroscience* 2.9, pp. 661–670.

[177] Giacomo Rizzolatti and Giuseppe Luppino (2001). “The cortical motor system”. In: *Neuron* 31.6, pp. 889–901.

[178] Giacomo Rizzolatti and Corrado Sinigaglia (2007). “Mirror neurons and motor intentionality”. In: *Functional neurology* 22.4, pp. 205–210.

[179] Giacomo Rizzolatti, Luciano Fadiga, Vittorio Gallese, and Leonardo Fogassi (1996). “Premotor cortex and the recognition of motor actions”. In: *Cognitive brain research* 3.2, pp. 131–141.

[180] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams (1986). “Learning representations by back-propagating errors”. In: *Nature* 323.6088, pp. 533–536.

[181] Sverker Runeson and Gunilla Frykholm (1983). “Kinematic specification of dynamics as an informational basis for person-and-action perception: Expectation, gender recognition, and deceptive intention.” In: *Journal of Experimental Psychology: General* 112.4, pp. 585–615.

[182] Joni N. Saby, Peter J. Marshall, and Andrew N. Meltzoff (2012). “Neural correlates of being imitated: An EEG study in preverbal infants”. In: *Social neuroscience* 7.6, pp. 650–661.

[183] Emilio Salinas and L.F. Abbott (2001). “Coordinate transformations in the visual system: How to generate gain fields and what to compute with them”. In: *Progress in brain research* 130, pp. 175–190.

[184] Stefan Schaal (2006). “Dynamic movement primitives – a framework for motor control in humans and humanoid robotics”. In: *Adaptive Motion of Animals and Machines*. Springer, pp. 261–280.

[185] Matthias Scholz and Ricardo Vigário (2002). “Nonlinear PCA: A new hierarchical approach”. In: *Proceedings of the European Symposium on Artificial Neural Networks*, pp. 439–444.

[186] Fabian Schrodt and Martin V. Butz (2014). “Modeling perspective-taking by forecasting 3D biological motion sequences”. In: *Cognitive Processing, Suppl. KogWis 2014*. Vol. 15, pp. 137–139.

[187] — (2015). “Learning conditional mappings between population-coded modalities”. In: *Machine Learning Reports* 3.2015, pp. 141–148.

[188] — (2016). "Just imagine! Learning to emulate and infer actions with a stochastic generative architecture". In: *Frontiers in Robotics and AI* 3.5.

[189] Fabian Schrodt, Johannes Lohmann, and Martin V. Butz (2016). "Mario becomes social!" In: *Video Proceedings of the 30th Conference of the Association for the Advancement of Artificial Intelligence*.

[190] Fabian Schrodt, Yves Röhm, and Martin V. Butz (2017). "An event-schematic, cooperative, cognitive architecture plays Super Mario". In: *Cognitive Robot Architectures* 1855, pp. 10–16.

[191] Fabian Schrodt, Georg Layher, Heiko Neumann, and Martin V. Butz (2014a). "Modeling perspective-taking by correlating visual and proprioceptive dynamics". In: *Proceedings of th 36th Annual Conference of the Cognitive Science Society*, pp. 1383–1388.

[192] — (2014b). "Modeling perspective-taking upon observation of 3D biological motion". In: *Proceedings of the 4th International Conference on Development and Learning and on Epigenetic Robotics*, pp. 328–333.

[193] — (2015). "Embodied learning of a generative neural model for biological motion perception and inference". In: *Frontiers in Computational Neuroscience* 9.79.

[194] Fabian Schrodt, Jan Kneissler, Stephan Ehrenfeld, and Martin V. Butz (2017). "Mario becomes cognitive". In: *Topics in Cognitive Science* 9.2, pp. 343–373.

[195] Stephen H. Scott (2003). "The role of primary motor cortex in goal-directed movements: Insights from neurophysiological studies on non-human primates". In: *Current Opinion in Neurobiology* 13.6, pp. 671–677.

[196] Stephen H. Scott, Paul L. Gribble, Kirsten M Graham, and D. William Cabel (2001). "Dissociation between hand motion and population vectors from neural activity in motor cortex". In: *Nature* 413.6852, pp. 161–165.

[197] Roger N. Shepard and Jacqueline Metzler (1971). "Mental rotation of three-dimensional objects". In: *Science* 171.3972, pp. 701–703.

[198] Shenna Shepard and Douglas Metzler (1988). "Mental rotation: Effects of dimensionality of objects and type of task". In: *Journal of Experimental Psychology: Human Perception and Performance* 14.1, pp. 3–11.

[199] Leonid Sigal and Michael J. Black (2006). "Measure locally, reason globally: Occlusion-sensitive articulated pose estimation". In: *Proceedings of*

the IEEE Conference on Computer Vision and Pattern Recognition. Vol. 2. IEEE, pp. 2041–2048.

[200] Niilo Sirola (2010). "Closed-form algorithms in mobile positioning: Myths and misconceptions". In: *Proceedings of the Workshop on Positioning, Navigation and Communication*. Ieee, pp. 38–44.

[201] Jessica A. Sommerville and Jean Decety (2006). "Weaving the fabric of social interaction: Articulating developmental psychology and cognitive neuroscience in the domain of motor cognition". In: *Psychonomic Bulletin & Review* 13.2, pp. 179–200.

[202] Jennifer A. Stevens, Pierre Fonlupt, Maggie Shiffrar, and Jean Decety (2000). "New aspects of motion perception: Selective neural encoding of apparent human movements". In: *Neuroreport* 11.1, pp. 109–115.

[203] Armin Stock and Claudia Stock (2004). "A short history of ideo-motor action". In: *Psychological Research* 68.2-3, pp. 176–188.

[204] Graham W. Taylor, Geoffrey E. Hinton, and Sam T. Roweis (2006). "Modeling human motion using binary latent variables". In: *Advances in Neural Information Processing Systems 19*. Ed. by Bernhard Schölkopf, John C Platt, and Thomas Hofman. MIT Press, pp. 1345–1352.

[205] Steven M. Thurman and Emily D. Grossman (2008). "Temporal 'bubbles' reveal key features for point-light biological motion perception". In: *Journal of Vision* 8.3, pp. 1–11.

[206] Michael Tomasello (1999). "The human adaptation for culture". In: *Annual review of anthropology* 28, pp. 509–529.

[207] Ivan Toni, Daniel Thoenissen, and Karl Zilles (2001). "Movement preparation and motor intention". In: *Neuroimage* 14.1, pp. 110–117.

[208] J. Gregory Trafton, Nicholas L. Cassimatis, Magdalena D. Bugajska, Derek P. Brock, Farilee E. Mintz, and Alan C. Schultz (2005). "Enabling effective human-robot interaction using perspective-taking in robots". In: *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 35.4, pp. 460–470.

[209] Anne Treisman (1998). "Feature binding, attention and object perception". In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 353.1373, pp. 1295–1306.

[210] Luca Turella, Moritz F. Wurm, Raffaele Tucciarelli, and Angelika Lingnau (2013). “Expertise in action observation: Recent neuroimaging findings and future perspectives”. In: *Frontiers in Human Neuroscience* 7.637.

[211] Erlinda R. Ulloa and Jaime A. Pineda (2007). “Recognition of point-light biological motion: Mu rhythms and mirror neuron activity”. In: *Behavioural brain research* 183.2, pp. 188–194.

[212] Kai Vogeley and Gereon R. Fink (2003). “Neural correlates of the first-person-perspective”. In: *Trends in Cognitive Sciences* 7.1, pp. 38–42.

[213] Kai Vogeley, Mark May, Afra Ritzl, Peter Falkai, Karl Zilles, and Gereon R. Fink (2004). “Neural correlates of first-person perspective as one constituent of human self-consciousness”. In: *Journal of Cognitive Neuroscience* 16.5, pp. 817–827.

[214] Stephen C. Want and Paul L. Harris (2002). “How do children ape? Applying concepts from the study of non-human primates to the developmental study of ’imitation’ in children”. In: *Developmental Science* 5.1, pp. 1–14.

[215] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh (2016). “Convolutional pose machines”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4724–4732.

[216] Paul J. Werbos (1988). “Generalization of backpropagation with application to a recurrent gas market model”. In: *Neural Networks* 1.4, pp. 339–356.

[217] Ronald J. Williams and David Zipser (1989). “A learning algorithm for continually running fully recurrent neural networks”. In: *Neural Computation* 1.2, pp. 270–280.

[218] Roy A. Wise (2004). “Dopamine, learning and motivation”. In: *Nature Reviews Neuroscience* 5.6, pp. 483–494.

[219] Yi Yang and Deva Ramanan (2011). “Articulated pose estimation with flexible mixtures-of-parts”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1385–1392.

[220] Jeffrey M. Zacks and Barbara Tversky (2001). “Event structure in perception and conception”. In: *Psychological Bulletin* 127.1, pp. 3–21.

[221] Jeffrey M. Zacks, Nicole K. Speer, Khena M. Swallow, Todd S. Braver, and Jeremy R. Reynolds (2007). “Event perception: A mind-brain perspective”. In: *Psychological Bulletin* 133.2, pp. 273–293.