Gestalt Perception of Biological motion with a Generative Artificial Neural Network Model

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von

Mahdi Sadeghi
aus Teheran, Iran

Tübingen
2021

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

Tag der mündlichen Qualifikation:	1.4.2022
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Martin Butz
2. Berichterstatter:	Prof. Dr. Siegfried Wahl

Abstract

In cognitive modelling understanding of biological motion by inference of own sensorimotor skills is extremely valued and is known as a fundamental element of social intelligence. It has been suggested that a proper Gestalt perception depends on suitably binding visual features, decently adapting the matching perspective, and mapping the bound features onto the correct Gestalt templates. This thesis introduces a generative artificial neural network model, which implements such Gestalt perception mechanisms proposing an algorithmic explanation. The architectural design of the model is an extension, modification and further investigation of previous work by Fabian Schrodt [102] which relies on the principle of active inference and predictive coding, coupled with suitable inductive learning and processing biases. At first we train the model to learn sufficiently accurate generative models of dynamic biological, or other harmonic, motion patterns. Afterwards we scramble the input and vary the perspective onto it. To be able to properly route the input and adapt the internal perspective onto a known frame of reference, the suggested modularized architecture propagates the prediction error back onto a binding matrix which consists of hidden neural states that determine feature binding, and further back onto perspective taking neurons, which rotate and translate the input features. The resulting process ensures that various types of biological motion are inferred upon observation, resolving the challenges of (I) feature binding into Gestalten, (II) perspective taking, and (III) behavior interpretation. Ablation studies underline that, 1. the separation of spatial input encodings into relative positional, directional, and motion magnitude pathways boost the quality of Gestalt perception, 2. population encodings implicitly enable the parallel testing of alternative interpretation hypotheses and therefore further improve accurate inference, 3. a temporal predictive processing module of the autoencoder-based compressed stimuli enables the retrospective inference of the unfolding behavior. I believe that similar components should be employed in other architectures where temporal bindings of information sources are beneficial. Moreover, given that binding, perspective taking, and intention interpretation are universal problems in cognitive science, our introduced mechanisms may be very useful for addressing similar challenges in other domains beyond biological motion patterns.

Kurzfassung

Das Verständnis biologischer Bewegungen durch die Rückschlüsse auf die eigenen sensomotorische Fähigkeiten gilt als ein grundlegendes Element der sozialen Intelligenz und ihm wird in der kognitiven Modellierung ein hoher Stellenwert beigemessen. Es wurde vorgeschlagen, dass eine korrekte Gestaltwahrnehmung von einer adäquaten Bindung visueller Merkmale, einer sinnvolle Übernahme der passenden Perspektive und der Zuordnung der gebundenen Merkmale zu den richtigen Gestaltschablonen abhängt. In dieser Arbeit wird ein generatives künstliches neuronales Netzwerkmodell vorgestellt, das solche Mechanismen der Gestaltwahrnehmung implementiert und eine algorithmische Erklärung hierfür vorschlägt. Das architektonische Design des Modells ist eine Erweiterung, Modifikation und Vertiefung der vorausgegangenen Arbeit von Fabian Schrodt [102], die auf dem auf dem Prinzip der aktiven Inferenz und der prädiktiven Kodierung basiert, gekoppelt mit geeigneten induktiven Lern- und Verarbeitungsverzerrungen. Zunächst trainieren wir das Modell darauf, ausreichend genaue generative Modelle von dynamischen biologischen oder anderen harmonischen Bewegungsmustern zu lernen. Anschließend wird die Eingabe umgeordnet und ihre Perspektive variiert. Um den Input richtig zu lokalisieren und die interne Perspektive an ein bekanntes Bezugssystem anpassen zu können, propagiert die vorgeschlagene modularisierte Architektur den Vorhersagefehler auf eine Bindungsmatrix zurück, die aus versteckten neuronalen Zuständen besteht, die die Merkmalsbindung bestimmen. Anschließend wird der Fehler weiter zurück auf perspektivische Neuronen propagiert, die die Eingangsmerkmale rotieren und räumlich verschieben. Dieser Prozess ermöglicht, dass verschiedene Arten von biologischen Bewegungen bei der Beobachtung erschlossen werden können, wobei die Aufgaben der (I) Merkmalsbindung in Gestalten, (II) Perspektivenübernahme und (III) Verhaltensinterpretation gelöst werden. Ablationsstudien unterstreichen, dass 1. die Trennung der räumlichen Eingangscodierungen in relative Positions-, Richtungs- und Bewegungsgrößenpfade die Qualität der Gestaltwahrnehmung erhöht, 2. Populationskodierungen implizit das parallele Testen alternativer Interpretationshypothesen ermöglichen und somit die Genauigkeit der Schlussfolgerungen verbessern, 3. ein zeitlich prädiktives Verarbeitungsmodul der durch einen Autoencoder komprimierten Stimuli ermöglicht die retro-

spektive Inferenz des beobachteten Verhaltens. Ich glaube, dass ähnliche Komponenten auch in anderen Architekturen eingesetzt werden sollten, in denen zeitliche Verknüpfungen von Informationsquellen von Vorteil sind. Da Bindung, Perspektivenübernahme und Absichtsinterpretation universelle Probleme in der Kognitionswissenschaft sind, können die von uns eingeführten Mechanismen zudem sehr nützlich sein, um ähnlichen Herausforderungen in anderen Bereichen als biologischen Bewegungsmustern anzugehen.

Acknowledgments

I would first like to express my deepest appreciation to my principal supervisor Prof. Dr. Martin V. Butz who directed my PhD studies. His guidance and mentorship along this journey was invaluable to me. My inspiration and passion to overcome challenges was, for the largest part, sparked and fueled by his previous cognitive modeling accomplishments. Being affiliated with him is a true honor indeed.

I would also like to express my sincere thanks to Prof. Siegfried Wahl for his support and valuable feedback during the project and taking the time to review this thesis. Without his participation and input, this research could not have been successfully conducted.

I am deeply grateful to Dr. Sebastian Otte and Dr. Fabian Schrodt who provided support, constructive feedback and nourishment throughout the development of this project.

Credits to all neuro-cognitive modeling lab members for their constant support and to help me enjoy the period of my doctoral studies.

Finally, I must express my very profound gratitude to my parents Mahnaz an Javad and to my sister Roya for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been achievable without them. Thank you.

Contents

List of Figures

List of Tables

Abbreviations

AE Autoencoder 30, 31, 74

AMC Acclaim Motion Capture 34

BCE Binary Cross-Entropy 50

CNN Convolutional Neural Network 22, 27, 76

FBE Feature Binding Error 41, 43, 56, 63

IFG Inferior Frontal Gyrus 9, 10

IPL Inferior Parietal Lobule 10

LSTM Long short-term memory 18, 23, 31, 34, 44, 47, 51, 60, 68, 75

MNS Mirror Neuron System 10

N2pc an Event-related potential component linked to selective attention 13

OD Orientation Difference 43, 64

PCA Principal Component Analysis 30, 74

PMC Premotor Cortex 10

RNN's Recurrent Neural Networks 25, 27

ST Spatial Transformers 22, 76

STS Superior Temporal Sulcus 9, 10

TD Translation Difference 43, 64

ToM Theory of Mind 6

VAE Variational Autoencoder 18, 30, 31, 36, 39, 44, 47, 49, 51, 52, 54, 60, 68, 73, 74, 77, 78

VR Virtual Reality 80

Chapter 1

Introduction

For many years, the development of humans has been examined within separate domains by cognitive psychologists. For instance, social cognition, which is always typified by theory of mind (ToM) tasks, has been explored separately from visual or motor skills as well as Gestalt Perception, Binding, and Perspective-Taking. Nevertheless, recent studies have shown that there might be links in the human brain and cognitive systems fundamental to these distinct skill sets. In this current study, the claim that social and motor skills are developing together is examined by exploring key cognitive domains including motor skill, imitation, and action understanding, all of which contribute, in distinct ways, to both social and motor behavior. There are suggestions in adult and neuroimaging research asserting a close link between imitation and action understanding, despite them being somehow independent of ToM and low-level motor control. Inspired by underlying mechanisms of action understanding in the brain, we suggest a generative recurrent neural network model, which resolves perspective taking, binding, and behavior interpretation problems concurrently by means of gradient-based inference.

1.1 Biological Background

Mirror neurons refer to the cells that trigger not only observation but also the execution of actions. While they were initially found in macaque monkey's premotor area F5, research has determined that they are also present in other brain areas such as the dorsal premotor cortex, and the lesser parietal lobule (Rizzolatti et al. [92]). Concerning human beings, similar cognitive neuroscience evidence regarding mirror neurons' presence has led to certain researchers' claims that mirror neurons play a role in action understanding (Johnson and Demiris [58]). Nevertheless, the actual meaning of action understanding is something that has been and continues to be disputed. Consequently, there will be a constant substantial confusion in research concerning both what action understanding as a process means and whether or not mirror neurons have a part to play in that particular process. Action understanding, however, has been, according to various scholars such as Gallese, and Fadiga, et al. [37], defined as an individual's ability to comprehend other people's actions, infer the objective of their action, and determine the intentions that motivate their actions. According to this definition, action understanding can be separated into three unique processes: identifying the cluster of body parts that play a part in a given action, its developmental progression, and the label of the individual's intention (Pellegrino et al. [25]). These processes, however, are distinct regarding the level of generalization that is called for across action facets. In particular, to identify an action, certain affective features need to be discriminated against based on the configural linkage between diverse parts of the body. The identification of objectives and intentions, on the other hand, requires simplification across the affective facets of the observed actions. According to Cook et al. [22], this is because an objective such as ”to grasp”, or an intention such as ”in order to eat,” can be reached through various grip types; significantly, a similar type of grip can be utilized in achieving numerous other intentions and goals. No one-to-one binding exists between the configurations of body parts, preferences, and goals; similar mirror neuronal triggering patterns cannot simultaneously represent the other's actions, objectives, and intentions. While this has been recognized in the cognitive neuroscience literature, scholars such as Arbib [8]; Rizzolatti et al. [93]; Gallese, and Fadiga, et al. [37], examining neurocognitive mechanisms fundamental to action understanding, often pay attention to just one of the three processes previously mentioned; problematically, however, all the three processes are considered as action understanding.

The non-specific utilization of the concept of “action understanding” has resulted in, as previously opined, contradicting results about whether or not mirror neurons (or brain areas in human that are considered to contain mirror neurons) play a role in action understanding. According to Johnson and Demiris [58], some researchers have claimed that the areas of the brain that contain mirror neurons are accountable for the identification of the intention of others. In his study, Johnson and Demiris found that interference by constant theta-burst stimulation of the premotor cortex impaired the precision on a task that involved an individual's intention. This task involved participants combining information about a hand configuration with contextual information to infer the intentions of the actor. Nevertheless, closer scrutiny of the findings indicated that participants matched a hand configuration illustrated in a video to that shown in an image, which was compromised to the exact degree as the intended task. As Cook et al. [22] assert, if mirror neurons play some part in identifying intentions, the stimulation should have lessened precision on the task that included the analysis of the intentions to a significant degree than the task that was purely perceptual. Thus, the findings only offer considerable evidence for mirror neurons' involvement in identifying the affective facets of actions.

1.1.1 Visual information Processing

Studies on brain imaging have shown that action observation in human beings activates the inferior frontal gyrus (IFG), the rostral area of the IPL, the lower zone of the precentral gyrus, and the parietal and occipital visual regions. The parietal and the frontal mirror neuron areas are somatotopically arranged (Bruce et al. [14]). The activation of the IFG's pars opercularis reveals the observation of the distal mouth and hand actions, while premotor cortex activation reveals proximal neck and arm movement. Unlike in monkeys, mirror neurons in human beings trigger even when intransitive (meaningless) trends are being observed (Oram and Perrett [81]). The observation of necessary actions leads to the triggering of the temporal and frontal nodes of the MNS; on the other hand, the word of pointless activities leads to the triggering of simply the frontal node. The primary function of the mirror neuron system, as earlier discussed, is action understanding. Every time an individual sees another individual perform a specific action, the mirror neurons representing the said action's performance are triggered. While performing an imitation, the observed information is transferred from the eyes to the visual cortex, which is then passed around the mirror neuron system (MNS) to the motor output to the muscles. Figure 1.1 provides a schematic view of the corresponding mirror neuron system (MNS) in the human brain.

Mirror neurons play a significant role in transforming visual observations into knowledge (Goodale and Milner [42]). Researchers on action observation have indicated that the inferior prefrontal lobe (IPL), the IFG, and an area within STS are triggered. Although action observation does not activate the primary motor cortex (PMC), it is somehow directly involved in action understanding (Pavlova [86]). Visual hypothesis, the "generate and test" framework, and the direct-match theory are key theories that explain the spectrum of action understanding. The visual theory is grounded on the effector's visual analysis, the object, and on the situation in which the action will infer a conclusion on the meaning of the action. The inferotemporal lobe, STS areas, and visual extrastriate regions make up the neural substrates (Downing et al. [26]).

On the other hand, the direct-match theory relies on observed mapping actions on one's motor depiction of the observed activity. Thus, it comprises a process of observation-triggered motor depiction, followed by matching this to the motor depiction generated during simulation. If there is correspondence between both motor depictions, action understanding results. A rather complicated theory posited for action understanding is the "generate-test" framework; according to this theory, action understanding needs to identify an "imaginary" objective that would create an achievement blueprint in the motor planning system of the observer (Ulloa and Pineda [117]). This objective is then matched with the actual action that has been observed. It is acknowledged that if the simulated motor action and the practical action do not match, another hypothesis is then generated and later assessed for unity with the effort that was observed. Therefore, actions are not only understood in terms of their end results but also in terms of mental states, particularly concerning the objectives that created them (Grossman et al. [45]).

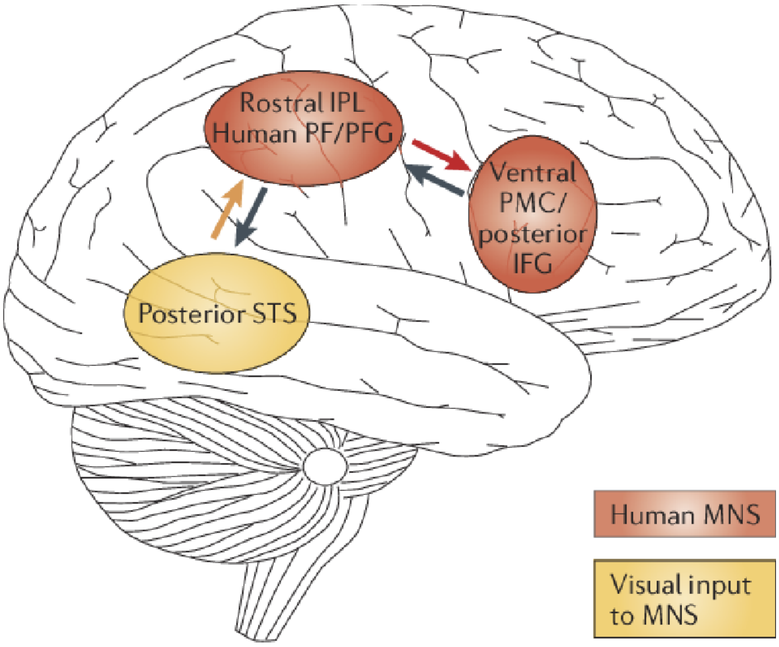


Figure 1.1: *A sketch of the frontoparietal mirror neuron system (MNS) (red) in the human brain and its major visual input (yellow) triggered by a gaze cue. Areas with mirror neuron properties are located at inferior frontal cortex(consist of posterior IFG and adjacent ventral PMC) as well as in the posterior area (e.g. rostral part of IPL). Temporal regions (e.g. STS) receive visual information and provide recognized input for the MNS. This is sent to parietal mirror neurons which are associated to imitation, sensorimotor integration, and spatial cognition, and relay it to the lower frontal cortex where the movement becomes correlated with a target. Figure from [54].*

1.2 Cognitive Background

Recognizing and making sense of other people's actions is called "action understanding" which is a challenging task due to the complexity of human behavior and movements. This task is performed by integrating visual and other sensory dynamics into the own sensorimotor system (i.e. simulation theory of social cognition [9, 38]).

1.2.1 Imitation, Social cognition, and Emotional Intelligence

Social cognition can be defined as a group of processes that range from perception to decision-making. It is fundamental to an individual's ability to decode other people's behaviors and intentions to organize actions aligning with moral and social, aside from economic and personal considerations. Its criticality in day-to-day life reveals the neural intricacy of social processing and the pervasiveness of social cognition deficits in various pathological situations (Carpenter et al. [20]). Social cognitive processes can be classified into three main areas linked with social perception - the affective handling of social cues such as emotional expression and faces, social understanding – the comprehension of other people's affective or cognitive states, and social decision-making – planning of behaviors considering not just one's objectives, but also those of others (Gallese et al. [39]).

Imitation, also known as cognitive imitation, refers to a type of social learning in which one copies another person's behavior. In biology, the importance of imitation is on its adaptive value regarding an organism's survival; in psychology, however, the significance is in development and growth of one's skills in order to produce more targeted actions (Elsner [29], Meltzoff[72]). The most significant imitation cases involve behavior that is demonstrated, which the imitator cannot see while performing the said behavior; a good example is a person scratching his or her head. Imitations of this kind are often considered opaque imitations since they are not easy to account for without theorizing about cognitive mechanisms, including perspective-taking, which most animals do not have. In this present study, the types of imitations of concern are social learning and social influence, which are not considered opaque imitations (Uithol and Paulus [116]). For example, mechanisms that are species-typical such as contagion and mimicry; mechanisms that are motivational such as fear transfer, incentive motivation, and social facilitation; attentional mechanisms that include stimulus and local enhancement; and affordance learning mechanisms that entail observational conditioning, imprinting, and understanding how the environment works (Przyrembel et al. [91]).

Nevertheless, imitation gives rise to a different problem that concerns how the individual imitating determines the pattern of motor activation that will mimic the model's action (Murphy [74]). While this challenge is unique, specialist theories have asserted that a corresponding unique solution exists, which claims the neurological and functional mechanisms meant for controlling imitation. Those general theories have suggested that such a problem can be resolved using action control and associative learning tools. Accordingly, this is in agreement with recent cognitive neuroscience research findings inspired by mirror neurons' discovery (Gallese and Goldman [38]). As such, imitation seems to be based on the mechanical stimulation of motor representations by movement observation; the externally prompted motor representations are then utilized to replicate the observed behavior. This capacity to imitate hinges on learned perceptual motor skills (Brass and Heyes [13]). This is how social cognition connects to action understanding. On the other hand, emotional intelligence specifies an individual's ability to utilize, comprehend, and manage his or her emotion in positive ways that support management of stress, effective communication, sharing in other people's pain, as well defusing conflicts and overcoming challenges (Goleman and Griese [41]). All these – imitation, social cognition, and emotional intelligence - are significant in cognitive neuroscience and psychology because they facilitate social and human mental health development.

1.2.2 Gestalt Perception

The visual system heavily relies on the organization's regularities of perceptual elements to generate transitive depictions of the world. An important illustration of such a function that has been formalized in the principles of Gestalt psychology is the affective grouping of simple visual cues such as arcs and lines into unitary objects, such as shapes and form (Jäkel et al. [59]). Recent studies on neuroimaging have determined posterior areas in the parietal and temporal lobes as being neuro-functional associates of Gestalt perception. In addition to prominent connection on a neuronal degree, the mechanisms are both extraordinarily the same on a behavioral degree depicting both a particular type of top-down visual dispensation in which individual objects are combined into a superordinate unit (Gallese, Keysers, et al. [39]). By observing a movement from a given perspective our brain employs its top-down embodied expectations in order to detect the respective features and group them to a proper Gestalt percept.. As argued by Herrmann and Bosch [50], the voxels triggered in global Gestalt perception must strongly react to configurations within than external to the range of subitizing. In characterizing automatic attentional capture and the associated cognitive processing of visual stimuli that are Gestalt like (Hartmann [46]), especially at the psychophysiological level using event-associated potential, it is apparent that Gestalt perception stimuli, when compared to non-Gestalt ones, are typified by a substantial N2pc together with improved event-associated potential amplitudes of non-literalized components (Wagemans et al. [120]).

1.2.3 Binding (correspondence) Problem

The determination of any criteria for imitation, social learning, mimicry, or even copying assumes a binding notion between separate and independent agents. They are adjudging whether or not a behavior has been socially transmitted calls for the observer to find a correspondence between the imitator and is being imitated (Butz and Kutter [19]). If the imitator and the one being imitated have similar bodies – for instance, they belong to the same species, are of the same gender, or age – then it would be evident to a human observer to simply map the body parts that correspond such as the right leg of the one demonstrating to the right portion of the imitator, and so forth (Nehaniv and Dautenhahn [75]). Also, an apparent correspondence of actions exists. However, if there are no obvious corresponding points between the demonstrator and the imitator, a correspondence problem arises.

The problem of binding (correspondence) is concerned with selecting and associating different visual facets into the correct combination (Heyes [51]; Butz and Kutter [19]). In nature, the bodies of many organisms change and grow over time. Still, those who socially learn can maintain and adapt those capabilities that have been socially transmitted regardless of the growth above and changes, whether natural or injurious, to their embodiment (Koffka [64]). Artificial intelligence agents such as robots and others that socially learn can face the correspondence problem and benefit from such massiveness to embodiment alterations. The correspondence problem is connected to attention (Treisman [114]), which is renowned to be determined by top-down signals from task requirements together with bottom-up signals derived from salient stimuli (Buschman and Miller [15]). For recognizing major scene information and their relations, the visual system leads attention to particular salience cues and binds them.

To solve the correspondence problem, other studies, for instance, Nehaniv and Dautenhahn [75] and Treisman [114], have demonstrated that loose perceptual synchronization and matching with the demonstrator each led to foster learning and significantly lower error rates. Recent artificial social learning mechanisms such as [5], can be utilized in population agents or robots to achieve cultural transfer in populations of such nature, even heterogeneous ones that comprise individuals with different embodiments.

1.2.4 Perspective Taking Challenge

Perspective-taking refers to a complicated cognitive process that has a vital role in social cognition, particularly regarding understanding other people's perspectives in a given situation (Schrodt et al. [104]). Clearly, there is some level of a conceptual connection between the cognitive and visual manifestations since they both involve identifying differing points of view. In this regard, Kessler and Thomson, [60] considered spatial perspective-taking as an embodied transformation in which the observer mentally adopts (i.e. rotates and translates) own body scheme onto an observed person. Studies have indicated that perspective-taking is a mental skill that develops early on in life; however, the challenge with it is that, with experience and age, it becomes more flexible and more complicated (Piaget and Inhelder [88]; Meltzoff and Prinz [71]). This means, although one can clearly recognize or comprehend another person's thinking in a given situation, they can either manipulate what they perceived or be manipulated (Johnson and Demiris [58]). Additionally, it also raises abiding by correspondence challenges about selecting and integrating perceived features in the right combination.

1.3 Technical Background

The way we efficiently perceive other people is in accordance with our observation and interpretation after years of evolution. As a result, making machines understand human behavior is a hallmark of cognition and artificial intelligence.

1.3.1 Learning models for action understanding

Deep learning and machine learning approaches have experimentally shown great success regarding learning image representation for specific tasks such as image captioning, action understating, semantic segmentation, and object identification (Ji et al. [57]; Nweke et al.[77]). Through convolutional neural networks, it is possible to capture the premise of the spatial locality of data structure concerning images via parameter sharing convolutions and through local invariance-constructing max-pooling neurons (Kleinlein et al. [63]). Understanding human action is a puzzling time series clustering task that comprises the prediction of a person's movement based on sensor data (He et al. [47]; Fu et al. [36]). Conventionally, it includes deep domain capability and techniques from signal processing to accurately bring about features from basic information in order to match the machine learning model.

Deep learning techniques such as recurrent and convolutional neural networks have recently shown that they can achieve great results by automatically identifying features from raw sensor information. Convolutional neural network models are a form of deep learning model designed for use with image data (Nweke et al. [77]; Oniga and Suto [80]; Guyen and Mirza [76]). They are highly effective in addressing interesting computer vision challenges, particularly when they are taught for activities such as localizing and identifying elements in images. Without supervision, they describe images' contents (Guyen and Mirza [76]; Layher et al. [65]; Tu et al. [115]). As Ji et al. [57] demonstrates in their consideration of a completely automated action recognition in an environment without supervision, CNNs, as deep learning models, are capable of directly acting on raw data inputs, thereby automating the entire feature reconstruction process. They however argue that models such as CNN are ill-equipped to handle inputs such as 2D. They therefore propose a CNN model that is 3D-based, which is capable of action recognition. Their proposed model has the capability of extracting features not only from spatial dimensions, but also from temporal dimensions through the performance of 3D convolutions; this enables the capturing of motion data that is encoded in various adjacent frames. This is possible because the ensuing model generates numerous information channels from frames of input, with the eventual feature representation being obtained by putting together information from all available channels. When applied to the real-world environment to identify human actions, its accomplishments are far superior in terms of performance, particularly given the fact that it does not rely on any handcrafted facets. On the other hand, recurrent neural networks are a form of deep learning model designed for learning from sequential data. The extended short-term memory network (LSTM) has remained considered the most effective form of recurrent neural network model because they have shown to be effective on sequence prediction problems that are challenging (Fu et al. [36]; Ma et al. [68]).

Layher et al.'s [65] focus on action recognition with regards to human agents on the basis of biologically motivated visual architecture of examining expressed motion, extends coarsely separated streams of sensory processing along diverse pathways that distinctively handle motion and form data. They show that the approach they propose aligns with other key pose suggestions in the literature that chose key pose frames based on the technique explained in past studies. Furthermore, in comparisons to the training on the complete frames set, depictions that are taught on the key pose frames lead to greater confidence with regards to cluster assignments. Layher et al. [65] also assert that key pose depictions illustrate auspicious capabilities for generalization in an evaluation of cross-dataset. However, none of the studies discussed above are able to tackle fundamental problems of action understanding (i.e. Feature Binding, Perspective-Taking, Gestalt Perception) simultaneous to inferring the intention of the occurring action.

1.4 Goal-directed Action, Kinematic Intention

Basic foundations of homeostatic behavior comprise the interaction with positive items while avoiding those considered negatives. For the longest time, kinematic bounds of any actions have been hypothesized to mirror the motor plan's content (Wise [122]). Kinematic perpetual of actions can be caught from various people carrying out a similar undertaking such as, for example, lifting up a cup from coffee table. Certainly, the kinematic facets of the upper-limb in the event of actions that are objective-driven have been shown to echo the intentions of the participant (indicating at or grasping something; fitting or lifting something; taking to throw) (Rizzolatti and Luppino [94]; Butterfill and Sinigaglia [16]). The same is true for the object's fundamental attributes (weight, shape, geometry, size, and texture) that the agent intermingles with. After all, motoric components and social intentions translate into particular kinematic facets.

Kinematic features associated with goal-focused actions not only mirror motor planning but are also controlled by intentions. The kinematics of activities directed towards stimuli considered pleasant could indicate facilitation (Hamilton and Grafton [23]). Consequently, the result of such a nature would be in tandem with the notion that motor plans include the outlays and the rewards of a particular action (Jacob and Jeannerod [55]). In general, behavior is goal-directed; thus, as the organization facets of objective-driven actions can be concluded from their movement structures, examining the kinematics of relating to emotion-loaded stimuli is essential. Esteves et al. [32] measured the kinematics of the movement of study participants' wrists in the task in which they were required to grasp stimuli that were emotion-loaded and carry them towards their bodies, found that there was an increase in the time-to-peak speed concerning fetching and moving pleasant stimuli towards one's body and a decrease when it came to bringing neutral or unpleasant stimuli. This suggested that resulting behavior aims to create situations linked with positive rewards, which in turn indicate that there is easiness in carrying out tasks with pleasant stimuli. Apparently, in reach-for-and-take and bring-towards-body motions, valence stimuli tend to impact the temporal and not the movement's spatial kinematic aspects.

1.5 Summary

The primary role of mirror neurons is action understanding and as discussed above, Binding, Perspective-Taking, Gestalt Perception, and behavior interpretation abilities are key aspects for having a competent social behavior.

Visual cues from others come in a different perspective than ones own body. Perspective taking is interpreted as a non-discrete mental transformation aiming at projecting our own perspective into the perspective of an observed person

[58, 105, 107] through priming corresponding motor simulations [28]. The binding problem addresses a major brain ability in which individual features such as motion direction, pattern, color or texture are integrated into one coherent entity, that is, a Gestalt [59]. Subsequently, to selectively bind observed features together, bottom-up saliency cues interact with top-down expectations in a Bayesian manner [15, 59].

We propose a generative recurrent neural network model, which solves the aforementioned problems simultaneously using gradient-based inference. Moreover, when the input stimulus is just partially present, the model is still able to provide biased imaginations and to perform Gestalt perception. The model contains multiple novel components such as a variational autoencoder (VAE for learning compressed spatial codes) along with a recurrent neural network module (a long short-term memory network, i.e. LSTM for learning temporally compressed codes, and to interpret intention).

Chapter 2

Related Mechanisms, Comparisons, Motivations

Our perception of an observed action is built upon distinct but firmly entangled cognitive domains that stem from our own embodied experiences. These cognitive domains represent motor, visual and intention-based encodings. In the vertebrates' nervous system, the encoding of sensory stimuli often happens via the combined effort of large population of neurons. These activity patterns have typically been considered as encoding the stimulus' value, while computation has been solemnized based on purpose estimation. Suggestions have in recent times indicated that neural computation is analogous to a procedure referred to as Bayesian inference, with population activity arrays depicting improbability concerning stimuli in the form of probability distributions. Pouget, P. Dayan, et al. [89] asserts that population codes are becoming valued as representational devices because there exists a globally accepted basic decoding and encoding framework whose properties are well understood. Nevertheless, very many areas of active examination persist. A good example is how continuous attractor networks are in an ideal world appropriate for implementing necessary computations with population codes, including approximations of basis function, removal of noise, and the integration of statistically sound cues (Pouget, P. Dayan, and R. Zemel [90]).

Emphasis has also been on using population codes for more apparent stimulus representation aspects, such as computational diversity and uncertainty. These proposals, however, are rather computational than mechanistic, except for the log-likelihood framework that depicts how filtering motion-energy offers an apt substrate for computations of statistical nature (Denève and Pouget [24]). Nevertheless, such models can further be developed and refined as propelled by the inevitable barrage of psychophysical findings demonstrating the sophisticated way through which observers attain, learn, and handle uncertainty acts.

2.1 Population Coding

Population coding refers to the quantitative study of which representation or algorithms are used by the human brain to associate together and assess the messages that different neurons carry (Pouget et al. [89]). The key issue in cognitive neuroscience concerning population coding concerns how information that a neural population carries can be quantified; it also concerns how to quantify the contribution by each member of the neural population or their interaction with each other regarding the general information that the considered neuron groups encode (Olshausen and Field [79]). There is a belief that neurons encode an animal's location concerning a global-centered frame of reference in environments, such as small mazes. An example of such an encoding is shown in Figure 2.1.

An essential element regarding population coding mechanism is because it is robust to the extent that damage to one cell cannot result in a catastrophic influence on the encoded representation (Pouget, P. Dayan, and R. Zemel [90]). This is because the information is often encoded across numerous cells.

2.2 Neural Dynamic Fields

Dynamic field hypothesis is an established theory for modeling embodied cognition. According to this theory, essential cognitive functions such as memory formation, attentional processes, formation of grounded depictions, adaptation, learning, and decision making all stem from neural dynamics (Schöner [101]; Erlhagen and Bicho [30]; Erlhagen and Schöner [31]). The elementary computation component of this model is the dynamic neural field. Under constraints on the dynamics time-scale, the dynamic neural field is computationally equal to a soft winner-take-all network that is believed to be one of the fundamental computational entities in neural processing.

Humans and animals are pretty remarkable regarding their ability to develop behavior in changing and complex environments (Martin, Scholz, et al. [70]). Their neural systems can resolve complex problems of movement creation and perception in the real world with adaptability, flexibility, and massiveness that are far beyond the technical capabilities of any system available today. The problem of how biological neural networks can continuously cope with the dynamics and complexities of real-world situations and still achieve their behavioral objectives is not easy to resolve. Processes including the formation of memory, adaptation, attention, and learning are all crucial in the biological problem solving of behavior generation in real-world settings. Neural dynamic fields are essential in the comprehension of how these process are realized by the biological brains' neural networks, and it is at the center of understanding cognition in humans and in the development of cognitive artifact that effectively contend with

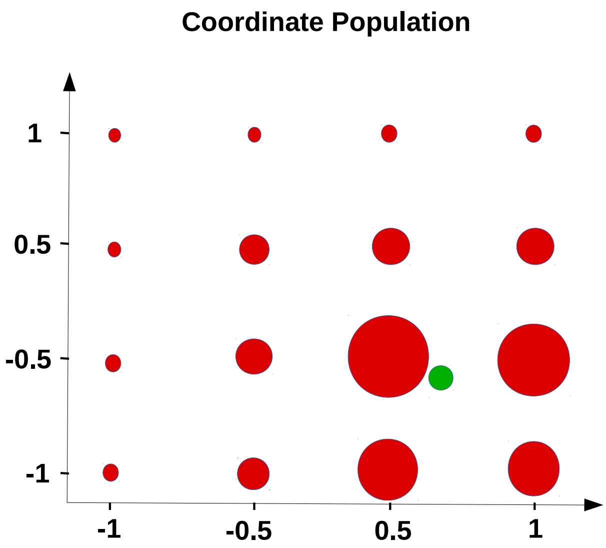


Figure 2.1: *Encoded population on 2D pendulum; Encoding result of 16 multivariate Gaussian neurons (red) which are uniformly distributed in an area all around the input stimulus (green) which is the coordinate of one joint of 2D pendulum at a given frame. Closer neurons to the observed stimulus have higher activations.*

the constraints in the real world (Sabinasz et al. [97]). Accordingly, dynamic neural fields have been suggested as a simplified mathematical framework for neural processing hinged on persistent interactions. These frameworks neglect individual neurons' temporal dynamics but rather utilize the averaged firing rate to transfer information and can create temporal bindings between different conceptual representations, for example, in the spatial arrangement of objects. The dynamic neural field model proposed by Amari [6] is exciting since it enables a fully analytical description of pattern formation dynamics under certain simplifying presumptions (Martin et al. [69]).

2.3 Transformer networks

While the convolutional neural networks (CNN) earlier discussed define a remarkably dominant cluster of models, they are still limited. Notably, this is because they cannot be spatially invariant to the input data in a productive way. This apparent limitation of convolutional neural networks results from them only having a somewhat inadequate, predefined strategy for pooling that is expected to handle variation in data's spatial arrangement pattern. Transformer networks, particularly spatial transformers, are known for enabling the network's input data to be spatially manipulated and subsequently executing data-dependent affine transformation on them [56].

As a result, they can be incorporated into the existing convolutional architecture, essentially enabling neural networks to actively alter feature maps conditional on the map itself, often without any additional modification of the process of optimization or any further training supervision.

Basic neural network architectures can be empowered with spatial transformation ability by making use of spatial transformers (ST). The utilization of spatial transformer networks, according to Jaderberg et al. [56], leads to the development of models that learn invariance to scale, translation, more general warping, and rotation, leading to an unrivaled performance on various benchmarks and for several clusters of transformations. However, it is essential to mention that the performance of spatial transformer networks depends on individual data samples, with the appropriate behavior acquired during training for the said task.

2.4 Attentional Mechanisms

The way we integrate and direct attention for an adequate perception, is affected by our expectations and embodied experiences. In transduction problems, for instance, machine translation and language modeling, and in sequence modeling, recurrent neural networks, gated recurrent, and LSTM neural networks have mainly been firmly developed as state-of-the-art methodologies. There has been continued effort to expand the boundaries of decoder-encoder architectures and recurrent language models. Typically, recurrent models include computation along with the position of symbols of output and input sequences (i.e. sequential computation) and subsequently do not perform well in terms of computational efficiency and learning duration (Vaswani et al. [118]).

Apparently, this fundamentally sequential nature prohibits parallelization within training instances that becomes essential at the longer sequence lengths where memory constraints restrict batching across cases. By utilizing conditional computation and factorization tactics, recent work has achieved significant performance enhancements. Nevertheless, the major constraint regarding sequential analysis persists. The attentional mechanism proposed by Vaswani et al. [118] is a framework architecture that eschews recurrence and instead relies entirely on an attention mechanism to infer global dependencies between output and input. Bring into line the positions to steps in computation results in a series of hidden states, which are a function of the initial state, and the input for position (Vaswani et al. [118]).

It allows for substantially more parallelization and can result in a better translation quality after very little training. The potential of attention-based frameworks is massive, particularly, if they can be extended to problems that involve output and input modalities apart from text to examine restricted, local attention mechanisms to effectively handle outputs and inputs such as audio, video, and images. It has been hypothesized that with this model, convolutions and recurrence can entirely be dispensed with (Olah, C. and Carter, [78]).

2.5 Capsule Networks

The vision of primates, particularly humans, often disregards details that are considered irrelevant by utilizing a cautiously determined order of fascination points to make sure that only a tiny part of the optic selection is eventually dealt with at the highest of resolutions. In trying to comprehend how much of our understanding concerning scenes emanates from the sequence of fascinations and how much they are due to a single fascination, using introspection is a very poor strategy. This is because a single fascination offers much more than a single observed object and its associated properties. As Sabour et al. [98] assert, the human multi-layered visual system always results in a parse tree-like configuration on each fascination. Parse trees are created on the fly by the dynamic allocation of memory. Sabour et al., however, assert that for one fixation, a parse tree is made from a fixed multi-layer neural network, the same way that a sculpture is carved out of a rock. Each layer is subdivided into various tiny clusters of neurons referred to as capsules, with each node in the parse tree corresponding to a capsule that is active. A capsule refers to a cluster of neurons whose trajectory of activity embodies the instantiation bounds of a certain type of unit, for instance, part of an object or the entire object itself. An active capsule individually chooses a capsule in a level beyond it to be its originator in the tree configuration through the repetitious routing method. This iterative process will address the challenge of allocating fragments to aggregates for advanced achromatic system levels.

The neurons' actions in any dynamic capsule embody a given unit's numerous features in the image. The elements can comprise various diverse forms of instantiation stricture, including posture (size, position, and alignment), speed, hue, deformation, albedo, amongst others (Sabour et al., [98]). One significant property is the presence of the instantiated unit in the image. A generic way of representing presence is using a separate logistic entity whose output relies on the chance that the unit is present. Sabour et al., however, propose another exciting alternative: the utilization of the entire length of the instantiation parameters' vector to depict the presence of the unit and force the vector's orientation to depict the unit's properties.

Since a capsule's output is a vector, it is conceivable to use a dominant active routing mechanism to guarantee that the capsule's output is directed to a suitable originator within the advanced layer. In the beginning, the result is routed to all potential originators but is cut back using coupling coefficients. This routing based on the agreement is far more efficient, especially when compared with the primary type of routing implemented by max-pooling, which permits neurons from one layer to disregard all except the most dynamic facet detector in a local group in the less advanced layer. This dynamic routing mechanism based on capsule networks is an efficient approach for segmenting highly overlapping objects.

Capsules utilize neural activities that differ with variation in viewpoints instead of eliminating the element of interpretation from activities. As a result, they are advantageous when compared to normalization techniques such as spatial transformer networks. They can handle multiple varying affine transformations of different object parts or whole objects simultaneously. Furthermore, they are also very efficient in handling segmentations, which is another difficult problem for vision, since the instantiation vector parameters enable them to utilize routing-by-agreement (Sabour et al. [98]). Additionally, regarding the significance of the dynamic routing process, there is backing from biologically plausible frameworks of the visual cortex's invariant pattern recognition. Therefore, there are vital representation reasons supporting capsule networks and the routing-by-agreement mechanism as a better methodology. However, further incremental insights are necessary before it can work exceptionally well in highly developed technologies.

2.6 Retrospective Inference models

As opposed to feed forward models that have limitations working with sequential data types, Recurrent neural networks (RNN's) [96] can process sequences of structure-like data with any length. Butz et al. [18] suggest a dynamic adaptive inference process based on a recurrent artificial neural network, which learns time-based analytical frameworks of dynamic systems. This process is referred as a Retrospective and Prospective Inference Scheme (REPRISE) process, which infers the contextual hidden state that best explains its new sensorimotor encounters and as associated, context-reliant temporal predictive frameworks, in retrospect. REPRISE combines learning of hidden state with goal-directed, active-inference-based control in which the activities of the parametric bias neurons are adapted retrospectively. For this purpose, the input, predictions, and hidden states are stored for a declared number of previous time steps. This period of past time steps is referred as tuning horizon. In order to update parameters of interest at each time step, a prediction error is computed and projected into the past horizon by means of back-propagation through time [121]. Using these updated parameters, the model then rolls out forward chain based on adapted hidden state, resulting in a corrected prediction for time step t. REPRISE provides a neural implementation, first of its kind, that demonstrates the emergent propensity to group different types of predictably coded sensorimotor dynamics into solid event codes. Although learning, the algorithm deduces established hidden states that tend to oppose discrete sensorimotor dynamics. REPRISE, Butz et al. note, is capable of differentiating, the three different types of vehicles, and effectively control them in a goal-focused way without any information regarding the identity of the cars or that three other vehicles are being offered. This is a remarkable feat that might be suitable for learning event-focused constructs and event hierarchies; however, there is still a need for additional work aimed at scaling the system for it to be applicable in more challenging scenarios.

A similar adaptation approach where gated autoencoder (or restricted Boltzmann machine) architectures are learned was employed by Memisevic [73]. According to this study, in many vision tasks such as stereopsis, invariant recognition, motion understanding, and visual odometry, one essential operation is establishing correspondences between images and between data from other modalities and pictures. Various recent studies have attempted to examine how to infer correspondences from data through Spatio-temporal, relational, and bilinear variants of deep learning techniques. These techniques utilize multiplicative interactions between features or pixels to represent patterns of correlation across numerous images. Memisevic, in his review of such work offers an examination of the part played by multiplicative interactions in learning associated with the encoding of relations and propose how complex cell and square-pooling models can be considered as a way of not only representing multiplicative interactions, but also of encoding relations (Memisevic [73]).

While various past studies have recommended that compositional theories be sensorimotor-based, achieving it is still somewhat unclear. Sugita et al. [109], however, suggests a second-order neural network that uses sensorimotor time-series data to learn compositional structures and tunes parametric bias neurons in a similar retrospective inference manner as discussed above.

Accordingly, Sugita et al. [109] demonstrate that various setups of second-order neural networks that have parametric biases can learn to compositionally emulate the interactions of objects beyond the interactions that had initially and particularly trained. This was impossible in previous architectures of neural networks such as recurrent neural networks (RNN's) and convolutional neural networks (CNN). These imitation capabilities, Sugita et al. asserts, are achieved through the development of self-structured, geometrically ordered compositional theory structure in parametric biases, and in task-based Braitenberg-like [12] sensory programming in hidden sensory layers. Second-order connections are crucial in the accomplishments of such tasks, connections of such nature are crucial in not only the learning in sensorimotor-based compositional structures, but also in the learning of Braitenberg-like, behavior-focused pro-presentations (Tani [110]; Battaglia et al. [11]).

Further, because other neural networks such as recurrent neural networks with parametric biases have not been able to show similar behavioral generalizations or compositional encodings in the same manner that second-order multiple time-scale recurrent neural networks have, it seems like multiplicative, second-order relations are necessary for learning compositional structure from patterns in sensorimotor interactions (Tani et al. [111]). However, as posited by Sugita et al., there are several ways in which this current approach should be refined, including ensuring that information on distinctness being relayed from internal reward and motivation signals, and the extension of the model such that it recursively acquires structured interaction concepts. Further, for the system's scalability to also be fostered to more complex and diverse interactions, the learning architecture needs to be further modularized and more obvious mechanisms of attention, goal-directedness, and focus need to be introduced.

2.7 Summary and Comparisons

Population coding supports having a consistent encoding over a neural population and is considered as a fundamental information processing property in our nervous system. Moreover, it provides a proper encoding even if some neurons are damaged. Our processing approach is inspired in a similar manner where we encode the input stimulus by a population of topological neurons with Gaussian tunings, which yield local responses to particular stimuli within a limited scope.

Neural Dynamic Fields in which different dynamics strive for persistent interactions, can create temporal bindings between various conceptual representations that are closely related to our possessing strategy.

Transformer network that models affine transformations works similar to our perspective-taking strategy. Although, they infer the reference frame purely stimulus driven and by means of a feed-forward procedure.

Attentional mechanisms that selectively process information can be considered related to our feature-binding adaptation procedure. However, our binding strategy routes information retrospectively rather than in a feed-forward approach.

Capsule Networks in which capsules (i.e. cluster of neurons) embody a given unit's features in the image, implements a dynamic routing mechanism so as to segment overlapping objects. This method is closely related to our flexible binding strategy.

Feed forward models cannot properly deal with sequential data, especially if the input sequences vary in length. Note that for our model we employ sequential motion capture data as input in which not only the information in one single sequence matter, but, moreover, the order in which individual elements occur within that sequence is vital for an adequate information transmission (Goodfellow et al. [43]). As in our approach neural codes are constantly attempting to match sensory inputs with top-down expectations and we require to associate observed visual information with the own sensorimotor system, to be able to solve feature-binding, spatial perspective-taking, and behavior interpretation problems, we employ similar adaptive retrospective inference processes like the ones represented in section 2.6. A summary of closely related approaches is given in Table 2.1.

Table 2.1: *Related Mechanisms*

	Feed-Forward	Retrospective	Flexible-Binding	Perspective-taking
Transformer Net.	✓	✗	✓	✓
Attention M.	✓	✗	✓	✗
Capsule Net.	✓	✗	✓	✗
Our approach	✓	✓	✓	✓

As opposed to other related approaches discussed in Table 2.1, our proposed architecture applies gradient-based (retrospective) inference in order to tune parametric bias neurons, which will be used to establish feature bindings and adapt the internal perspective onto the observed features.

2.8 Motivations

We introduce a generative neural network model, which employs its own generated imaginations to solve the Gestalt perception, perspective taking, feature binding, and behavior interpretation problems. The model is based on the dissertation of Fabian Schrodt [102, 103]. Although, it consists of multiple novel modules as well as a recurrent artificial neural network module with a different strategy for recognizing occurring behavior, and a variational autoencoder (VAE) for spatial code reconstruction. Additionally, it is easily scalable to other generic data-types like 2D pendulum and interaction scenarios. Moreover, it is capable of enabling/disabling population coding and activating/deactivating additional motion features, which provides adequate infrastructures for different use-cases other than biological motion, along with equipping the analysis with sufficient ablation studies. Following, we address the main characteristics of our proposed architecture.

- ***Retrospective Inference***; In contrast to other similar works discussed in Section 2.7, our model uses retrospective latent state inference (i.e. using gradient descent on convex error functions for optimizing parameters), which helps us with efficient calculation of gradients for the model's adaptation purposes (Sugita, Tani, and Butz [109]; Tani et al. [111]; Tani [110]; Rumelhart, Hinton, and Williams[96]).

- ***Binding, Perspective-Taking***; Retrospective inference of binding matrix activities has the potential as a universal unit that correctly directs attention [99, 102]. The same method is used to propagate the prediction error further back onto perspective taking neurons, which translate and rotate the input features onto a known frame of reference [105]. As a consequence, the introduced generative neural network architecture is capable of solving binding, and perspective taking in a similar efficient retrospective manner.

- ***Population Coding, Redundant encodings***; We will further demonstrate that in order to obtain encodings that are capable of accurately and robustly infering the correct hidden motion patterns, population encodings and redundant relative spatial encodings (i.e. supplementary information from motion direction, posture, and motion magnitude) are highly useful. In this respect, the current model is also able to asses the performance on the raw submodal information (i.e. population encoding is disabled) as well as on the parts of the input information channels (i.e. just selected spatial encodings are enabled).

- ***Spatial Reconstruction***; In order to learn compressed representations of input and as compared to [102], we use a variational autoencoder (VAE) [62], which learns the parameters of a probability distribution representing the data rather than learning just a function that represents the data. VAE samples from it's learnt distribution and generates new data samples. VAE does not have limitations of standard autoencoders for generation as standard AE's use a latent space for generation which is not continuous and makes interpolation harder. Note that with AE's we are simply performing a non-linear extension of PCA. As a consequence, using VAE will ensure a suitable sampling and data generation and will foster the development of more balanced latent state encodings.

- ***Temporal Prediction***; For the purpose of empowering the architecture to be able to rely on its own imagination when the bottom-up input is not available, missing or considered unreliable, we equip the architecture with a LSTM [53]. Which is a prosperous option for learning time-series sequential data that efficiently tackles the vanishing gradient problem [52] and stores previous time-step information in order to bridge very long time lags.

- ***Behavior Interpretation, Code Imagination***; Our LSTM-based temporal predictions generate closed-loop predictions and a consequent behavior interpretation. In comparison with [102] (where three different temporal AE was used; one for each submodality), our behavior type inference strategy is different and is explained in Section 3.4. Subsequently, we are able to feed the VAE with a whole Gestalt, which is a concatenation of individual submodal encodings, and which uses the corresponding compressed code as input for the Behavioral module that will result in a smaller number of trainable weights and consequently a faster inference and Gestalt Perception. Additionally, even when the input stimulus is only partially present, the proposed behavioral module can correctly distinguish different Gestalt patterns and infer the actual behavior.

- ***Generalisation***; We furthermore enhance previous work by [102] in that the current architecture is capable of switching between 2D and 3D cases and is shown to be able to handle other motion patterns (i.e. not just biological motion). Subsequently, it adapts the distributed topological neurons automatically to efficiently cover areas surrounding the input stimuli. To confirm the generality of our results, we evaluate binding performance on other scenarios like a 2D pendulum and an agents chasing scenario.

- ***Other Investigations***; Implementing such a robust model will make us able to investigate other hallmarks of cognition and intelligence. For instance, what happens to Gestalt perception and intention interpretation if the observed data is rotated far away from the known perspective? The achieved observations are reported in Section 4.

- ***Employed Mechanism for Transformer Networks***; Additionally, in Section 5.2 we will hypothesize an affine transformation mechanism (similar to our retrospective binding strategy) by making use of our applied perspective taking process combined with the original work from Jaderberg et al. [56].

- ***Successful binding with distractor existence***; The model's performance will further be validated when different types of distractors that are not part of the perceived Gestalt are added to the visual input. We will demonstrate that the proposed adaptive perceptual mechanism will be tuned according to its self perceptual experiences and will successfully ignore other dynamics.

We further expect that our binding approach will be beneficial in other domains, wherever input information requires to be flexibly bound and correlated to other data on the fly, fostering overall consistency. Moreover, given that binding, perspective taking, and intention interpretation are common universal problems in cognitive science, our introduced mechanisms may be very useful for addressing similar challenges in other domains beyond dynamic motion patterns.

Chapter 3

The proposed architecture

As outlined in previous chapters, in addition to Gestalt perception, perspective taking and behavior interpretation abilities are crucial for a competent social behavior. In this chapter [1], we propose a generative, autoencoder-based neural network model, which tackles the challenges of (I) feature binding into Gestalten, (II) perspective taking, and (III) behavior interpretation simultaneously by making use of retrospective, prediction error-minimizing inference [99, 102]. Additionally, when the input stimulus is not entirely present, the model is still able to produce biased imaginations and to perform Gestalt perception. The encoder part of the architecture is equipped with a rotation matrix and a transition vector for perspective taking, as well as a binding matrix for flexibly integrating input features into one Gestalt percept. Respective neural codes are constantly striving to align sensory inputs with top-down expectations [21] and the parameters of these modules are tuned online through retrospective, gradient based latent state inference [18]. Consequently, the model attempts to integrate all bottom-up visual cues into a Gestalt from a canonical perspective by mimicking an approximate top-down inference.

Our proposed modularized architecture infers spatio-temporal relations of visual features and Gestalt templates (similar to parietal and temporal cortex), binds input features into Gestalt templates, and performs perspective-taking (similar to dorsal stream areas and parietal cortex). Therefore, the model can

[1] Some parts of this chapter is based on my published papers (see publications on 21).

partially be considered to reflect the functionalities of parietal and temporal regions of the mirror neuron system. Moreover, the employed temporal predictive processing module is closely related to kinematic intentions and frontal-lobe encodings.

For stimulating a self-grounded Gestalt perception [39], we initially train the model on a canonical perspective of an ordered set of motion features where each visual feature corresponds to a location. The said location is represented by a Cartesian coordinate with respect to a global frame of reference in order to encode origin and orientation. Eventually, the model learns a generative, autoencoder-based model of motion patterns, as well as a LSTM-based temporal prediction of the submodal codes. Next, feature binding and perspective taking are performed by propagating the reconstruction error back onto the binding and perspective parameters, where this procedure can be regarded as specialized parametric bias neurons [109, 110]. Lastly, by making use of a temporal predictive processing module we are able to distinguish different Gestalt patterns and infer the actual behavior.

Our evaluations demonstrate that it is extremely beneficial (i) to utilize population encodings of the individual features and (ii) to segregate the motion feature information into relative position, motion direction, and motion magnitudes. We assess the model's capabilities on two dimensional cases such as two joint pendulum and agent interaction scenarios, as well as on three dimensional cyclic dynamical motion patterns of different acting subjects (for instance, basketball playing, dancing, jumping, walking). Additionally, we present further investigations and provide ablation studies in chapter 4 .

3.1 Sub-Modal Population Encoding

The proposed recurrent neural network architecture (cf. Fig. 3.1) proceeds with a number of visually perceived salient features and learns a generative model of compressed embodied action patterns. Each Cartesian input coordinate x_i at time step t refers to a joint location of Acclaim Motion Capture (AMC) data [2] and will be separated into distinct sub-modalities: relative position, motion magnitude, and motion direction. (i.e. p_i, m_i, and d_i in Fig. 3.1).

The relative position of a visual feature is dependent on both the choice of origin and the orientation of the coordinate frame whereas the motion direction depends only on the orientation but not on the origin. However, unlike the relative position and motion direction, motion magnitude is totally independent of perspective.

Accordingly, given a visual input coordinate and its velocity, three distinct types of submodal information are derived and individually transformed via a translation bias b and a rotation matrix R, which specify origin and orientation. Subsequently, relative position $P_i(t)$, motion direction $d_i(t)$ and motion magnitude $m_i(t)$ signals are extracted.

At each point in time t, the following transformations are applied:

$$P_i(t) = R(t) \cdot X_i(t) + b(t), \tag{3.1}$$

determining the relative position $P_i(t)$ of the input feature $X_i(t)$;

$$m_i(t) = \|R(t) \cdot V_i(t)\|\,, \tag{3.2}$$

where $V_t = X_t - X_{t-1}$ is the velocity of each visual feature and $m_i(t)$ denotes the absolute motion magnitude;

$$d_i(t) = \frac{R(t) \cdot V_i(t)}{m_i(t)}, \tag{3.3}$$

computing the relative motion direction $d_i(t)$.

A representation of the proposed architecture in a connectivity graph, as well as the processing pipeline for a single three dimensional visual feature is depicted in figure 3.1. Following the extraction of relevant submodal information and projecting them onto a specific visual frame of reference, we encode each submodality individually by one population of topological neurons with Gaussian tuning curves. Individual Gaussian neuron centers in each submodal population are evenly distributed in the expected range of the stimulus in accordance to its range, dimension, and configuration of the perceived submodal stimuli. Such encodings are analogous to encodings found in the visual cortex and beyond

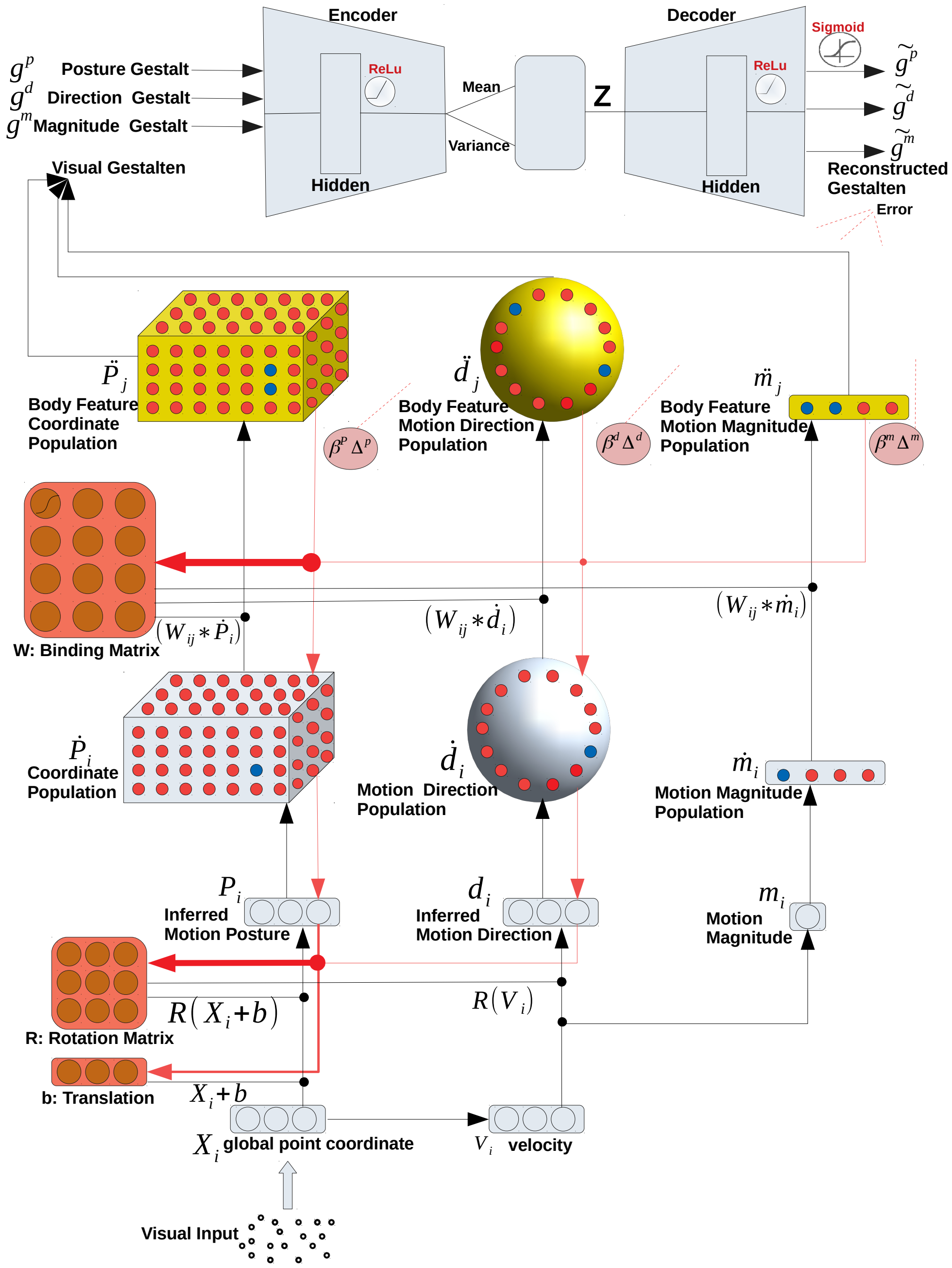


Figure 3.1: *Our generative, modularized neural network model proceeds with a visual feature i; Initially, its corresponding Cartesian coordinate, velocity, and magnitude are calculated. Next, the observed visual features are transformed by applying possible translation and rotation operations. Afterwards, each input feature is encoded into redundant posture, motion direction, and magnitude-respective population codes with 64, 32, and 8 neurons, respectively. Subsequently, a neural gating matrix selects and assigns the observed feature i to an autoencoder input slot j, which represents i's correlated bodily feature. Variational autoencoders attempt to reconstruct the posture, motion direction, and magnitude patterns. Based on the VAE's derived reconstruction loss (squared difference between input Gestalt and its reconstruction), the parameters of the gating and rotation matrices as well as of the translation vector are adapted with gradient-based, retrospective inference.*

[89]. Moreover, to evaluate the efficiency of population encoding, we also assess the model's performance on the raw submodal information, that is, without population encoding.

The response of α-th neuron associated with the i-th visually observed feature in a population that encodes the position p is calculated by:

$$\dot{P}_{\alpha,i}(t) = (r^p)^{D^p} \cdot N\left(P_i(t); c^p_\alpha, \Sigma^p\right) \tag{3.4}$$

In the equation above each neuron has a specific center c^p_α and a response variance Σ^p. Furthermore, the density of the multivariate Gaussian distribution at l with mean μ and a D^p-dimensional diagonal covariance matrix Σ is:

$$N(l;\mu,\Sigma) = \frac{1}{\sqrt{det(2\pi\Sigma)}} exp\left[-\tfrac{1}{2}(l-\mu)^T\Sigma^{-1}(l-\mu)\right] \tag{3.5}$$

The factor $(r^p)^{D^p}$ scales the neural activities, dependent on the relative distance r^p between neighboring neurons. The relative distance r^p is also used to specify the diagonal variance entries:

$$\sigma^p = \zeta^p \cdot (r^p)^2 \tag{3.6}$$

where $\zeta^p \in (0,1]$ denotes the breadth of the cell tunings.

Likewise, topological neurons' activations for direction and magnitude submodalities are determined by:

$$\dot{d}_{\alpha,i}(t) = (r^d)^{D^d} \cdot N\left(d_i(t); c^d_\alpha, \Sigma^d\right), \tag{3.7}$$

$$\dot{m}_{\alpha,i}(t) = (r^m)^{D^m} \cdot N\left(m_i(t); c^m_\alpha, \Sigma^m\right), \tag{3.8}$$

Posture neurons ($D^p = 3$; 2 for pendulum) are evenly distributed in a specific range on a grid. Direction neurons are evenly scattered on the surface of a unit sphere (D^d = 3; 2 for pendulum), while neurons that encode the motion magnitude of the observed feature are distributed linearly (D^m = 1). Based upon the motion capture data and the applied skeleton, posture, direction, and magnitude, populations are configured to have 64, 32, and 4 neurons, respectively (for 2D pendulum we distributed 8 direction neurons on a unique circle, 16 posture neurons on a rectangle area around the stimuli, and 4 linear magnitude neurons. Similarly, for interaction scenario, respective neurons are configured to automatically and efficiently cover areas surrounding the acting agents).

In sum, processing pathways of our introduced model consist of bottom-up, perceptual processing, which continuously attempts to adjust sensory inputs to top-down expectations through online adaptations in order to end up having suitable spatial predictive encodings.

3.2 Gestalt Perception and Feature Binding

While seeing an unknown motion sequence, interpreting the observed motion dynamics requires some mechanism to 'bind' the information relating the observed features to the respective learned body features and to distinguish it from other features [113]. The binding problem [114] arises when we select separate visually observed features and integrate them in a correct order.

We approach this problem by selectively routing respective feature patterns and motion dynamics, ensuring that rerouted feature patterns match expected Gestalt dynamics. Choosing the suitable observed features $i \in \{1...N\}$ (i.e. $\dot{p}_j$, $\dot{d}_j$ and $\dot{m}_j$ in Fig. 3.1) and allocating them to the correct neural processing pathway or bodily features $j \in \{1...M\}$) (i.e. $\ddot{p}_j$, $\ddot{d}_j$ and $\ddot{m}_j$ in Fig. 3.1) is addressed by an adaptive connectivity matrix (W: binding matrix), such that:

$$\ddot{p}_j(t) = \sum_{i=1}^{N} w_{ij}(t) \cdot \dot{p}_i(t), \tag{3.9}$$

$$\ddot{d}_j(t) = \sum_{i=1}^{N} w_{ij}(t) \cdot \dot{d}_i(t), \tag{3.10}$$

$$\ddot{m}_j(t) = \sum_{i=1}^{N} w_{ij}(t) \cdot \dot{m}_i(t). \tag{3.11}$$

Population-encoded activations of the i-th observed or unassigned features for posture, motion direction, and magnitude are represented by $\dot{p}_j$, $\dot{d}_j$, and $\dot{m}_j$, respectively. Their relative j-th bodily or assigned features are indicated by $\ddot{p}_j$, $\ddot{d}_j$ and $\ddot{m}_j$.

$$w_{ij} = \frac{1}{1 + exp(-w^b_{ij})} \in (0, 1) \tag{3.12}$$

denotes the corresponding assignment strength, while the adaptive parametric bias neuron activity is represented by w^b_{ij}.

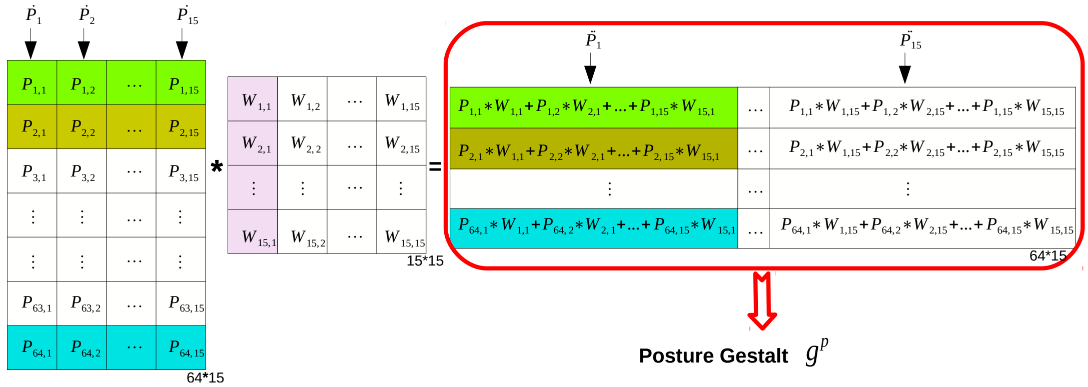
Each set of submodal bodily feature populations is concatenated into a Gestalt vector g^x:

$$g^p(t) = (\ddot{p}_1(t), \ddot{p}_2(t), ..., \ddot{p}_M(t)), \tag{3.13}$$

$$g^d(t) = (\ddot{d}_1(t), \ddot{d}_2(t), ..., \ddot{d}_M(t)), \tag{3.14}$$

$$g^m(t) = (\ddot{m}_1(t), \ddot{m}_2(t), ..., \ddot{m}_M(t)), \tag{3.15}$$

yielding Gestalt vectors for the postural, motion direction, and magnitude submodalities, respectively. A matrix visualisation of the the aforementioned routing procedure is provided in Figure 3.2.



$\dot{P}_1$
$\dot{P}_2$
$\dot{P}_{15}$
$P_{1,1}$ $P_{1,2}$ ⋯ $P_{1,15}$
$P_{2,1}$ $P_{2,2}$ ⋯ $P_{2,15}$
$P_{3,1}$ $P_{3,2}$ ⋯ $P_{3,15}$
$P_{63,1}$ $P_{63,2}$ ⋯ $P_{63,15}$
$P_{64,1}$ $P_{64,2}$ ⋯ $P_{64,15}$
64*15
*
$W_{1,1}$ $W_{1,2}$ ⋯ $W_{1,15}$
$W_{2,1}$ $W_{2,2}$ ⋯ $W_{2,15}$
$W_{15,1}$ $W_{15,2}$ ⋯ $W_{15,15}$
15*15
=
$\dot{P}_1$
$\dot{P}_{15}$
$P_{1,1}*W_{1,1}+P_{1,2}*W_{2,1}+...+P_{1,15}*W_{15,1}$ ⋯ $P_{1,1}*W_{1,15}+P_{1,2}*W_{2,15}+...+P_{1,15}*W_{15,15}$
$P_{2,1}*W_{1,1}+P_{2,2}*W_{2,1}+...+P_{2,15}*W_{15,1}$ ⋯ $P_{2,1}*W_{1,15}+P_{2,2}*W_{2,15}+...+P_{2,15}*W_{15,15}$
$P_{64,1}*W_{1,1}+P_{64,2}*W_{2,1}+...+P_{64,15}*W_{15,1}$ ⋯ $P_{64,1}*W_{1,15}+P_{64,2}*W_{2,15}+...+P_{64,15}*W_{15,15}$
64*15
Posture Gestalt g^p

Figure 3.2: *Population encoded activations of each observed joint ($\dot{p}_i$) are placed in separate columns (for this example posture submodality was considered). Afterwards, a binding matrix W selects the observed features $\dot{p}_i$ and assigns them to their respective bodily features $\ddot{p}_j$. In the end, the complete set of perceived bodily feature populations is joined into a posture Gestalt vector (g^p).*

For every submodal Gestalt $x \in \{p, d, m\}$, one variational autoencoder (VAE) [62] was used in order to learn predictive Gestalt encodings. Thus, the bottom-up submodal Gestalt vectors g^x are passed through the autoencoder, resulting in generating reconstructions of the Gestalt perceptions. Taken together, the respective autoencoders learn distributed, predictive encodings of actions for each submodal Gestalt vector and subsequently they generate posture, motion direction, and magnitude predictions. Thereby, a sufficiently trained autoencoder always infers the closest known stimulus pattern based on its self-perceptual, or embodied experience, even when it observes actions with unknown identity of the features and perspectives.

After the model has learned, the difference between the Gestalt input to the variational autoencoder and the reconstructed Gestalt output is to be minimized by adapting the parametric bias neurons' activities of the perspective taking and feature binding modules.

We indicate the relative squared losses by $\mathcal{L}_p$, $\mathcal{L}_d$, and $\mathcal{L}_m$, respectively. We scale the loss signals by corresponding adaptation factors β^p, β^d, and β^m to balance the error signal impacts. The particular adaptation of the parametric bias neurons' activities $w^b_{ij}(t)$ is calculated by typical gradient descent with momentum:

$$\Delta w^b_{ij}(t) = \; -\eta^f \frac{\partial \mathcal{L}(t)}{\partial w^b_{ij}(t)} + \gamma^f (w^b_{ij}(t-1) - w^b_{ij}(t-2)), \tag{3.16}$$

where η^f denotes the learning rate, and γ^f the momentum; for the feature binding adaptation procedure, and the loss signal $\mathcal{L}(t)$ yields:

$$\mathcal{L}(t) = \beta^p \mathcal{L}_p(t) + \beta^d \mathcal{L}_d(t) + \beta^m \mathcal{L}_m(t) \tag{3.17}$$

During training, for all $i = j$ the assignment biases w^b_{ij} are set to 1000 (resulting in $w_{ii} \approx 1$) and for all $i \neq j$ to -1000 (resulting in $w_{ij} \approx 0$), since the assignment is fixed during self-observations. During testing, all assignment biases are initialized at -5 (which leads to an initial subtle mixture of all possible assignments) and are updated through time by making use of Eq 3.16.

It should be noted that the ideal assignment strength w_{ij} yields an activation of 1 for the bias neuron that connects an observed feature with its respective body feature, and an activation of 0 otherwise. In that regard, we introduce Feature

Binding Error (FBE) to evaluate the binding progress over time;

$$FBE(t) = \sum_{j=1}^{M} \sqrt{(w_{jj}(t) - 1)^2 + \sum_{i=1, i \neq j}^{N} w_{ii}(t)^2}. \tag{3.18}$$

FBE measures the sum of Euclidean distances between the model's inferred assignment and the correct assignment.

3.3 Perspective Taking

Spatial perspective taking allows humans to recognize movement patterns in the motion of an observed person by mentally adopting their perspective [60] and therefore, can be considered as a mental transformation process that aligns the observer's perspective with a self-centered perspective.

The employed mechanism consists of a translation followed by a rotation of all visually observed features and is based on [106]. Subsequently, our proposed model learns and applies derived signals by which a mental perspective should be adopted and thus it accurately attempts to co-adapt the top-down autoencoder-based reconstructions, and the bottom-up input routing, manifested in translation- and rotation-based perspective taking and binding, that is, the fundamental neurocomputational mechanism of perspective-taking and mental rotation.

Translation encodes the origin of the model's internal, imagined frame of reference in addition to the center of rotation. It is determined by 0-initialized bias neurons b_a, which are adapted to minimize the top-down loss signal $\mathcal{L}(t)$ by means of gradient descent with momentum;

$$\Delta b_a(t) = \; -\eta^b \frac{\partial \mathcal{L}(t)}{\partial^t b_a(t)} + \gamma^b (b_a(t-1) - b_a(t-2)) \tag{3.19}$$

where $a \in \{x, y, z\}$ represents the affected axis, γ^b the momentum term, and η^b the adaptation rate. It should be noted, however, that motion magnitude and direction are invariant to translations. Accordingly, the adaptation is determined by the posture-respective weighted error signals $\beta^p \Delta^p_{1...M}$ only (cf. also

Figure 3.1).

Rotation is performed by making use of a neural 3×3 matrix R, which is driven by three Euler angles α_x, α_y , and α_z, each of which indicates a rotation activation around a specific Cartesian axis;

$$R = R_x(\alpha_x(t))R_y(\alpha_y(t))R_z(\alpha_z(t)) \tag{3.20}$$

is the corresponding connectivity structure. Correspondingly, the respective matrices of activation functions are:

$$R(\alpha_z) = \begin{bmatrix} cos\,\alpha_z & -sin\,\alpha_z & 0 \\ sin\,\alpha_z & cos\,\alpha_z & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.21}$$

$$R(\alpha_y) = \begin{bmatrix} cos\,\alpha_y & 0 & sin\,\alpha_y \\ 0 & 1 & 0 \\ -sin\,\alpha_y & 0 & cos\,\alpha_y \end{bmatrix} \tag{3.22}$$

$$R(\alpha_x) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos\,\alpha_x & -sin\,\alpha_x \\ 0 & sin\,\alpha_x & cos\,\alpha_x \end{bmatrix} \tag{3.23}$$

Note that, visuo-spatial perspective-taking is considered to be a mental procedure that rotates, and translates the observed perspective onto the vantage points of others (Hegarty and Waller, [49]). Likewise, our proposed architecture imagines the entire visual percept from the perspective of another actor.

Analogous to translation, rotation is represented by zero-initialized bias neurons, which can be adapted online by means of gradient descent. The adaptation through time follows the rule:

$$\Delta\alpha_a(t) = \ -\eta^r \frac{\mathcal{L}(t)}{\partial^r \alpha_a(t)} + \gamma^r(\alpha_a(t-1) - \alpha_a(t-2)), \tag{3.24}$$

where $a \in \{x, y, z\}$, η^r indicates the adaptation rate, and γ^r the momentum. Magnitude is invariant to rotation by nature, and therefore it is not considered for rotation adaptation. As a result, we consider posture and direction respective weighted error signals throughout the rotation adaptation process (i.e. $\beta^p \Delta^p_{1...M}$ and $\beta^d \Delta^d_{1...M}$).

In order to evaluate the transformation development, we introduce orientation difference (OD) measure to be

$$OD(t) = \frac{180}{2\Pi} Acos(\frac{tr(A^{model}(t)A^{data}(t)) - 1}{2}) \text{ in}^\circ, \tag{3.25}$$

where A^{data} is a constant rotation matrix applied to all visual inputs before testing and A^{model} is the dynamic, currently inferred rotation matrix. Moreover, we employ a translation difference (TD) in order to monitor the translation adaptation progress over time;

$$TD(t) = \left\| b^{data}(t) - b^{model}(t) \right\| \text{ in cm}, \tag{3.26}$$

However, it is vital to note that the model has no knowledge concerning FBE, OD nor TD instead it employs its self-generated submodal expectations (embodied imagination) for its adaptation intentions. Additionally, I provided some important mathematical equations and partial derivatives of the employed Gradient Descent processes in Appendices A and B.

3.4 Temporal Predictive Processing Module

In addition to learning compressed spatial codes of the population encoded stimuli, our architecture also requires to learn temporal representations that predict the progress of the developing codes over time. Accordingly, our proposed model adds a recurrent behavioral module (cf. Figure 3.3), which enables the imagination of temporal dynamics (that is, the progress of the multimodal Gestalt stimuli whenever it is not present) and the inference of the type of observed behavior. To interpret the observed action types during testing, the trained behavioral module will be utilized as an inference mechanism.

3.5 Behavior Interpretation

A one-hot encoded action type vector is given as input to the behavioral module during training. Furthermore, as depicted in Figure 3.3, we feed the concatenated submodal populations (i.e. whole Gestalt) as input to the variational autoencoder and train the behavioral module according to the acquired compressed code (latent vector Z).

$$g^{whole}(t) = (g^p(t), g^d(t), g^m(t)) \tag{3.27}$$

The behavioral module contains an LSTM followed by a linear layer. It receives the latent vector Z in addition to the actual one-hot encoded action as input and predicts the latent vector at the next time step.

We will further demonstrate that this module has the ability to infer the current behavior that is being observed, as well as to imagine temporal behavioral dynamics when the bottom-up input deemed unreliable or is missing (i.e. Stimulus Presence factor is turned to zero). During our conducted experiments we represent *Stimulus Presence* as the number of time steps the input stimulus is present, accessible and reliable. As an example, $StimulusPresence = 100$ indicates that

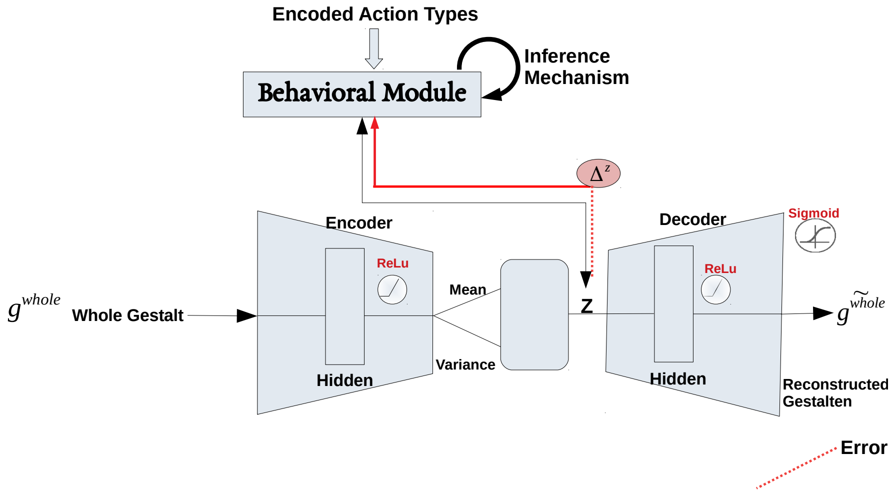


Figure 3.3: *A behavioral module makes use of its prediction error of the VAE's latent vector (Z; which is also referred to as the compressed code) to bias the imagination of submodal codes, in addition to interpret the actual perceived behavior.*

the input stimulus for a particular action is available for the first 100 frames and during the rest, the model has to employ its own embodied imagination.

Note that, once the *Stimulus Presence* factor turns to zero, the VAE is not receiving the input Gestalt any longer. Instead, the behavioral module switches to closed-loop behavioral processing without input-signal-dependent adaptations of the latent state Z, which enables the architecture to generate the respective compressed Gestalt code Z.

Nevertheless, we still utilize the bottom-up submodal populations to compare the imagined reconstructed Gestalt with the bottom-up input, for adjusting binding, rotation matrix and translation vector activities. During testing and by the time when the stimulus is no longer available or when the entire stimuli is not shown, the trained behavioral module's temporal prediction enables consistent imaginations that leads to proper inference of kinematic intentions.

The implemented inference procedure does a Grid search over all possible one-hot encodings and selects the one that results in smallest mean prediction error over previous time steps (although also gradient-based inference adaptation procedure may be applied in follow-up studies). Eventually, the respective imagined Gestalt percept will be employed for the model's further top-down perspective taking and feature binding processes.

In sum, the extracted submodal information from the input bodily features will be separately encoded into populations of topological neurons. Afterwards, the acquired compressed code will be developed and predicted in accordance with its corresponding spatial and temporal contingencies (i.e. embodied learning of submodal spatial codes). The resulting signal will further be used to learn predictions of succeeding codes. Subsequently, the predictive processing component can be seen as a module that provides the driving signal that enables embodied, intention-specific simulations, and behavior interpretations.

Taken to together, the proposed model attempts to:

1. Learn action patterns of submodal perceptions in compressed, generative formats.

2. Learn spatio-temporal imagination of submodal compressed codes and ac-

tion embeddings.

3. Generate embodied expectations in order to bias the simulation of submodal codes.
4. Resolve the feature binding perspective taking challenges, while simultaneously interpreting the actual observed behavior.
5. Selectively deactivate parts of the redundant relative spatial encodings, and the population encodings in order to highlight their significance.

Chapter 4

Experimental Results

In this chapter[1], I evaluate the introduced architecture in different scenarios and provide ablation studies in order to investigate the motivations addressed in section 2.8. I initially outline the formats of the used data and the properties of the employed topological parameters. Then the model's performance will be examined while it is carrying out feature binding, perspective taking, and behavior interpretation tasks. In the end, we attempt to explore the proposed model further by presenting ablation studies and other use cases.

4.1 Data Configurations

To assess the models' performance we utilized human motion capture data provided by the Graphics Lab of the Carnegie Mellon University [2]. The employed 3D cyclic and continuous motion capture data is represented in Table 4.1 from which we selected 15 limbs (out of 30 available limbs in the skeleton files).

A snapshot and a schematic of chosen body joints are illustrated in Figure 4.1

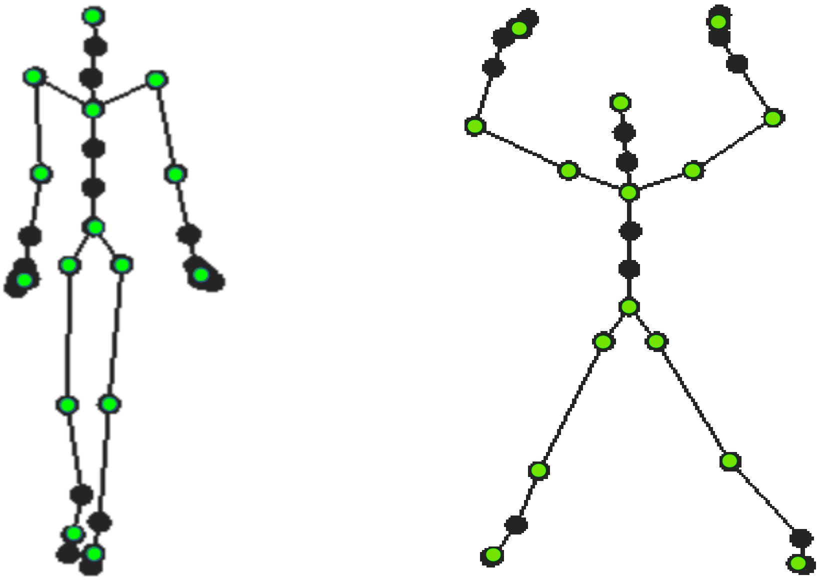
4.2 VAE and LSTM Prediction Errors

For adapting LSTM and VAE modules, TensorFlow and PyTorch libraries of Python are used. Moreover, feature binding and perspective taking adaptations

[1] Some parts of this chapter is based on my published papers (see 21).

Table 4.1: *Employed data for training and testing of the architecture*

	Training			Testing		
Motion Type	Subject Nr.	Trial Nr.	Time Steps	Subject Nr.	Trial Nr.	Time Steps
3D Walking	35	7	1030	5, 12	1, 1	598, 523
3D Jumping Jack	23	15	495	22	16	432
3D Dancing	55	2	1:1000	55	2	1000:2180



(a) *Walking Actor*

(b) *Jumping actor*

Figure 4.1: *Green dots indicate the 15 selected visual inputs from all available body features of motion capture data.* ***(a)*** *Skeleton data of a walking subject.**(b)*** *Skeleton data of a subject performing jumping jacks.*

are obtained by minimizing self-generated prediction errors through gradient descent. During training the input stimulus is fully accessible by the model and is perceived from an egocentric viewpoint. The training target is to:

1. Examine the compression quality of the autoencoder for learning action patterns of submodal perceptions in generative and compressed formats.

2. Evaluate the prediction quality of the LSTM in order to learn spatio-temporal

imagination of action embeddings and submodal compressed codes.

Note that feature binding and perspective taking adaptations are disabled during the model's training. To identify ideal parameter settings, Grid search was used and the configured network parameters are listed in Table 4.2. For training the LSTM a learning rate of 7.10^{-2} together with 60 cells were used. Furthermore, both LSTM and VAE were adapted using ADAM optimisation algorithm [61].

Additionally, the raw Cartesian data was fed into another VAE module for making the architecture able to assess the influence of population coding. As shown in Figure 4.2, the model considerably improves its reconstruction error over time while activating as well as deactivating the population coding.

Although we can not directly compare cases with and without population encoding as they have different loss measures, it is conspicuous that learning occurs in both cases. It is also well-noticeable that when applying the population encoding the respective reconstruction error improvement specifically for posture and motion direction is much stronger; (see the different ranges of the y-axes in Figure 4.2).

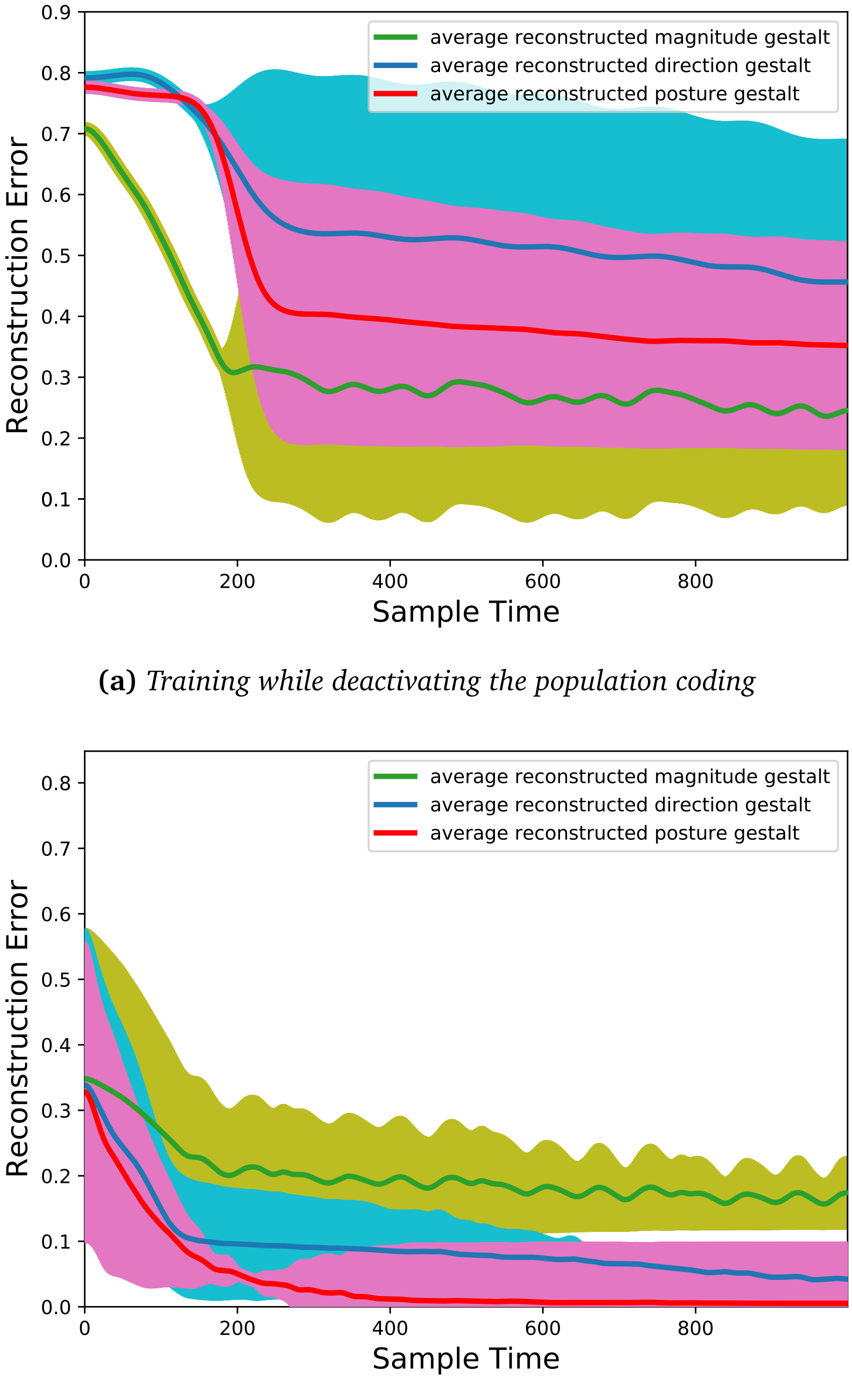
It is vital to consider that all results described in this chapter demonstrate averages (and standard deviations where possible) over ten independently trained neural networks (for 1000 epochs and 1000 time steps), unless stated otherwise.

Table 4.2: *Assigned parameters for variational autoencoder's training*

Experiment	Learning Rate Pos	Learning Rate Dir	Learning Rate Mag	Opt.	Hidden Size	Latent Size	ζ^p	ζ^d	ζ^m
Motion (pop. coding)	$1 \cdot 10^{-3}$	$8 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	Adam	45	25	0.85	0.85	0.95
Motion (raw data)	$1 \cdot 10^{-3}$	$2 \cdot 10^{-5}$	$8 \cdot 10^{-4}$	Adam	25	10	-	-	-
Pendulum	$1 \cdot 10^{-2}$	$1 \cdot 10^{-2}$	$1 \cdot 10^{-3}$	Adam	45	25	0.85	0.85	0.95
Experiments 1-6	$1 \cdot 10^{-2}$	$9 \cdot 10^{-1}$	$1 \cdot 10^{-2}$	Adam	20	30	0.85	0.85	0.95

The temporal code imagination error while seeing three different types of actions (dancing, walking, and jumping) during training is indicated in Figure 4.3 where every 400 time frames a new action was presented to the model.

Note, though, that the final aim of training is not to minimize the reconstruction error as much as possible but instead its goal is to train the VAE sufficiently



(a) *Training while deactivating the population coding*

(b) *Training while activating the population coding*

Figure 4.2: *Visual spatial reconstruction error (i.e. Binary Cross-Entropy (BCE) Loss) of magnitude, direction, and posture Gestalt averaged over 10 independently trained neural networks* ***up****: trained without using the population encoding;* ***down****:trained while using the population encoding.*

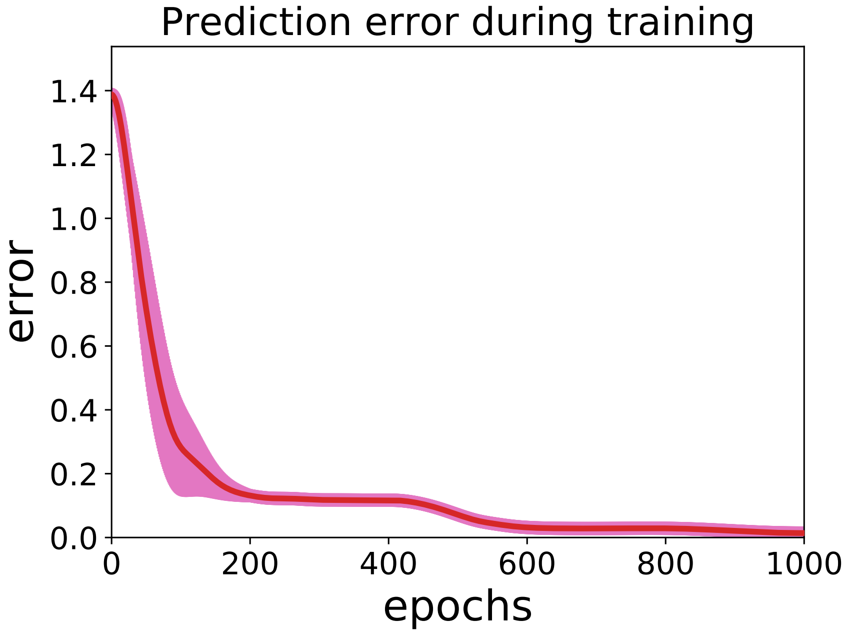


Figure 4.3: *Depicts the training error; the behavioral module learns to predict the latent vector Z of the variational autoencoder during training. This will add the imagination capability to the model during testing and once the stimulus is not entirely present or deemed to be unreliable.*

enough so that its generated error can be exploited to adjust the network's parameterization; that are its binding matrix, rotation, and translation values. Subsequently, we are required to train the LSTM accurate enough such that its prediction error over the latent vector Z, can be used to interpret the actual observed behavior during testing.

4.3 Adaptation of Feature Binding

The results reported in the previous section imply that LSTM learns an adequate prediction and VAE learns good encodings. The main question, though, is whether the resultant loss signals are beneficial in order to perform perspective taking and feature binding tasks, as well as to interpret the observed behavior.

To evaluate the feature binding ability of the proposed model, perspective taking is disabled. For each test trial, bias activities of neural feature binding were

Table 4.3: *Architectural hyperparameters used for feature binding adaptations*

Experiment 1	Experiment 2	Experiment 3	Experiment 4
No Pop Code	With Pop Code	With Pop Code	2D Pendulum
β^{pos}= 5	β^{pos}= 6	β^{pos}= 8	β^{pos}= 1
β^{dir}= 1	β^{dir}= 0	β^{dir}= 2	β^{dir}= 8
β^{mag}= 0.125	β^{mag}= 0	β^{mag}= 0.125	β^{mag}= 2
η^{f} = 1	η^{f} = 1	η^{f} = 1	η^{f} = 0.1
γ^{f} = 0.9	γ^{f} = 0.9	γ^{f} = 0.9	γ^{f} = 0.9

reset to -5 (i.e. $w^b_{ij} = -5$), leading to an assignment strength of $w_{ij} \approx 0.0067$ and uniform distribution of all value information efficiently over the VAE inputs.

It should be noted, however, that since all elements of the binding matrix were initialized with the same value, the model has no knowledge concerning the correct assignments. Consequently, shuffling the order of the observed features for evaluation intentions is not needed. Architectural hyperparameters chosen for all experiments are shown in Table 4.3.

The applied gradient-based, retrospective inference procedure denotes the implemented adaptive processes of binding, perspective taking, and Behavior type inference by attempting to minimize the reconstruction loss over time as much as possible.

The Feature Binding mechanism starts with assigning very low initial activities to all elements in the binding matrix W. Then it develops an adaptive assignment of the input features to respective (learned) body features where during this procedure the top-down Gradient descent based prediction error, which is generated by VAE, adapts entire activities in the binding matrix. Thereafter, the translation vector b and the rotation matrix R activities are adapted by backpropagating the achieved reconstruction error further on. Thereby, the visual input features are translated and rotated onto a known frame of reference.

Accordingly, the model maps the unknown allocentric view point onto the known egocentric frame of reference it was already trained on.

Figure 4.4 illustrates that in all encoding cases the feature binding error decreases. Additionally, it depicts the resulting confusion (binding) matrix values at the end of the binding adaptation process (that is, after 1000 time steps).

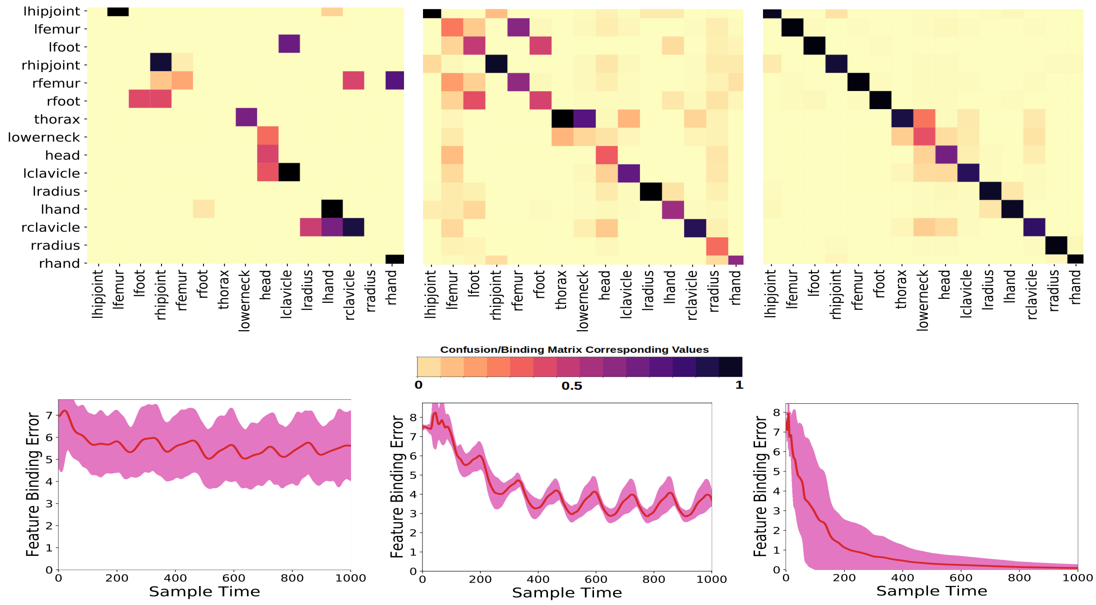


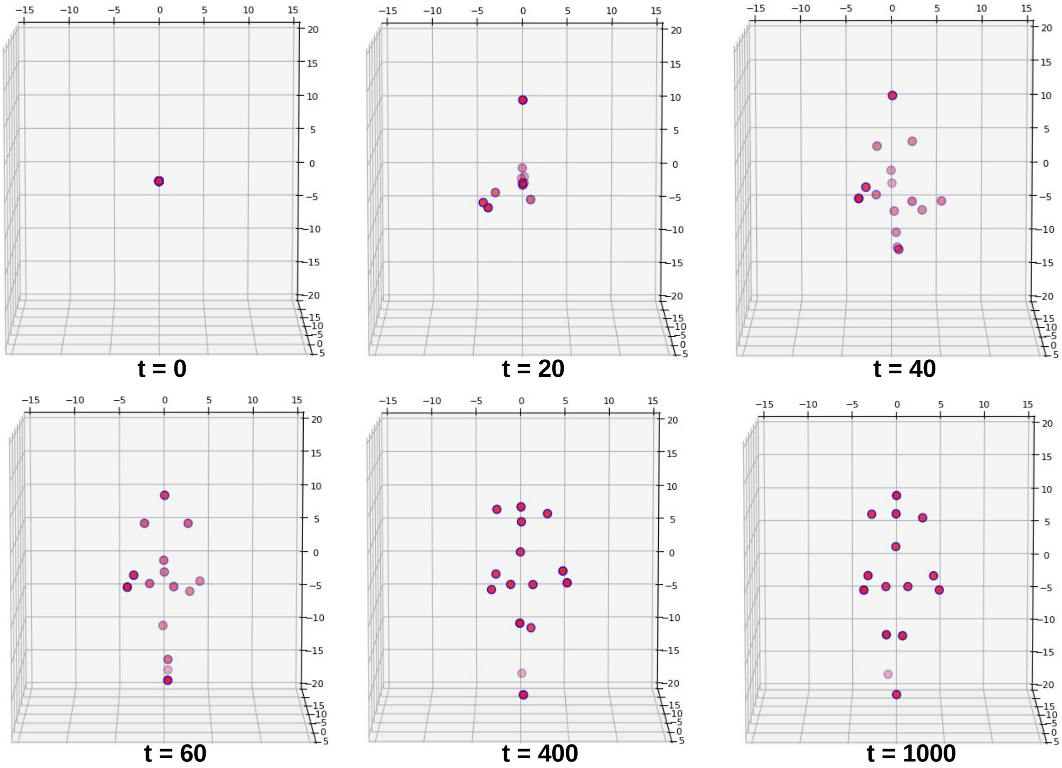
Figure 4.4: *Feature binding progress and adaptations according to experiments 1 to 3 (see Table 4.3). Each column indicates the outcome of respective experiments.* ***Upper Row*** *Demonstrates the resulting confusion matrices in which input features (columns) were assigned to the corresponding body features (rows).* ***Lower Row*** *Shows the respective feature binding error progress over time.*

As can be seen from Figure 4.4, the employed population encoding is highly effective in decreasing the feature binding error which leads to a more diagonal binding matrix (experiments 2 and 3) while without population encoding the confusion matrix does not entirely converge (experiment 1).

The confusion matrix in the left of Figure 4.4 represents experiment 1, in which the population encoding was disabled. Inside the resulting matrix the thorax joint is confused with the neck, femur is confused with the hip joint and even strong confusions can be noticed such as the left clavicle with the left foot. When providing the posture population encoding, right-left and adjacent limb confusions remain (cf. middle confusion matrix in Fig. 4.4). Consequently, when entire motion information is added in population encoded format (cf. right confusion matrix in Fig. 4.4), barely any confusions remain, denoting full Gestalt perception.

In sum, although some feature binding constellations can be found to be similarly efficient, when enabling the population encoding the feature binding error drops considerably lower. It should be noted, however, that the likelihood of estimating the correct assignment by chance is practically impossible (i.e. 225 choose 15 leads to $9.1 \cdot 10^{22}$ possibilities).

Figure 4.5 shows the progress of feature binding of the model on an observed actor. The reported results clearly indicate that the VAE is capable of identifying the correct bindings, especially when population encoding is used and supplementary feature information is provided.



t = 0

t = 20

t = 40

t = 60

t = 400

t = 1000

Figure 4.5: *Adapting the feature binding. When population encoding is enabled and complementary feature information is available the model is able to associate each input feature to its correct matching body feature as early as ∼400 time steps.*

4.4 Feature Binding with Distractor Existence

To evaluate the models' reaction to distractor features that do not belong to the perceived gestalt, two additional points were added to the observed data during testing. Disabled biological features were selected as dynamic distractors and two stationary points were used as static distractors.

4.4.1 Two Static Points as Distractor

In the following experiment, the model perceived two extra stationary features. As shown in Figure 4.6, albeit small confusions exist, the model does not assign the observed distractors to any bodily features and correctly neglects them.

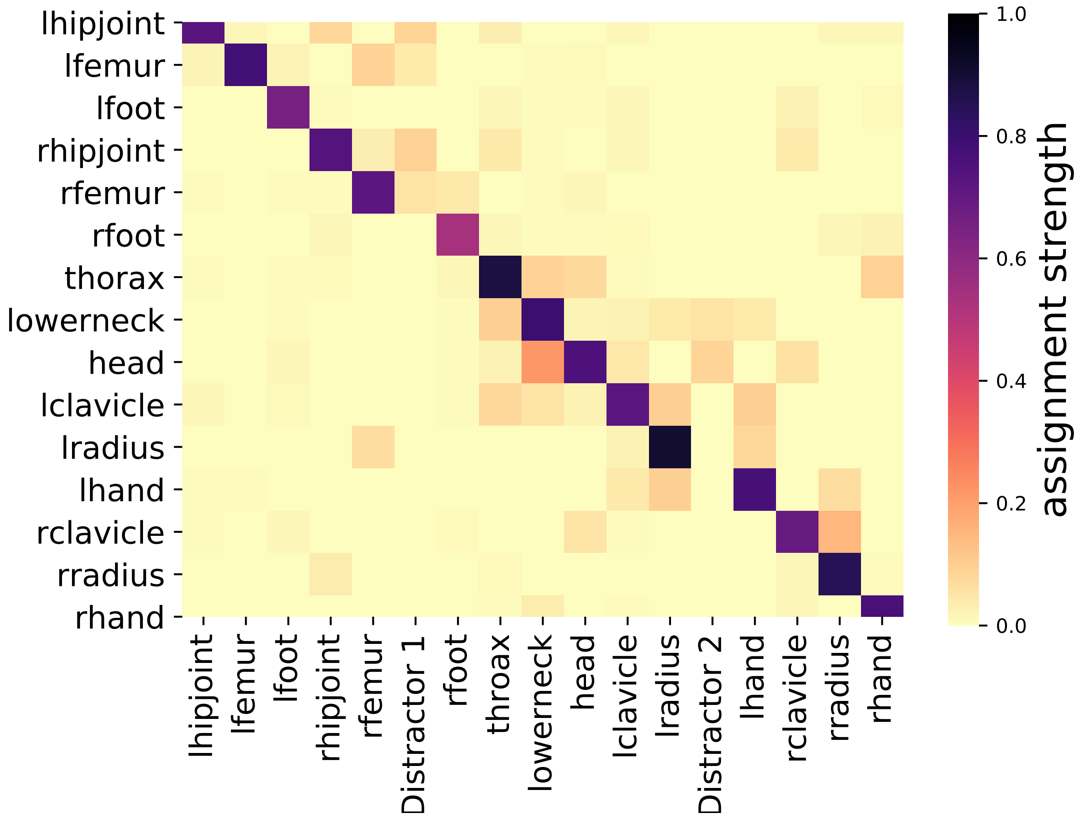


Figure 4.6: *Adapting the feature binding while two static points were inserted as distractors. Distractors 1 and 2 are fixed coordinates near hip joint, and above head joint, respectively.*

4.4.2 Two Rhythmic Input Features as Distractor

In the following experiment, the model was presented with two additional periodically moving distractors. We chose two biological features (distractor 1: right toe, and distractor 2: left wrist) that were not assigned to any bodily features throughout the training. These biological distractors mimic motion dynamics of neighboring joints and produce perceptual ambiguities. As indicated in Figure 4.7, despite existing some minor confusions between distractor 1, right hand, and right femur as well as between distractor 2, left hand, and left radius, the model appropriately assigns input features to their corresponding body features.

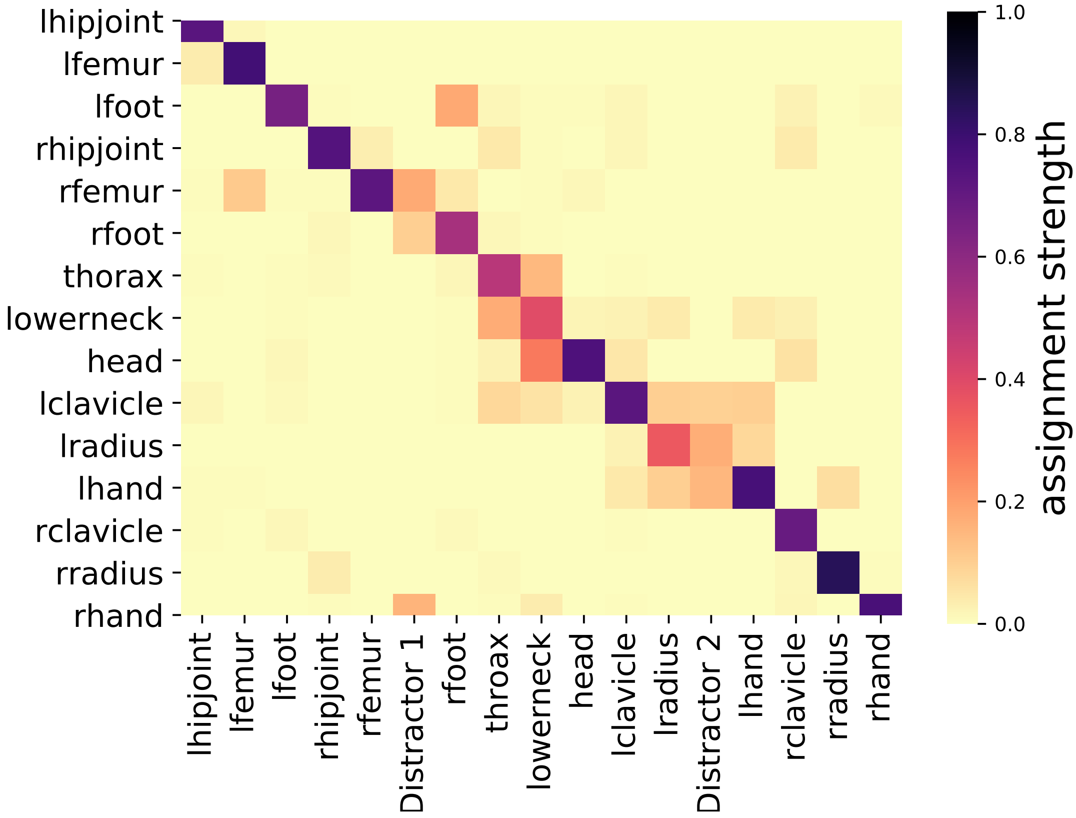


Figure 4.7: *Adapting the feature binding while two biological input features were used as distractors. Two disabled observed features (irrelevant features) during training, were employed as dynamic distractors.*

Table 4.4 denotes the duration which took until the respective confusion matrix reaches a sufficently good convergence (i.e. FBE less than 3). As Table

4.4 implies, the average convergence duration while injecting two biological distractors took the longest.

Experiment	Average Convergence Duration
Without distractors	$\sim$400 time steps
Static distractors	$\sim$650 time steps
Dynamic distractors	$\sim$900 time steps

Table 4.4: *Feature Binding duration*

Taken together, the proposed model's performance was more robust to non-biological distractors that are not part of the perceived gestalt, than to biological distractor features. The above findings are in agreement with results reported in other psychometric studies [85] in which Pavlova and Sokolov assessed the success rate of multiple subjects in detecting a point light walker within a point light display that included biological distractor features.

4.5 Adaptation of Perspective Taking

To focus on perspective taking, feature binding was disabled. The development of spatial translation and orientation was assessed, while the confusion matrix was set to the correct diagonal binding values [2]. For this purpose, we make use of our fully trained model and assess the progress of spatial translation and orientation parametric biases while the binding matrix is fixed to diagonal.

During testing and before feeding the data into the model, we transform each trial's observed data by a constant translation vector followed by a constant rotation matrix. As a consequence, the model observes the data from an unknown frame of reference with an unknown viewpoint and should transfer it to its known egocentric perspective, which is derived from the model's self-perceptual experiences during training.

Chosen parameters for perspective taking adaptation processes are indicated in Table 4.5.

[2] A video of the perspective taking and feature binding adaptation procedures and development is available here: https://uni-tuebingen.de/de/206397

Table 4.5: *Hyperparameters used for perspective taking adaptations*

Parameter	η^r	γ^r	η^b	γ^b	β^{pos}	β^{dir}	β^{mag}
Value	1.10^{-2}	9.10^{-1}	8.10^{-2}	9.10^{-1}	8	3	0

Figure 4.8 demonstrates the inference progression of the rotation matrix, when the input data was rotated by randomly assigned angles (that are, α_x, α_y, and α_z; corresponding values are shown in the legend). Analogously, Figure 4.9 illustrates the respective dynamic inference progression of the translation vector. It may be noted that the reported rotation and translation progression results are obtained based on one of the trained networks (other trained networks had qualitatively similar results).

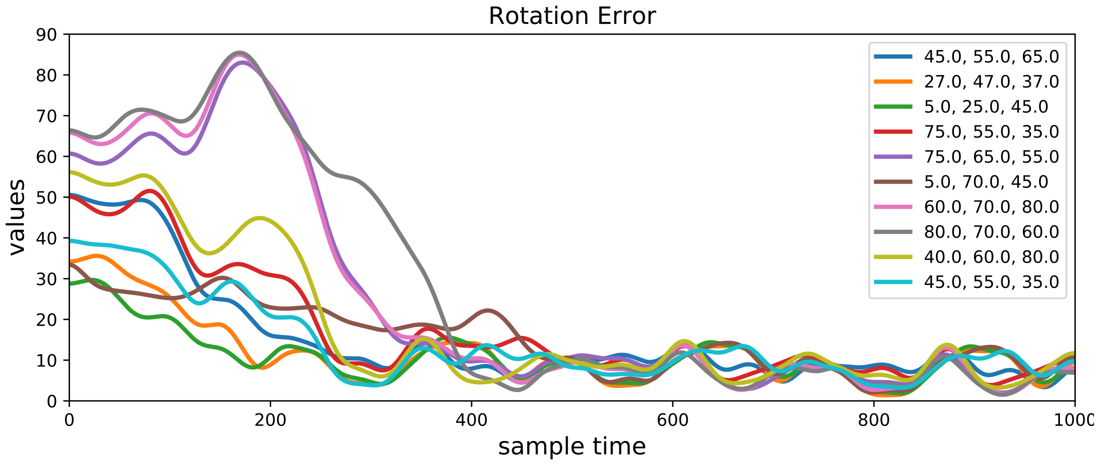


Figure 4.8: *Dynamic inference of the rotation matrix; the legend shows the initial input data rotations over each axis. The model was capable of inferring the orientation of the previously unseen test data.*

Based on the reported results for both translation and rotation, the duration to infer the correct perspective is in accordance with the perspective disturbance (that is, the duration to infer the correct perspective takes longer when the perspective is disturbed at a stronger level). Similar findings are reported by Pavlova and Sokolov [85]. Rather intense rotations of close to $90°$ on all three axes are hardest. Likewise, extreme translations result in delayed convergence, as can be expected.

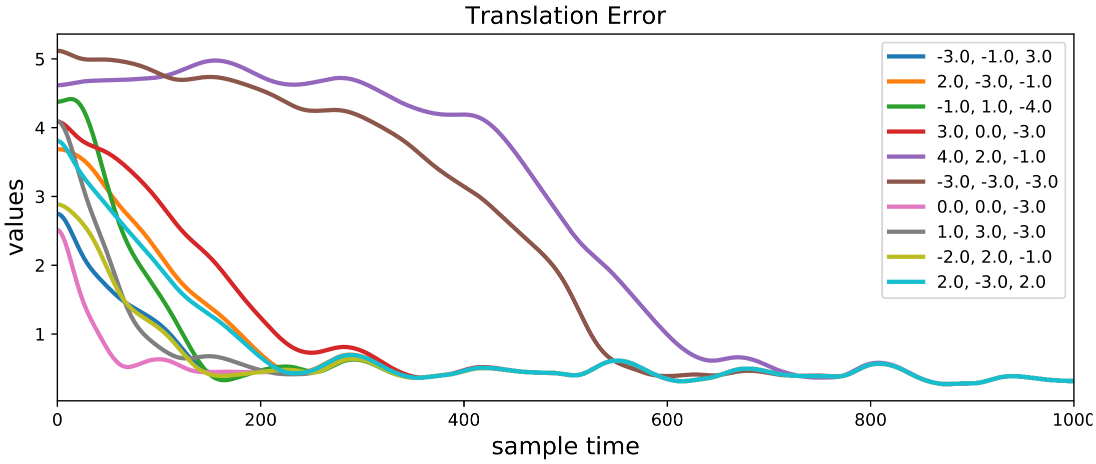


Figure 4.9: *Translation adaptation progress over time; the initial translation biases assigned to the test set are indicated in the legend (that are, b_x, b_y, and b_z). The model was successfully able to infer the translation of the formerly unseen test data.*

In this respect, Figure 4.10 illustratively shows the development of perspective taking on an observed actor from an allocentric point of view where the entire input data was translated and rotated with ($b_x = 0$, $b_y = 5$, $b_z = 0$, $R_x = 0°$, $R_y = 60°$, $R_z = 0°$).

To examine the perspective taking capabilities further, we exemplarily look at the performance of a case where both translations and rotations are applied to the input data (we set $b_x = -2$, $b_y = 2.5$, $b_z = -4$, $R_x = 25°$, $R_y = 35°$, $R_z = 45°$). Figure 4.11 indicates that also in the case of perspective taking the population encoding is extremely beneficial. Moreover, Figures 4.10 and 4.11 confirm that rotation and translation distortions can be optimized concurrently.

4.6 Behavioral Module and Ablative Studies

In this section I will provide ablation studies. In order to highlight the significance of population encodings and redundant relative spatial encodings, we will selectively disable population encodings and eliminate parts of the redundant spatial encodings. Table 4.6 summarizes all of the experiments that were con-

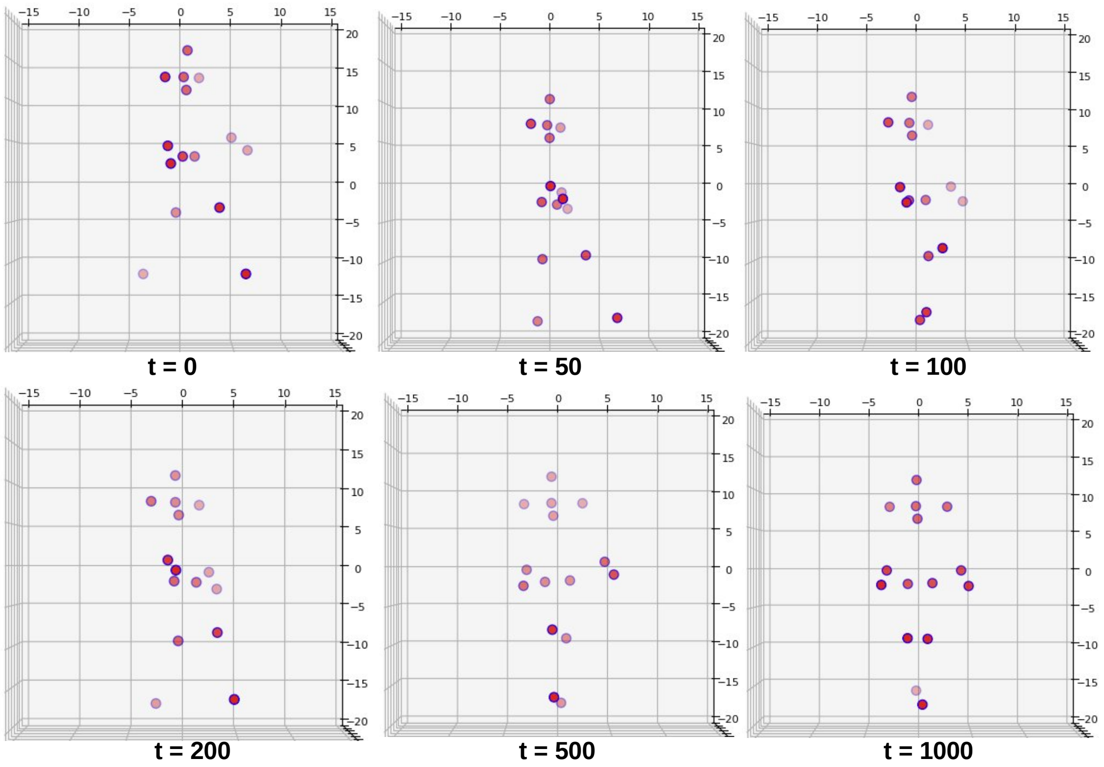


Figure 4.10: *Adapting the perspective taking; The model accurately processes a transformation that aligns the observer's view point with a self-centered, canonical perspective.*

ducted; first investigated experiment contains the most accurate performance evaluation when all information is available. During other conducted experiments, the bottom-up stimulus activity of each biological motion subsequence is limited to the first 100 of the 400 time steps.

Experiment 3 measures the perspective taking capability of the model. Within experiments 4, 5, and 6, we evaluate the model's feature binding ability by inferring the binding matrix W activities (that is, assignment's strength).

Besides, the importance of the redundant relative spatial encodings and the population encodings for feature binding and action type inference are confirmed by experiments 5 and 6.

Employed hyperparameters for training VAE and LSTM modules are described in Section 4.2 and in Table 4.2. It should be noted, that the input stimuli was

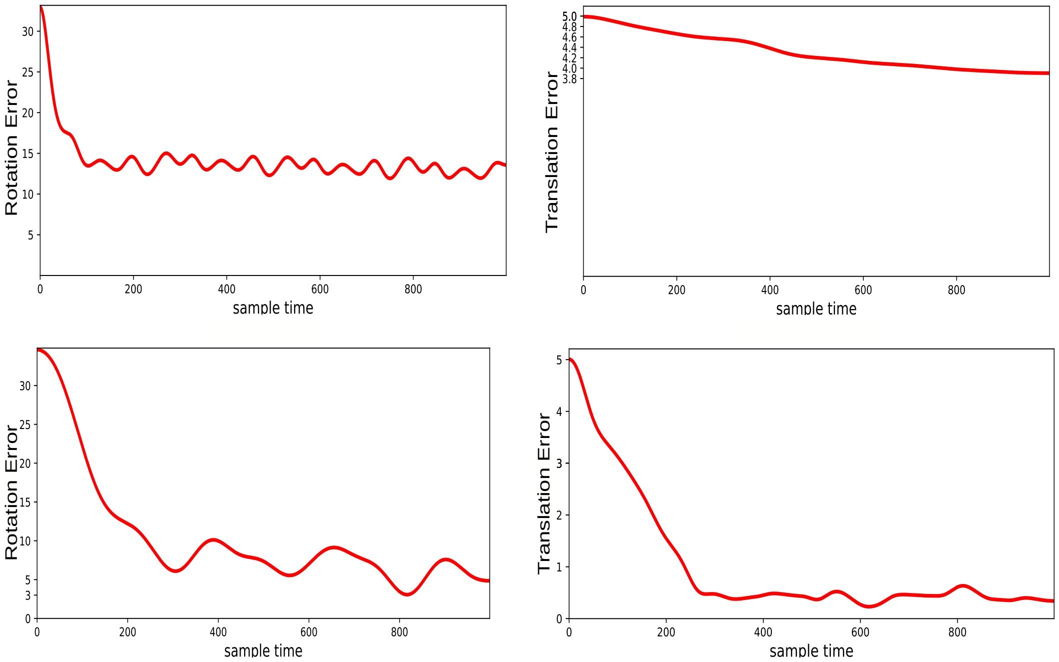


Figure 4.11: *Adapting translation and rotation;* ***Upper-Row:*** *While population coding is disabled;* ***Lower-Row:*** *While population coding is enabled*

Table 4.6: *Explored experiments*

Exp. no.	Stimulus Encoding	Stimulus Presence	FB	Perspective
1	pop code	always	provided	provided
2	pop code	100 Frames	provided	provided
3	pop code	100 Frames	provided	inferred
4	pop code	100 Frames	inferred	provided
5	scalar	100 Frames	inferred	provided
6	posture p. c.	100 Frames	inferred	provided

always present during training or self perceptual experiences. After an arbitrary number of frames (here 100) during testing, the Stimulus Presence factor will become zero. Subsequently, the bottom-up input data is no longer fed into the VAE which leads to enabling the closed-loop processing where the behavioral module generates latent states Z.

Figure 4.12 indicates the code prediction error during testing the model with experiment four and upon observing three different types of behaviors (dancing, jumping, walking) listed in Table 4.1 where every 400 time frames a new action was fed into the model. Additionally, the provided testing action-type sequence for experiments presented in Table 4.6 is first Walking afterwards Jumping followed by Dancing, 400 frames each. In this regard, Figure 4.13 indicates the output of behavior type inference on a model which is trained on three types of actions (Dancing, Walking, and Jumping).

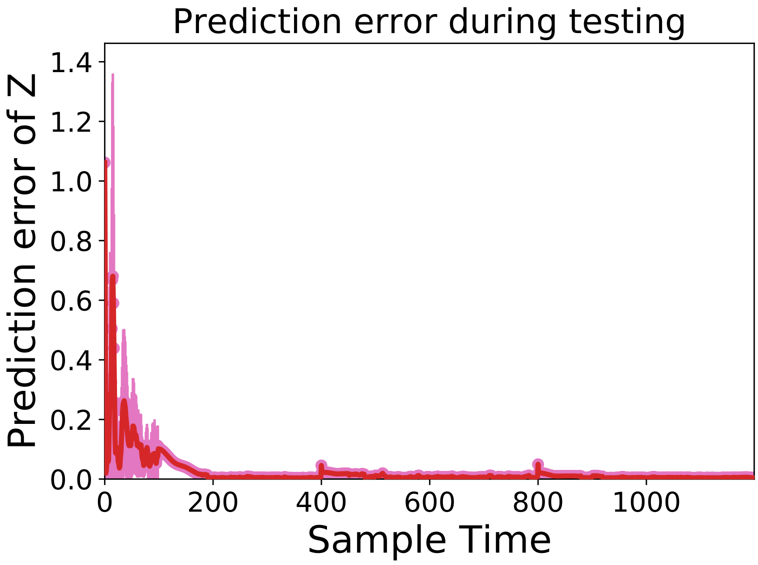


Figure 4.12: *Indicates the consequent prediction error during model's testing with experiment four (see Table 4.6).*

The mean inferred action values within every 400 time frames is illustrated by means of the heat-map-like confusion matrix. As can be seen from Figure 4.13, the entry in the first, second, and third row of the matrix is correct in the corresponding first, second, and third column, encoding walking, jumping jack, and dancing action types.

Figure 4.14 describes the results of experiment 2 in which the setting are analogous to experiment 1 except for the *Stimulus Presence* duration. Throughout the experiment 2, each action's input stimulus was provided for the first 100 frames only (open loop), switching to closed-loop for the remaining 300 frames while

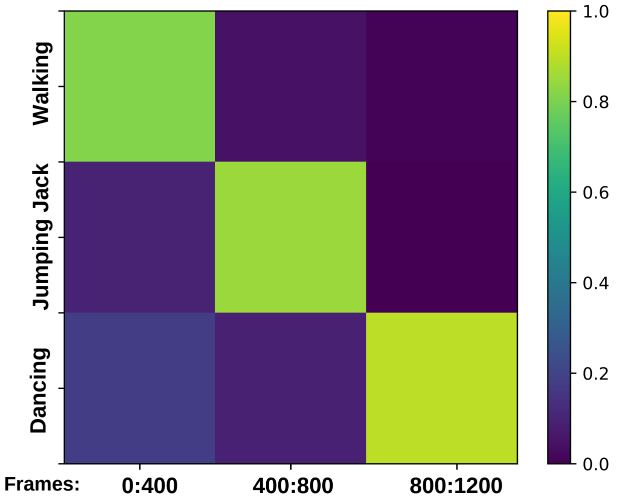


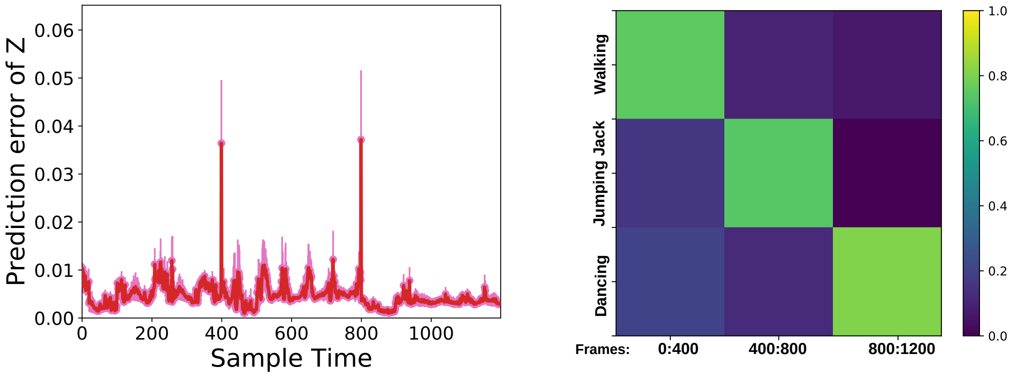
Figure 4.13: *Experiment 1; the model accurately infers the action types during testing. Based on the given stimulus the first 400 frames, the model predicts 1,0,0:walking as the occurring action type then 0,1,0:Jumping jacks followed by 0,0,1:Dancing.*

the input stimulus was no longer present.

As part **(b)** of Figure 4.14 indicates, the model is able to infer the actual behavior while it relies on its own imagination. As shown in part **(a)**, the two peaks appear when the data switches to the next action.

4.6.1 Feature Binding and Perspective Taking

Suitable perspective taking and feature binding adaptations facilitate Gestalt perception and behavior interpretation. Through experiments 4, 5, and 6, perspective taking and feature binding adaptation progress were examined relative to the temporally imagined code Z, after inferring the action type simultaneously during the first 100 time steps of each 400 step subsequence. Employed hyperparameters for feature binding ablation studies are according to given values for Exp. 3 of Table 4.3 and used architectural parameters for perspective taking are similar to Table 4.5. The Gestalt perception progress within these experiments is assessed by making use of Feature Binding Error (FBE), Translation



(a) *Test error*

(b) *Inferred action*

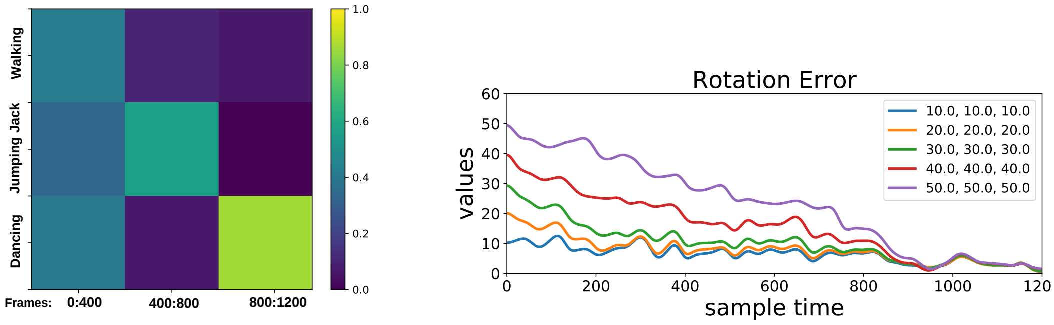
Figure 4.14: *Experiment 2: Evaluating the model's performance while the input stimulus is present for 100 time frames at the beginning of each observed stimulus.* ***(a)****Illustrates the consequent LSTM's prediction error of latent vector Z while* ***(b)*** *Shows the resulting inferred behavior.*

Difference (TD), and Orientation Difference (OD) as described in Sections 3.2 and 3.3.

According to five different initial viewpoints applied during experiment 3, the adaptation progress of perspective taking is illustrated in Figure 4.15 part **(b)**.

Spatial abilities are known to be iterative and incremental [107]. Hereof, part **(a)** of the above figure indicates the fact that the resulting inference becomes correctly identified once the relative orientation difference converges; attesting that the initial input data rotations was significantly associated with appropriate Gestalt perception, that is, the more the observed data is rotated away from a canonical perspective, the tougher the Gestalt perception (in terms of recognition time and performance). The above findings are similar to psychological results reported by Pavlova and Sokolov [85] as well as Shepard and Metzler[108] where they assessed the orientation specificity and mental rotation of human subjects.

As can be observed from Figure 4.16, adapting the feature binding and concurrently inferring the occurring action type (Experiment 4), as anticipated, makes it harder for the architecture to detect the actual behavior as long as the feature



(a) *Inferred action* **(b)** *Test error*

Figure 4.15: *Exp. 3: Rotation adaptation progress while inferring the actual observed intention **(a)** Indicates the mean inferred behavior over all five rotation trials. **(b)** Depicts the relative orientation difference progress which is being adapted toward the model's perceptual experience during training, where the initial input data rotations over each axis are demonstrated in the legend.*

binding error is still high.

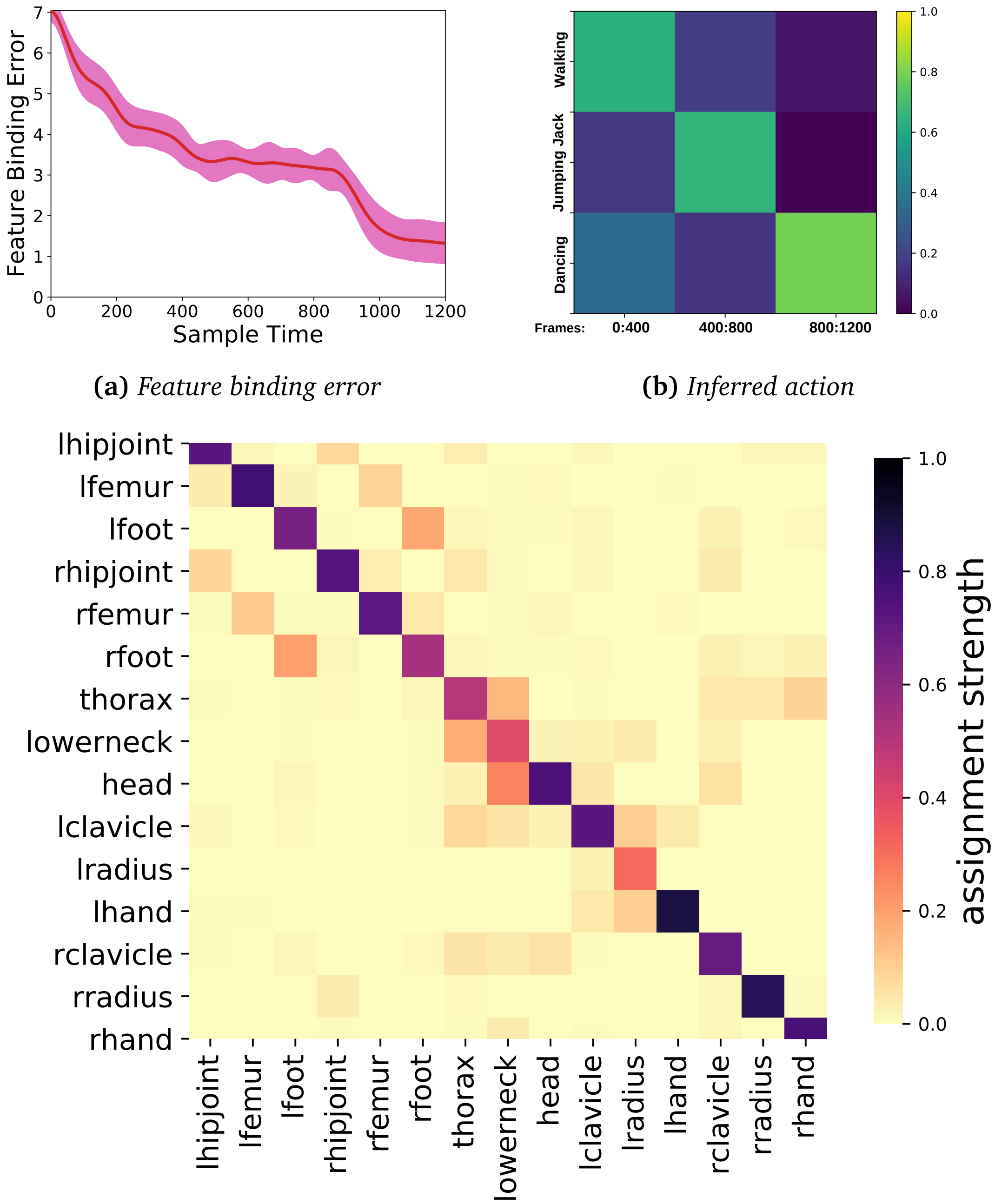
Nevertheless, the proposed model is still capable of inferring the correct feature binding assignments even when the bottom-up signal is given to the variational autoencoder (VAE) for only the first 100 time steps.

4.6.2 Ablation Studies

Here I attempt to explore the effectiveness of the population encodings and the redundant spatial input encodings throughout temporal prediction and Gestalt perception processes.

During experiment 5, the population encodings was disabled simply after training the VAE to learn suitable compressions in order to encode the corresponding scalar values. Within experiment 6, parts of the the modal input were removed, just the posture submodality was trained and processed while the population encoding was activated. Feature binding and Gestalt perception development was examined throughout both experiments.

Based on the results described in Figures 4.16 and 4.17, it is clear that during experiment 4, where population encodings was activated and information of all



(a) *Feature binding error*

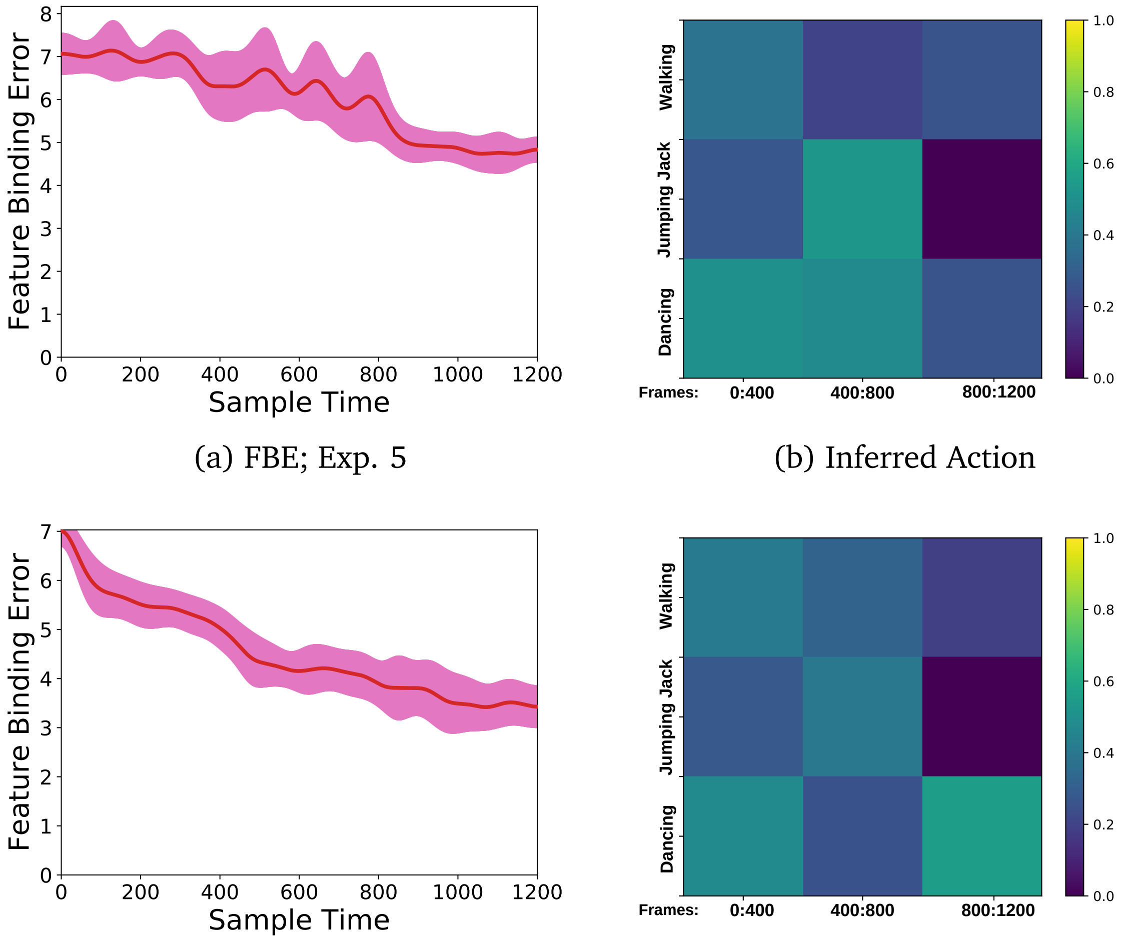
(b) *Inferred action*

(c) *Binding matrix assignment*

Figure 4.16: *Exp. 4:* ***(a)*** *Demonstrates the feature binding progress;* ***(b)*** *Indicates the corresponding inferred behavior, binning the estimates over the first, second, and third 400 steps.* ***(c)*** *Shows the assignment of observed features to the correct corresponding bodily features at the end of experiment.*

sub-modalities were fed into the variational autoencoder and subsequently to our retrospective inference procedure, we achieved better feature binding and behavior inference results.

Therefore, we can conclude that in order to infer the unfolding behavior employing the population encoding is extremely beneficial. Similarly, providing appropriate redundant information by means of separating the observed spa-



(a) FBE; Exp. 5

(b) Inferred Action

(c) FBE; Exp. 6

(d) Inferred Action

Figure 4.17: ***Upper-Row:*** *Experiment 5; population coding is deactivated.* ***Lower-Row:*** *Experiment 6; magnitude and direction sub-modalities are deactivated (that is, β^{pos}=8 , β^{mag}=0, β^{dir}=0).*

Table 4.7: *Used pendulum data for training and testing the model*

Motion-Type	Training	Testing
2D Pendulum	seven cycles 1000 frames	same amount of data (different cycles)

tial location data into directional, positional, and motion magnitude encodings is hugely advantageous for a more reliable action type inference. It should be noted, however, that none of the error measures demonstrated in this chapter are known to the model. Indeed, these losses are being converged by minimizing embodied expectation errors and by using of predictive encoding [21].

4.7 Further Investigations

4.7.1 2D Pendulum

In order to ensure the generality of our findings, the feature binding performance of the model was assessed based on the pendulum data which we adapted from animation example of mathplotlib [1]. The two joint pendulum had two joints with 1.25 and 1 kg mass, and 0.8 and 0.6 meter length, respectively.

The training and testing data basically encode the 2D-pendulum swinging back and forth for seven cycles with completely loose lower limb and are described in Table 4.7.

Architectural parameter values used for training LSTM and VAE modules are represented in Section 4.2 and in Table 4.2. Additionally, hyperparameters used for feature binding adaptation of 2D pendulum are given in Table 4.3.

Figure 4.18 verifies that even for two joint pendulum data feature binding works completely robust and effective. As shown by Figure 4.18, although some minor disruption exist (that is, some latent activity of Joint 2 is added to Joint 1) the model performs a robust Gestalt perception.

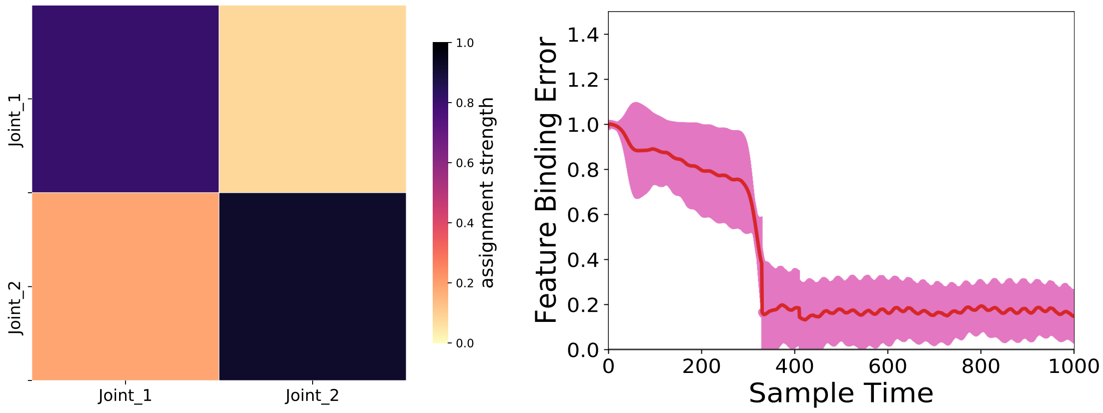


Figure 4.18: *Feature binding development and adaptation based on two jointed pendulum data*

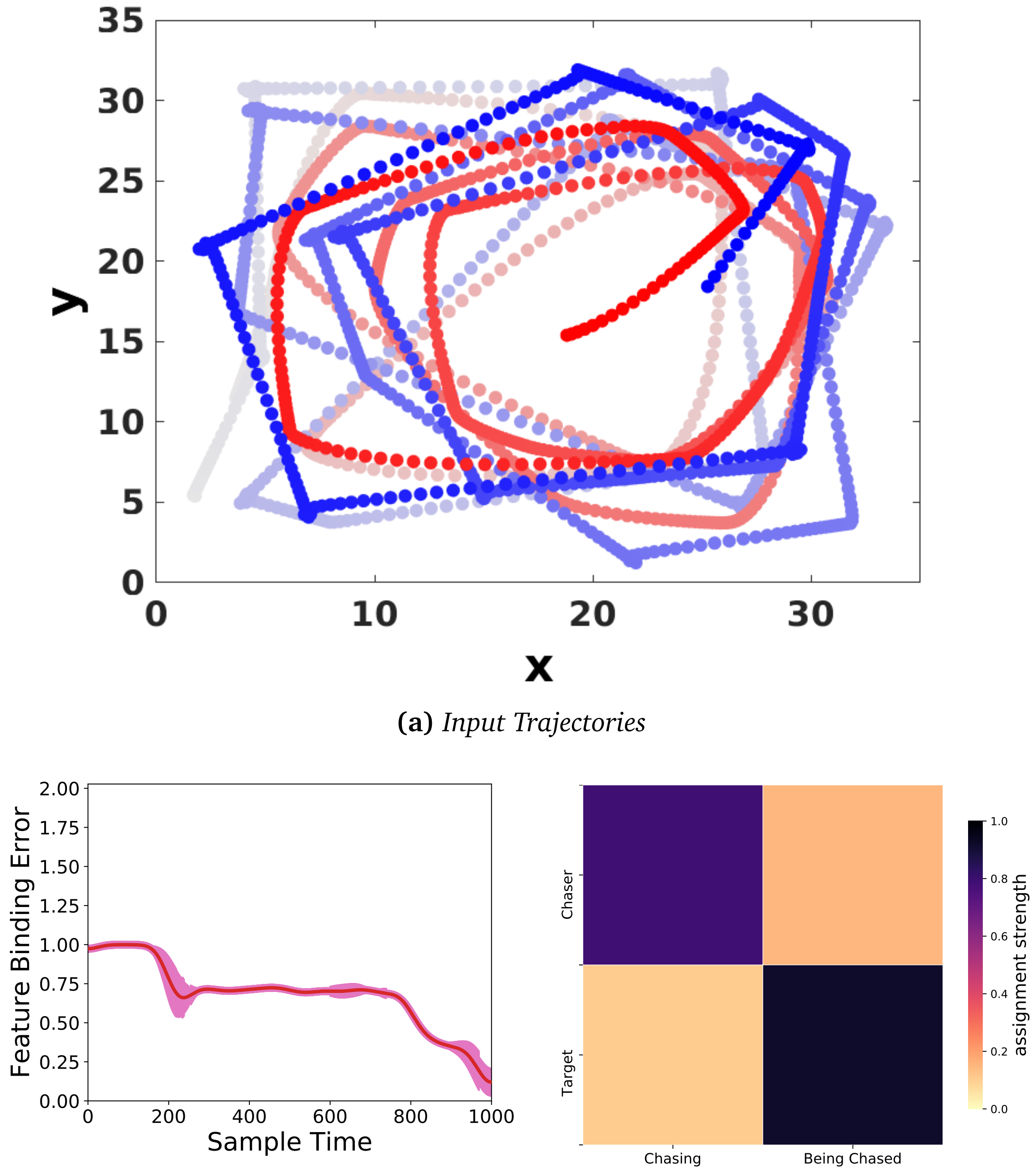
4.7.2 Chasing Scenario

We have additionally scrutinized our model's performance on other types of Gestalt perception, as well as social agent interactions (like a group of people interacting while pursuing a cooperative or competitive task).

For this purpose, we generated chasing agents' interactions data by making use of a generative model of social interactions introduced by Salatiello et al. [100].

During the chasing experiment one agent is continuously chasing another agent. An exemplary scenario of chasing agents' trajectories is illustrated in part (a) of Figure 4.19. Once the model has been sufficiently trained (1000 epochs and 10 differently initialized neural networks), we evaluate our binding approach by assigning each agent to its correct behavior. The averaged feature binding results over those 10 different trained neural network are shown in parts (b) and (c) of Figure 4.19 and as can be seen the model is capable of reducing the feature binding error over time by means of gradient based inference and establishing a correct binding between chaser and chasee.

Taken together, the results reported in this chapter confirm that the model is able to correctly bind features into Gestalt templates, and to infer the perspective of an actor robustly. Moreover, the model adequately ignores the perceived features that do not belong to those features seen during training which is similar to neuroscientific discoveries that assert, mirror neurons merely respond to



(a) *Input Trajectories*

(b) *Feature binding error of chasing*

(c) *Inferred action of chasing*

Figure 4.19: ***(a)*** *Shows the Trajectories of the two agents during an example chase scenario (Red:chaser, and Blue:chased);* ***(b)*** *Indicates the relative feature binding error.* ***(c)*** *Shows the assignment of observed behavior to the correct acting agents.*

perceived movements that are part of the motor repertoire of the observer. Additionally, even when the bottom-up input is not available or is not reliable the models' embodied imaginations are appropriate for perceptual inferences.

Furthermore, our approach may be suitably applied to other spatio-temporal interaction events, attention-demanding predictions, simulation, forecasting, and encoding problems [17, 19]. Additionally, an overview of some of the capabilities of the implemented architecture [3] is listed in Appendix C.

[3] The implemented code and its documentations is available at [4]
https://github.com/CognitiveModeling/Gestalt-Perception ; As of 24.7.2021

Chapter 5

Conclusion and Discussion

The main conclusions of this thesis are drawn together and presented in this chapter. This thesis primarily aimed to introduce a generative artificial neural network model which was based on the previous work from [102]. The inspiration behind the introduced architectural design came from predictive coding [34], active inference [18, 33, 82, 83], and brain's sub-symbolic information processing concepts [10, 17, 21, 35].

It is worth remarking here that the neocortex is not a feed-forward architecture [48]. Consequently, the proposed model in this thesis suggests an architecture that depends on the interaction of bottom-up saliency cues and top-down expectations in a Bayesian manner. Such interactions will result in inferring the respective adaptation parameters over time.

5.1 Observations and Assertions

The outcome of various experiments based on the motivations and hypotheses outlined in Section 2.8 lead to the following conclusions which will be explained individually.

- ***Retrospective Inference***

A major aspect of this research as opposed to other similar works reported in Chapter 2 was retrospective inference of latent states by means of stochastic gradient descent [95], which resulted in the model suitably coping with motion sequence data. The gradient-based, retrospective latent state inference approach [18, 110] fits best with the introduced model, given that in our employed rhythmic action pattern data both individual elements as well as their occurrence order transmit critical information regarding the perceived input stimuli. Most notably, the architecture consistently attempts to match the sensory input with the top-down expectations by making use of an adaptive retrospective inference procedure as discussed in sections 3.2 and 3.3. Such a procedure adapts parametric biases in the past while monitoring the unfolding sensorimotor dynamics.

- ***Binding, Perspective-Taking***

The *Perspective Taking* challenge arises when the model perceives the input stimuli from an unknown allocentric viewpoint and attempts to transfer it to its known self-centered perspective. This cognitive ability enables humans to mentally adopt their spatially observed perspective and subsequently recognize the corresponding movement patterns [60]. The *Feature Binding* problem [114] concerns selecting and integrating distinct visually observed features into their right order. Selectively routing the motion dynamics and assigning them to respective neural processing pathways in a way that they match the expected Gestalt dynamics can be considered as a way to resolve this problem.

As described in Chapter 3, the model uses the top-down reconstruction error that originates from the VAE to reduce the discrepancy between the observed and the predicted sensory signals [40]. This approach is in agreement with predictive coding theory [34] based on which the human brain continuously predicts the sensory causes forward in time. As presented in Chapter 3, for an adequate routing of the input features and the adaptation of the perceived perspective onto a known reference frame, the proposed modularized architecture propagated the prediction error back onto a binding matrix, and further back onto perspective taking neurons, which rotated and translated the observed input to a known perspective. Eventually, the model was able to establish a suitable correspondence between input features and bodily features in nearly all feature binding constellations (see Chapter 4). Furthermore, the transformation of the observed stimuli onto a known canonical perspective was successful within all investigated perspective taking experiments (see Sections 4.5 and 4.6.1).

- ***Population Coding, Redundant encodings***

In order to resolve perceptual conflicts and ambiguities within all experiments, the described population encoding approach was essential. Another critical component was separating the visually observed input into three distinct motion modalities (i.e. posture, magnitude, and motion direction) and weighting the corresponding submodal expectation losses by their respective adaptation factors. According to the conducted ablative experiments (described in Tables 4.3; 4.6) and their outcomes, population encoding and redundant submodal components provided constructive information so as to establish the correspondence between the perceived and the known frame of references, as well as to bind the observed features into Gestalt templates.

- ***Spatial Reconstruction***

For learning the compressed formats of the visually observed stimulus a VAE module was used, modified from [102], who used a different autoencoder implementation. Instead of learning a function that merely represents the data, this module made the architecture capable of learning the parameters of a probability distribution that represented the data (i.e. variance and mean of the latent space variables).

The employed VAE does not have the standard AE's drawbacks (e.g. standard AE's simply perform a non-linear extension of PCA; their utilized latent space tends to be not continuous and may result in a harder interpolation while generating biased data).

The major problem with autoencoders, for generation, is that the latent space they transform their inputs to and where their encoded vectors lie, may not be continuous or allow easy interpolation.

The reported results in Section 4.2 demonstrate that the VAE was able to foster the development of more balanced latent state encodings and to perform a proper sampling and data generation that supported the architecture in solving the perspective taking, feature binding and Gestalt perception challenges.

- **Temporal Prediction**

A temporal predictive processing module was used to learn temporal correlations of the compressed spatial codes. Based on the reported results in Section 4.6, this module was successfully providing the temporal code imagination whenever the bottom-up input was missing. Additionally, it helped the architecture to store the previous time-steps' information and subsequently tackle the vanishing gradient problem.

- **Behavior Interpretation**

The introduced LSTM-based temporal prediction procedure helped the model to interpret and understand the unfolding behavior. Compared to [102], the intention interpretation and Gestalt Perception procedures were significantly faster as they had lower number of trainable weights. Moreover, the results in Section 4.6 indicate that the model was able to correctly infer various action types.

Once the input stimulus is missing or considered unreliable, the model switches to closed loop processing and effectively relies on its own embodied simulation for performing a successful behaviour interpretation (see the corresponding results of experiments 2 to 6 listed in Table 4.6).

- **Successful binding despite distractors**

The results reported in Section 4.4 show that the employed adaptive mechanism was capable of establishing a robust and correct routing from observed features to their respective body features, effectively ignored other dynamics and distractors that did not belong to the perceived Gestalt during training. It should be noted here that the architecture's performance was more robust to static distractors than to biological and rhythmic distractors, which is in agreement with previous psychological study findings[85].

- **Generalisation**

As compared to [102], the proposed model is able to switch between 2D and 3D visual inputs. Moreover, it automatically distributes the topological Gaussian neurons around the observed stimulus. The experimentation results shown in Section 4.7 further support this claim.

- **Other Investigations**

The model's performance was further scrutinized on related Gestalt perception and recognition problems, as well as on social agent interaction patterns. In

Section 4.7.2, we exemplary investigated the binding performance of the model while detecting the intention of chasing agents and, as the results indicate, the model was able to select and assign the corresponding behavior to their respective actors.

Other exploratory experiments in Sections 4.5 and 4.6.1 show that the more the visual input is rotated away from a known perspective, the more difficult and delayed the Gestalt perception and recognition. This is analogous to the psychometric study results reported by Pavlova and Sokolov [85].

5.2 A Transformation Mechanism

Despite the success of Convolutional Neural Networks (CNN) in certain learning frameworks [66], they still suffer from not being spatially invariant to the input data [67]. Jaderberg et al. [56] presented a Spatial Transformer (ST) architecture, which can be incorporated into various neural network models in order to provide spatial transformation abilities. Based upon the previous work by Jaderberg et al. [56], I hypothesize an affine transformation mechanism that utilizes an adaptation technique similar to the described perspective taking procedure in Section 3.3, and subsequently resolves the limitations of the previous feed-forward architecture by computing the gradients efficiently (as outlined in Section 2.6). For this purpose, parametric bias neurons' activities will be adapted retrospectively by making use of back-propagation through time.

Networks that employ ST are able to select the most relevant parts of a distorted input image (i.e. attention), as well as to transform those selected parts to a canonical point of view [56]. The Spatial Transformer's core architecture is depicted in the Figure 5.1 [56].

Based on the retrospective inference adaptation procedures described in Section 3.3, I propose an affine transformation mechanism in which we aim at using a similar perspective taking process inside the localisation network of the existing Spatial Transformer's framework. Figure 5.2 indicates the same procedure in a connectivity graph.

The localisation network within the spatial transformers receives the input feature map U and outputs the transformation parameter θ. Figure 5.2 illus-

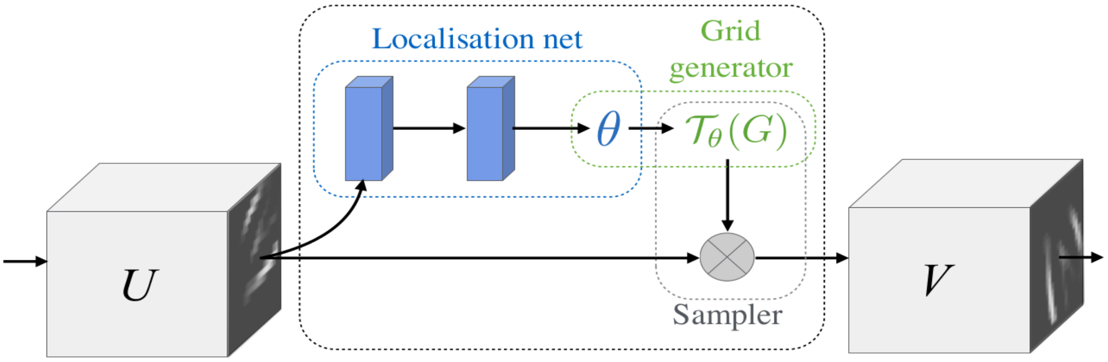


Figure 5.1: *A Spatial transformer consists of three major components;* ***1)*** *A localisation network which computes transformation parameters θ in accordance with the input feature map U,* ***2)*** *A Grid generator, which uses the predicted transformation parameters ($\mathcal{T}_\theta(G)$ denotes the spatially generated Grid over V), and* ***3)*** *A sampling mechanism that draws samples from the generated Grid and produces the output feature map V.*

trates an alternative localisation network which takes the image's feature map $U \in \mathbb{R}^{H*W}$ with height H and width W as input and outputs the inferred translation and rotation activities. As stated by Jaderberg et al. [56], the parameters of the affine transformation can eventually be used as:

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \tau_\theta\left(G_i\right) = A_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} r_{11} & r_{12} & b_1 \\ r_{21} & r_{22} & b_2 \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{5.1}$$

where r and b elements are the respective parametric bias neurons' activities for rotation matrix and translation vector; $\mathcal{T}_\theta$(G) denotes the spatially generated Grid; A_θ stands for the affine transformation of the corresponding perspective taking procedure; (x_i^s, y_i^s) is the source coordinate, and (x_i^t, y_i^t) yield the target coordinate in the output feature map.

As shown in Figure 5.2, every black pixel may be separately encoded with one population of postural neurons with Gaussian turnings as previously explained in Section 3.1. Afterwards, the concatenation of all population encoded activities across all black pixels would be forwarded as input to the variational autoencoder. VAE would then learn the spatial representation of input features in a compressed and generative format.

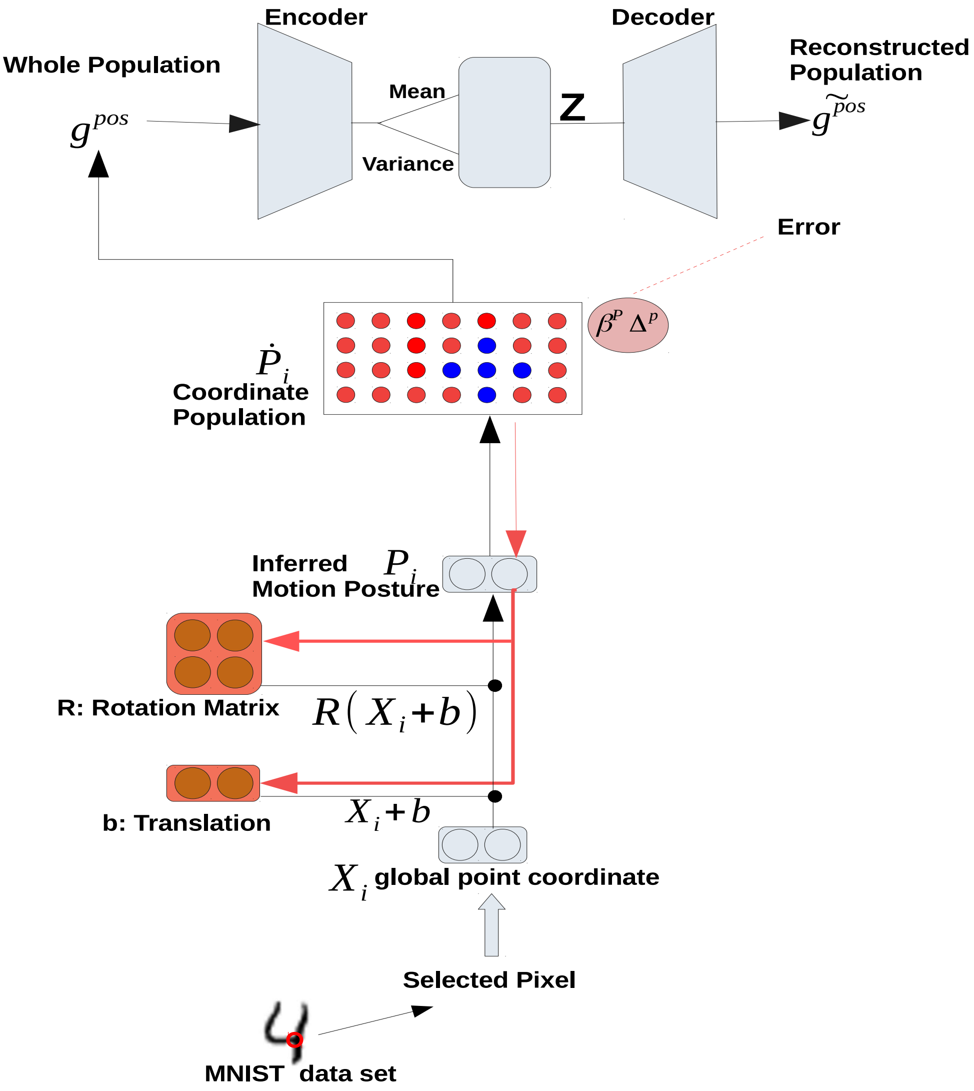


Figure 5.2: *The proposed localisation network; during training the architecture will observe a visual input and finds out the respective encoded population of Gaussian neurons for all candidate pixels. Then, a VAE will perceive the concatenation of population-encoded activities for all black pixels and will learn their corresponding compressed formats. During testing, the top-down reconstruction error that originates from the VAE helps the localisation network to transform the distorted digit's image back onto its canonical perspective.*

As a consequence, when during testing the architecture observes an input digit that is rotated and translated away from a canonical perspective, VAE generates biased reconstruction error of the input population. Subsequently, the model may employ the obtained top-down reconstruction error in order to provide an adequate canonical representation of the input stimulus, effectively transferring it back onto its known, self-centered perspective. Moreover, the above described mechanism may be extended to 3D or even higher dimensional transformations.

5.3 Future Prospects and Summary

As outlined in Chapter 4, foreseen applications of the proposed architecture are not only in the area of biological motion interpretation but also in other domains such as 2D motion recognition tasks (see Section 4.7), as well as interpreting multiple agents that are interacting with each other (see Section 4.7.2).

Furthermore, given that binding and perspective taking are global problems in cognitive science, the suggested retrospective inference adaptation mechanisms can potentially be used as an efficient alternative for solving related challenges. For instance, in Section 5.2 we proposed a mechanism that employs a transformation procedure similar to our perspective taking approach. Moreover, the obtained feature binding results justify further usage scenarios such as cases where information needs to be robustly bound together and assigned to other on the fly data, for accomplishing overall consistency.

Nevertheless, in follow-up research, several interesting aspects of the proposed model may be explored further by evaluating other aspects of human cognition and making use of other available databases like the Facial Expression Database [84] and the Emotional Body Motion Database [3, 119], which could potentially give us the ability to explore human facial and emotional expressions while performing an action (i.e. happiness while acting jumping jacks; laughing while dancing and so on).

Future studies seeking to use and build upon the methodology introduced in this thesis can attempt to further explore the proposed model's effectiveness while performing the following investigations:

- Understanding the behavior of social actors within complicated scenarios such as a group of people interacting with each other while pursuing a competitive or cooperative task.
- As a robust inference mechanism for attention-based architectures [118].

- Improving the inference of modern object representation learning models (like the framework introduced by Greff et al.[44]) while observing and interpreting even complex visual scenes.
- As an intermediate visualization stage (for providing a better representation of the observed scene) within an interactive Virtual Reality (VR) environment.
- Neuroimaging studies that aim at exploring the activation of visual cortex and mirror neurons during the involvement of feature-binding and perspective-taking tasks.

Summary

Action understanding is a challenging task due to the complexity of human movements and behavior. Human cognition has an adaptive nature [7] during which top-down expectations interact with bottom-up saliency cues [48]. Inspired by the underlying principles of action understanding in the brain, this thesis and its methodological approach contributed to the debates on fundamental human cognitive challenges which are Gestalt perception, feature binding and perspective taking. The suggested model was capable of learning compressed dynamic biological motion or other rhythmic motion patterns according to its self perceptual experiences. Initially, the model was trained to learn sufficiently accurate generative models of the observed action patterns. After training, the model propagated the prediction error back onto a binding matrix, and further back onto perspective taking neurons. During this process the model was pursing the goal to minimize reconstruction errors over time. Consequently, this adaptation mechanism resulted in correctly routing the input features to their respective bodily features and also adapting the internal perspective onto a known reference frame, while properly interpreting the actual observed motion dynamics. Moreover, the employed gradient-based, retrospective inference could potentially be used within attention-based architectures. Additionally, ablation studies revealed that the population encodings and complementary spatial encodings enhanced the model's performance. Eventually, the introduced architecture aims at providing insights for improving the behavior of intelligent agents.

Appendices

Appendix A

Multivariate Gaussians

As previously outlined in Section 3.1, each individual submodal feature is encoded by one population topological Gaussian neurons. Here I provide some details of the employed encoding approach.

A.1 Basics

The density of a multivariate normal distribution [112] at x with mean m and variance Σ (i.e. $x \sim (m, \Sigma)$), is calculated by:

$$p(x) = \frac{1}{\sqrt{det(2\pi\Sigma)}} exp\left[-\tfrac{1}{2}(x-m)^T\Sigma^{-1}(x-m)\right] \tag{A.1}$$

It should be noted, however, that for a d-dimensional x we have [87]

$$det(2\pi\Sigma) = (2\pi)^d det(\Sigma) \tag{A.2}$$

A.2 Derivation

The first order derivative of the aforementioned density yields [87]:

$$\frac{\partial p(x)}{\partial x} = -p(x)\Sigma^{-1}(x-m) \tag{A.3}$$

Appendix B

Rotation Matrices

As described in Section 3.3, for adapting the visual frame of reference we employed neural rotation matrix R as formulated in equation 3.20. Here we use the notation $s_r = sin(\Theta_r)$ and $c_r = cos(\Theta_r)$ where $r \in (x, y, z)$.

B.1 Basics

The resulting rotation matrix R will be:

$$\begin{bmatrix} e_{0,0} & e_{0,1} & e_{0,2} \\ e_{1,0} & e_{1,1} & e_{1,2} \\ e_{2,0} & e_{2,1} & e_{2,2} \end{bmatrix} = \begin{bmatrix} c_y c_z & -c_y s_z & s_y \\ c_z s_x s_y + c_x s_z & c_x c_z - s_x s_y s_z & -c_y s_x \\ -c_x c_z s_y + s_x s_z & c_z s_x + c_x s_y s_z & c_x c_y \end{bmatrix} \tag{B.1}$$

The respective derivative for each rotation angle is then computed by $\frac{\partial R}{\partial x}$, $\frac{\partial R}{\partial y}$, and $\frac{\partial R}{\partial z}$.

B.2 Rotation Matrix to Euler Angles

In order to determine the resulting Euler angles based on the achieved rotation matrix [27], at each time step we consider three cases: 1. $\Theta_y \in (-\frac{\pi}{2}, \frac{\pi}{2})$; and as a result, we have $c_y \neq 0$ and subsequently, Euler angles are:

$$\Theta_x = atan2(-e_{12}, e_{22}), \Theta_y = asin(e_{02}), \Theta_z = atan2(-e_{01}, e_{00}) \tag{B.2}$$

2. $\Theta_y = \frac{\pi}{2}$; and therefore, $s_y = 1$ and $c_y = 0$ which will result in:

$$\Theta_x + \Theta_z = atan2(e_{10}, e_{11}) \tag{B.3}$$

3. $\Theta_y = -\frac{\pi}{2}$; which yields $s_y = -1$ and $c_y = 0$

$$\Theta_z - \Theta_x = atan2(e_{10}, e_{11}) \tag{B.4}$$

The Algorithm 1 below, indicates the resulting pseudo-code.

Algorithm 1: Rotation Matrix to Euler Angles

```
Input  : Rotation Matrix R
Output: $\theta_x, \theta_y, \theta_z$
$R = \begin{bmatrix} e_{0,0} & e_{0,1} & e_{0,2} \\ e_{1,0} & e_{1,1} & e_{1,2} \\ e_{2,0} & e_{2,1} & e_{2,2} \end{bmatrix}$
/* The rotation Matrix is available at each time step          */
if $e_{02} < +1$ then
    if $e_{02} > -1$ then
        $\theta_x \leftarrow atan2(-e_{12}, e_{22})$
        $\theta_y \leftarrow asin(e_{02})$
        $\theta_z \leftarrow atan2(-e_{01}, e_{00})$
    end if
    else
        // In this case $e_{02} = -1$
        $\theta_x \leftarrow -atan2(e_{10}, e_{11})$
        $\theta_y \leftarrow -\frac{\pi}{2}$
        $\theta_z \leftarrow 0$
    end if
else
    // In this case $e_{02} = +1$
    $\theta_x \leftarrow atan2(e_{10}, e_{11})$
    $\theta_y \leftarrow \frac{\pi}{2}$
    $\theta_z \leftarrow 0$
end if
return $\theta_x, \theta_y, \theta_z$
```

Appendix C

Implementation Details

Following Table C.1 lists all functionalities within the implemented framework.

Table C.1: *Implemented functionalities*

Flag name	Functionality
-Train	To activate the training mode.
-Test	To activate the testing mode.
-FB	To enable feature binding during testing.
-PT	To enable perspective taking during testing.
-VAE	To activate the variational autoencoder.
-N	Indicates the number of training or testing sample frames.
-Shuffle	Randomly shuffles the input sequence.
-Video	To create output videos.
-F	Starting frame for video creation.
-E	Ending frame while creating the video.
-i	Input data file; if 3D: .amc orientations, if 2D: .txt coordinates.
-j	Input skeleton; .asf file; just in 3D input cases.
-c	Indicates the widths and heights of Gaussian neurons.
-NoPopCode	To disable the population encoding of input sequences.
-2D	To enable the 2-dimensional pendulum case.
-2DAnim	To enable the 2-dimensional two-agents interaction case.
-Stim_Reliability	Denotes the duration when the input stimulus is still available; after this time frame the model relies on its own embodied imagination (that is, closed loop).
-Intention	To create video during intention inference processes.

Here I summarize some of the command lines and their respective use cases:

- **Command to receive help and see all flag's functionalities**

```
python3.7 -m Population_Coding.py -h
```

- **Basic training of the model**

```
python3.7 -m Population_Coding.py -Train -VAE -N130 -iS35T07.amc -jS35T07.asf
```

- **Test with Variational Autoencoder and adapt the feature binding**

```
python3.7 -m Population_Coding.py -Test -VAE -FB -N500 -iS35T07.amc -jS35T07.asf
```

- **Test with Variational Autoencoder and adapt the perspective taking**

```
python3.7 -m Population_Coding.py -Test -VAE -PT -N500 -iS35T07.amc -jS35T07.asf
```

- **Train the model and create variational autoencoder's reconstructed videos**

```
python3.7 -m Population_Coding.py -Train -VAE -N135 -iS35T07.amc -jS35T07.asf -Video -F28 -E110
```

- **Custom adjustments for feature binding**.
The corresponding elements of FBGridList are: [Learning rate, Momentum Rate, posture adaptation factor, direction adaptation factor, magnitude adaptation factor, all initial values for binding matrix elements]; This is specially useful while doing the grid-search for hyperparameter tunings.

```
python3.7 -m Population_Coding.py -Test -VAE -FB -N785 -iS35T07.amc -jS35T07.asf -FBGrid
-FBGridList=[1,0.9,20,0.1,0.5,-1]
```

- **Custom adjustments for perspective taking**.
The corresponding elements of PTGridList are: [data bias(x), data bias(y), data bias(z), translation learning rate, translation momentum, data rotation(x), data rotation(y), data rotation(z), rotation learning rate, rotation momentum, posture adaptation factor, direction adaptation factor].

```
python3.7 -m Population_Coding.py -Test -PT -N584 -iS35T07.amc -jS35T07.asf -VAE -PTGrid
-PTGridList=[-2,-2,-2,0.01,0.9,15,10,5,0.01,0.9,15,1]
```

- **Custom adjustments for training the variational autoencoder**.
corresponding elements of VAETrainGridList are: [direction learning rate, posture learning rate, magnitude learning rate, latent space size, hidden layer size].

```
python3.7 -m Population_Coding.py -Train -VAE -N145 -iS35T07.amc -jS35T07.asf -VAETrainGrid
-VAETrainGridList=[5e-3,5e-4,1e-4,50,200]
```

- **Training the variational autoencoder while deactivating the population coding**.

```
python3.7 -m Population_Coding.py -Train -VAE -N155 -iS35T07.amc -jS35T07.asf -VAE -NoPopCode
-VAETrainGrid -VAETrainGridList=[2e-5,1e-3,4e-3,10,32]
```

- **Testing the feature binding on the trained variational autoencoder while disabling the population coding**.

```
python3.7 -m Population_Coding.py -Test -VAE -N600 -iS35T07.amc -jS35T07.asf -NoPopCode -FB
```

- **Testing the perspective taking on the trained variational autoencoder while deactivating the population coding**.

```
python3.7 -m Population_Coding.py -Test -VAE -N600 -iS35T07.amc -jS35T07.asf -NoPopCode -PT
-PTGrid -PTGridList=[3,3,3,0.02,0.4,0,0,0,0,0,1,0]
```

- **Creating a video of the binding progress and development over time**.

```
python3.7 -m Population_Coding.py -Test -VAE -N255 -iS35T07.amc -jS35T07.asf -FB -Video -F50
-E100
```

- **Performing the training on 2D two jointed pendulum data**.

```
python3.7 -m Population_Coding.py -Train -VAE -2D -N135 -i2D_Pendulum.txt -VAETrainGrid
-VAETrainGridList=[8e-4,1e-3,5e-3,50,100]
```

- **Testing the feature binding on the trained two jointed pendulum**.

```
python3.7 -m Population_Coding.py -Test -VAE -2D -FB -N1000 -i2Dtest_set.txt -FBGrid
-FBGridList=[0.01,0.15,0.05,20,20,-1]
```

- **Testing the feature binding on the trained 2D Pendulum and creating a video from confusion matrix progress over time**.

```
python3.7 -m Population_Coding.py -Test -VAE -2D -FB -N100 -i2Dtest_set.txt -Video -F0 -E100
```

- **Testing the temporal imagination ability of the model**.
When after some time steps (i.e. -Stim_Reliability=200) the bottom up stimulus information is missing and the model is not receiving the input data. Then, the model has to rely on its own imagination biased from its trained temporal autoencoder (which uses an LSTM-based architecture for its prediction purposes).

```
python3.7 -m Population_Coding.py -Test -VAE -N1000 -iS35T07.amc -jS35T07.asf -VAETrainGrid
-VAETrainGridList=[4e-3,7e-4,8e-4,26,36] -Stim_Reliability=200
```

- **Testing the imagination, feature binding, and the perspective taking capabilities of the model**.
The stimulus was available for the first 200 time steps and from then on the model employs the trained LSTM module to come up with an imagined latent space to be able to implement efficient perspective taking and feature binding tasks.

```
python3.7 -m Population_Coding.py -Test -VAE -N1000 -iS35T07.amc -jS35T07.asf -Stim_Reliability=200
-FB -FBGrid -FBGridList=[1,0.9,20,0.1,0.5,-1]
```

```
python3.7 -m Population_Coding.py -Test -VAE -N1000 -iS35T07.amc -jS35T07.asf -PT -PTGrid
-Stim_Reliability=100 -PTGridList=[0,5,0,0.8,0.9,0,60,0,0.08,0.9,8,1]
```

- **Training and testing the model on some other non-cyclic/non-standard types of data**.
For instance on the dataset introduced by Salatiello et. al. [100] in which we have two agents interacting with each other.

```
python3.7 -m Population_Coding.py -Train -VAE -2D -N800 -iCH02_Tra_RR.csv -jCH02_Tra_BC.csv -2DAnim
-VAETrainGrid -VAETrainGridList=[8e-4,1e-3,5e-3,50,100]
```

```
python3.7 -m Population_Coding.py -Test -VAE -2D -N1000 -iCH02_Tra_RR.csv -jCH02_Tra_BC.csv -2DAnim
-FB -FBGrid -FBGridList=[0.1,0.9,10,1,0.5,-2]
```

- **Testing the model's behavior interpretation capability while visualising the development of inferred behaviour or intention**.

```
python3.7 -m Population_Coding.py -Test -VAE -N400 -Stim_Reliability=250 -iS35T07.amc -jS35T07.asf
-Intention
```

Bibliography

[1] Mathplotlib animation examples; double pendulum release: 2.0.2. https://matplotlib.org/examples/animation/, May 2017.

[2] Carnegie mellon university motion capture database. http://mocap.cs.cmu.edu, 2018.

[3] Emotional body motion database. http://ebmdb.tuebingen.mpg.de/index.php, 2021.

[4] Implemented architecture, framework. https://github.com/CognitiveModeling/Gestalt-Perception, 2021.

[5] Aris Alissandrakis, Chrystopher L Nehaniv, and Kerstin Dautenhahn. Imitation with alice: Learning to imitate corresponding actions across dissimilar embodiments. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 32(4):482–496, 2002.

[6] Shun-ichi Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological cybernetics*, 27(2):77–87, 1977.

[7] John R Anderson. Is human cognition adaptive? *Behavioral and Brain Sciences*, 14(3):471–485, 1991.

[8] Michael A Arbib. Mirror system activity for action and language is embedded in the integration of dorsal and ventral pathways. *Brain and language*, 112(1):12–24, 2010.

[9] L. W. Barsalou. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22:577–600, 1999.

[10] Lawrence W Barsalou. Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645, 2008.

[11] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *arXiv preprint 1806.01261*, 2018.

[12] Valentino Braitenberg. *Vehicles: Experiments in synthetic psychology*. MIT press, 1986.

[13] Marcel Brass and Cecilia Heyes. Imitation: is cognitive neuroscience solving the correspondence problem? *Trends in cognitive sciences*, 9(10):489–495, 2005.

[14] Charles Bruce, Robert Desimone, and Charles G Gross. Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of neurophysiology*, 46(2):369–384, 1981.

[15] Timothy J. Buschman and Earl K Miller. Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science*, 315(5820):1860–1862, 2007.

[16] Stephen A Butterfill and Corrado Sinigaglia. Intention and motor representation in purposive action. *Philosophy and Phenomenological Research*, 88(1):119–145, 2014.

[17] Martin V. Butz. Towards a unified sub-symbolic computational theory of cognition. *Frontiers in Psychology*, 7(925), 2016.

[18] Martin V. Butz, David Bilkey, Dania Humaidan, Alistair Knott, and Sebastian Otte. Learning, planning, and control in a monolithic neural event inference architecture. *Neural Networks*, 117:135–144, 2019.

[19] Martin V. Butz and Esther F. Kutter. *How the Mind Comes Into Being: Introducing Cognitive Science from a Functional and Computational Perspective*. Oxford University Press, Oxford, UK, 2017.

[20] Malinda Carpenter, Katherine Nagell, Michael Tomasello, George Butterworth, and Chris Moore. Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the society for research in child development*, pages i–174, 1998.

[21] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204, 2013.

[22] Richard Cook, Geoffrey Bird, Caroline Catmur, Clare Press, and Cecilia Heyes. Mirror neurons: From origin to function. *Behavioral and Brain Sciences*, 37:177–192, 2014.

[23] Antonia F de C. Hamilton and Scott T Grafton. Action outcomes are represented in human inferior frontoparietal cortex. *Cerebral Cortex*, 18(5):1160–1168, 2008.

[24] Sophie Denève and Alexandre Pouget. Bayesian multisensory integration and cross-modal spatial links. *Journal of Physiology - Paris*, 98:249–258, 2004.

[25] Giuseppe Di Pellegrino, Luciano Fadiga, Leonardo Fogassi, Vittorio Gallese, and Giacomo Rizzolatti. Understanding motor events: a neurophysiological study. *Experimental brain research*, 91(1):176–180, 1992.

[26] Paul E Downing, Yuhong Jiang, Miles Shuman, and Nancy Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473, 2001.

[27] David Eberly. Euler angle formulas. *Geometric Tools, LLC, Technical Report*, pages 1–18, 2008.

[28] Martin G Edwards, Glyn W Humphreys, and Umberto Castiello. Motor facilitation following action observation: A behavioural study in prehensile action. *Brain and Cognition*, 53(3):495–502, 2003.

[29] Birgit Elsner. Infants' imitation of goal-directed actions: The role of movements and action effects. *Acta psychologica*, 124(1):44–59, 2007.

[30] Wolfram Erlhagen and Estela Bicho. The dynamic neural field approach to cognitive robotics. *Journal of neural engineering*, 3(3):R36, 2006.

[31] Wolfram Erlhagen and Gregor Schöner. Dynamic field theory of movement preparation. *Psychological review*, 109(3):545, 2002.

[32] PO Esteves, LAS Oliveira, AA Nogueira-Campos, G Saunier, Thierry Pozzo, JM Oliveira, EC Rodrigues, E Volchan, and CD Vargas. Motor planning of goal-directed action is tuned by the emotional valence of the stimulus: a kinematic study. *Scientific reports*, 6(1):1–7, 2016.

[33] Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, Giovanni Pezzulo, et al. Active inference and learning. *Neuroscience & Biobehavioral Reviews*, 68:862–879, 2016.

[34] Karl Friston and Stefan Kiebel. Predictive coding under the free-energy principle. *Philosophical transactions of the Royal Society B: Biological sciences*, 364(1521):1211–1221, 2009.

[35] Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. *Journal of physiology-Paris*, 100(1-3):70–87, 2006.

[36] Qi Fu, Shiwei Ma, Lina Liu, and Jinjin Liu. Human action recognition based on sparse lstm auto-encoder and improved 3d cnn. In *2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pages 197–201. IEEE, 2018.

[37] Vittorio Gallese, Luciano Fadiga, Leonardo Fogassi, and Giacomo Rizzolatti. Action recognition in the premotor cortex. *Brain*, 119(2):593–609, 1996.

[38] Vittorio Gallese and Alvin Goldman. Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, 2(12):493–501, 1998.

[39] Vittorio Gallese, Christian Keysers, and Giacomo Rizzolatti. A unifying view of the basis of social cognition. *Trends in Cognitive Sciences*, 8(9):396–403, 2004.

[40] Paweł Gładziejewski. Predictive coding and representationalism. *Synthese*, 193(2):559–582, 2016.

[41] Daniel Goleman and Friedrich Griese. *Emotionale intelligenz*. Hanser München, 1996.

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

[43] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[44] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Daniel Watters, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *International Conference on Machine Learning*, pages 2424–2433. PMLR, 2019.

[45] Emily Grossman, Michael Donnelly, R Price, D Pickens, V Morgan, G Neighbor, and Randolph Blake. Brain areas involved in perception of biological motion. *Journal of cognitive neuroscience*, 12(5):711–720, 2000.

[46] George W Hartmann. Gestalt psychology: A survey of facts and principles. 1935.

[47] Dongliang He, Zhichao Zhou, Chuang Gan, Fu Li, Xiao Liu, Yandong Li, Limin Wang, and Shilei Wen. Stnet: Local and global spatial-temporal modeling for action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8401–8408, 2019.

[48] David J Heeger. Theory of cortical function. *Proceedings of the National Academy of Sciences*, 114(8):1773–1782, 2017.

[49] Mary Hegarty and David Waller. A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, 32(2):175–191, 2004.

[50] Christoph S Herrmann and Volker Bosch. Gestalt perception modulates early visual processing. *Neuroreport*, 12(5):901–904, 2001.

[51] Cecilia Heyes. Causes and consequences of imitation. *Trends in cognitive sciences*, 5(6):253–261, 2001.

[52] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

[53] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[54] Marco Iacoboni and Mirella Dapretto. The mirror neuron system and the consequences of its dysfunction. *Nature Reviews Neuroscience*, 7(12):942–951, 2006.

[55] Pierre Jacob and Marc Jeannerod. The motor theory of social cognition: a critique. *Trends in cognitive sciences*, 9(1):21–25, 2005.

[56] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. *Advances in Neural Information Processing Systems*, (28):2017–2025, 2015.

[57] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[58] Matthew Johnson and Yiannis Demiris. Perceptual perspective taking and action recognition. *International Journal of Advanced Robotic Systems*, 2(4):301–309, 2005.

[59] Frank Jäkel, Manish Singh, Felix A. Wichmann, and Michael H. Herzog. An overview of quantitative approaches in gestalt perception. *Quantitative Approaches in Gestalt Perception Vision Research*, 126:3–8, 2016.

[60] Klaus Kessler and Lindsey Anne Thomson. The embodied nature of spatial perspective taking: embodied transformation versus sensorimotor interference. *Cognition*, 114(1):72–88, 2010.

[61] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[62] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[63] Ricardo Kleinlein, Álvaro García-Faura, Cristina Luna Jiménez, Juan Manuel Montero, Fernando Díaz-de María, and Fernando Fernández-Martínez. Predicting image aesthetics for intelligent tourism information systems. *Electronics*, 8(6):671, 2019.

[64] Kurt Koffka. *Principles of Gestalt psychology*, volume 44. Routledge, 2013.

[65] Georg Layher, Tobias Brosch, and Heiko Neumann. Real-time biologically inspired action recognition from key poses using a neuromorphic architecture. *Frontiers in neurorobotics*, 11:13, 2017.

[66] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[67] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 991–999, 2015.

[68] Miao Ma, Naresh Marturi, Yibin Li, Ales Leonardis, and Rustam Stolkin. Region-sequence based six-stream cnn features for general and fine-grained human action recognition in videos. *Pattern Recognition*, 76:506–521, 2018.

[69] Valère Martin, Hendrik Reimann, and Gregor Schöner. A process account of the uncontrolled manifold structure of joint space variance in pointing movements. *Biological Cybernetics*, 113(3):293–307, 2019.

[70] Valère Martin, John P Scholz, and Gregor Schöner. Redundancy, self-motion, and motor control. *Neural Computation*, 21(5):1371–1414, 2009.

[71] A. N. Meltzoff and W. Prinz. *The imitative mind: Development, evolution and brain bases*. Cambridge University Press, Cambridge, 2002.

[72] Andrew N Meltzoff. Infant imitation after a 1-week delay: long-term memory for novel acts and multiple stimuli. *Developmental psychology*, 24(4):470, 1988.

[73] R. Memisevic. Learning to relate images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1829–1846, Aug 2013.

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

[75] Chrystopher L Nehaniv, Kerstin Dautenhahn, et al. The correspondence problem. *Imitation in animals and artifacts*, 41, 2002.

[76] Tam V Nguyen and Bilal Mirza. Dual-layer kernel extreme learning machine for action recognition. *Neurocomputing*, 260:123–130, 2017.

[77] Henry Friday Nweke, Ying Wah Teh, Mohammed Ali Al-Garadi, and Uzoma Rita Alo. Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. *Expert Systems with Applications*, 105:233–261, 2018.

[78] Chris Olah and Shan Carter. Attention and augmented recurrent neural networks. *Distill*, 1(9):e1, 2016.

[79] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: a strategy employed by v1? *Vision Res*, 37:3311–25, 1997.

[80] Stefan Oniga and Jozsef Suto. Activity recognition in adaptive assistive systems using artificial neural networks. *Elektronika ir Elektrotechnika*, 22(1):68–72, 2016.

[81] MW Oram and DI Perrett. Responses of anterior superior temporal polysensory (stpa) neurons to "biological motion" stimuli. *Journal of cognitive neuroscience*, 6(2):99–116, 1994.

[82] S Otte and MV Butz. Active tuning: signal denoising, reconstruction, and prediction with temporal forward model gradients (2019). pct. *EP2019/069659, patent pending*.

[83] Sebastian Otte, Matthias Karlbauer, and Martin V Butz. Active tuning. *arXiv preprint arXiv:2010.03958*, 2020.

[84] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*, pages 5–pp. IEEE, 2005.

[85] Marina Pavlova and Alexander Sokolov. Orientation specificity in biological motion perception. *Perception & Psychophysics*, 62(5):889–899, 2000.

[86] Marina A Pavlova. Biological motion processing as a hallmark of social cognition. *Cerebral Cortex*, 22(5):981–995, 2012.

[87] Kaare Brandt Petersen and MS Pedersen. The matrix cookbook, version 20121115. *Technical Univ. Denmark, Kongens Lyngby, Denmark, Tech. Rep*, 3274, 2012.

[88] JEAN PlAGET and BARBEL Inhelder. The child's conception of space. *London: Roudedge & Kegan Paul*, 1956.

[89] Alexandre Pouget, Peter Dayan, and Richard Zemel. Information processing with population codes. *Nature Reviews Neuroscience*, 1(2):125–132, 2000.

[90] Alexandre Pouget, Peter Dayan, and Richard S. Zemel. Inference and computation with population codes. *Annual Review of Neuroscience*, 26:381–410, 2003.

[91] Marisa Przyrembel, Jonathan Smallwood, Michael Pauen, and Tania Singer. Illuminating the dark matter of social neuroscience: Considering the problem of social interaction from philosophical, psychological,

and neuroscientific perspectives. *Frontiers in Human Neuroscience*, 6:190, 2012.

[92] Giacomo Rizzolatti and Laila Craighero. The mirror-neuron system. *Annu. Rev. Neurosci.*, 27:169–192, 2004.

[93] Giacomo Rizzolatti, Luciano Fadiga, Vittorio Gallese, and Leonardo Fogassi. Premotor cortex and the recognition of motor actions. *Cognitive brain research*, 3(2):131–141, 1996.

[94] Giacomo Rizzolatti and Giuseppe Luppino. The cortical motor system. *Neuron*, 31(6):889–901, 2001.

[95] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[96] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[97] Daniel Sabinasz, Mathis Richter, Jonas Lins, and Gregor Schöner. Speaker-specific adaptation to variable use of uncertainty expressions. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pages 620–627, 2020.

[98] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. (30):3856–3866, 2017.

[99] Mahdi Sadeghi, Fabian Schrodt, Sebastian Otte, and Martin V Butz. Binding and perspective taking as inference in a generative neural network model. *arXiv preprint arXiv:2012.05152*, 2020.

[100] Alessandro Salatiello, Mohammad Hovaidi-Ardestani, and Martin A Giese. A dynamical generative model of social interactions. *Frontiers in Neurorobotics*, 2021.

[101] Gregor Schöner. The dynamics of neural populations capture the laws of the mind. *Topics in cognitive science*, 12(4):1257–1271, 2020.

[102] Fabian Schrodt. *Neurocomputational Principles of Action Understanding: Perceptual Inference, Predictive Coding,and Embodied Simulation*. PhD thesis, University of Tübingen, 2018.

[103] Fabian Schrodt and Martin V Butz. Just imagine! learning to emulate and infer actions with a stochastic generative architecture. *Frontiers in Robotics and AI*, 3:5, 2016.

[104] Fabian Schrodt, Georg Layher, Heiko Neumann, and Martin Butz. Modeling perspective-taking by correlating visual and proprioceptive dynamics. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, 2014.

[105] Fabian Schrodt, Georg Layher, Heiko Neumann, and Martin V Butz. Modeling perspective-taking upon observation of 3d biological motion. In *4th International Conference on Development and Learning and on Epigenetic Robotics*, pages 305–310. IEEE, 2014.

[106] Fabian Schrodt, Georg Layher, Heiko Neumann, and Martin V. Butz. Embodied learning of a generative neural model for biological motion perception and inference. *Frontiers in Computational Neuroscience*, 9(79), 2015.

[107] Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971.

[108] Shenna Shepard and Douglas Metzler. Mental rotation: effects of dimensionality of objects and type of task. *Journal of experimental psychology: Human perception and performance*, 14(1):3, 1988.

[109] Yuuya Sugita, Jun Tani, and Martin V Butz. Simultaneously emerging braitenberg codes and compositionality. *Adaptive Behavior*, 19:295–316, 2011.

[110] Jun Tani. *Exploring Robotic Minds*. Oxford University Press, Oxford, UK, 2017.

[111] Jun Tani, Masato Ito, and Yuuya Sugita. Self-organization of distributedly represented multiple behavior schemata in a mirror system: reviews of robot experiments using rnnpb. *Neural Networks*, 17(8-9):1273–1289, 2004.

[112] Yung Liang Tong. *The multivariate normal distribution*. Springer Science & Business Media, 2012.

[113] Anne Treisman. The binding problem. *Current opinion in neurobiology*, 6(2):171–178, 1996.

[114] Anne Treisman. Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 353(1373):1295–1306, 1998.

[115] Zhigang Tu, Wei Xie, Qianqing Qin, Ronald Poppe, Remco C Veltkamp, Baoxin Li, and Junsong Yuan. Multi-stream cnn: Learning representations based on human-related regions for action recognition. *Pattern Recognition*, 79:32–43, 2018.

[116] Sebo Uithol and Markus Paulus. What do infants understand of others' action? a theoretical account of early social cognition. *Psychological Research*, 78(5):609–622, 2014.

[117] Erlinda R Ulloa and Jaime A Pineda. Recognition of point-light biological motion: mu rhythms and mirror neuron activity. *Behavioural brain research*, 183(2):188–194, 2007.

[118] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, (30):5998–6008, 2017.

[119] Ekaterina Volkova, Stephan De La Rosa, Heinrich H Bülthoff, and Betty Mohler. The mpi emotional body expressions database for narrative scenarios. *PloS one*, 9(12):e113647, 2014.

[120] Johan Wagemans, James H Elder, Michael Kubovy, Stephen E Palmer, Mary A Peterson, Manish Singh, and Rüdiger von der Heydt. A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological bulletin*, 138(6):1172, 2012.

[121] Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[122] Roy A Wise. Dopamine, learning and motivation. *Nature reviews neuroscience*, 5(6):483–494, 2004.

Publications

Full papers published in conference proceedings

[1] **Sadeghi. Mahdi**, Schrodt. Fabian, Otte. Sebastian, Butz. Martin. ”Gestalt Perception of Biological Motion: A Generative Artificial Neural Network Model.”. In *International Conference on Development and Learning (ICDL) (pp. 1-7) - IEEE*, 2021. Link

Sadeghi; Contributed to methodological design, architecture implementation , data analysis, and wrote the first draft of manuscript.

Schrodt; Contributed to methodological design.

Otte; Contributed to writing and revising the manuscript.

Butz; Contributed to methodological design, data analysis, and writing and revising the manuscript.

[2] **Sadeghi. Mahdi**, Schrodt. Fabian, Otte. Sebastian, Butz. Martin. Binding and Perspective Taking as Inference in a Generative Neural Network Model. In *30th International Conference on Artificial Neural Networks, ICANN*, (pp. 3-14), Springer, 2021. Link

Sadeghi; Contributed to methodological design, architecture implementation , data analysis, and wrote the first draft of manuscript.

Schrodt; Contributed to methodological design.

Otte; Contributed to writing and revising the manuscript.

Butz; Contributed to methodological design, data analysis, and writing and revising the manuscript.