

# Perception of Human Movement Based on Modular Movement Primitives

**Kumulative Dissertation**  
zur Erlangung des Grades eines  
DOKTOR DER NATURWISSENSCHAFTEN  
(Dr. rer. nat.)  
des Fachbereichs Psychologie  
der Philipps-Universität Marburg

Vorgelegt von  
Benjamin Knopp, M.Sc.  
Aus Wuppertal

Marburg an der Lahn, 2021



Die vorliegende Dissertation wurde von April 2016 bis Februar 2021 am  
Fachbereich Psychologie, Philipps-Universität Marburg unter Leitung von Prof.  
Dominik Endres angefertigt.

Vom Fachbereich Psychologie  
der Philipps-Universität Marburg (Hochschulkenziffer 1180)  
als Dissertation angenommen am 22.02.2021.

Erstgutachter: Prof. Dominik Endres  
Zweitgutachter: Prof. Gunnar Blohm  
Tag der Disputation: 30.04.2021

Originaldokument gespeichert auf dem Publikationsserver der  
Philipps-Universität Marburg  
<http://archiv.ub.uni-marburg.de>



Dieses Werk bzw. Inhalt steht unter einer  
Creative Commons  
Namensnennung  
Keine kommerzielle Nutzung  
Weitergabe unter gleichen Bedingungen  
3.0 Deutschland Lizenz.

Die vollständige Lizenz finden Sie unter:  
<http://creativecommons.org/licenses/by-nc-sa/3.0/de/>

## Perception of Human Movement Based on Modular Movement Primitives

### ABSTRACT

People can identify and understand human movement from very degraded visual information without effort. A few dots representing the position of the joints are enough to induce a vivid and stable percept of the underlying movement. Due to this ability, the realistic animation of 3D characters requires great skill. Studying the constituents of movement that looks natural would not only help these artists, but also bring better understanding of the underlying information processing in the brain.

Analogous to the hurdles in animation, the efforts of roboticists reflect the complexity of motion production: controlling the many degrees of freedom of a body requires time-consuming computations. Modularity is one strategy to address this problem: Complex movement can be decomposed into simple primitives. A few primitives can conversely be used to compose a large number of movements. Many types of movement primitives (MPs) have been proposed on different levels of information processing hierarchy in the brain. MPs have mostly been proposed for movement production. Yet, modularity based on primitives might similarly enable robust movement perception.

For my thesis, I have conducted perceptual experiments based on the assumption of a shared representation of perception and action based on MPs. The three different types of MPs I have investigated are temporal MPs (TMP), dynamical MPs (DMP), and coupled Gaussian process dynamical models (cGPDM).

The MP-models have been trained on natural movements to generate new movements. I then perceptually validated these artificial movements in different psychophysical experiments. In all experiments I used a two-alternative forced choice paradigm, in which human observers were presented a movement based on motion-capturing data, and one generated by an MP-model. They were then asked to choose the movement which they perceived as more natural.

In the first experiment I investigated walking movements, and found that, in line with previous results, faithful representation of movement dynamics is more important than good reconstruction of pose. In the second experiment I investigated the role of prediction in perception using reaching movements. Here, I found that perceived naturalness of the predictions is similar to the perceived naturalness of movements itself obtained in the first experiment.

I have found that MP models are able to produce movement that looks natural, with the TMP achieving the highest perceptual scores as well highest predictiveness of perceived naturalness among the three model classes, suggesting their suitability for a shared representation of perception and action.

# Contents

<b>1</b>	<b>INTRODUCTION</b>	<b>1</b>
1.1	Perception . . . . .	3
1.1.1	Computational models of the visual system . . . . .	3
1.1.2	Biological Movement Perception . . . . .	4
1.2	Action . . . . .	7
1.2.1	Muscle Synergies . . . . .	7
1.2.2	Kinematic Primitives . . . . .	8
1.2.3	Dynamical Primitives . . . . .	8
1.2.4	Complexity estimation . . . . .	9
1.3	Perception and Action . . . . .	10
1.3.1	A shared representation for perception and action . . . . .	10
1.4	Research Rationale . . . . .	11
<b>2</b>	<b>SUMMARIES OF PUBLISHED PAPERS</b>	<b>14</b>
2.1	“The Variational Coupled Gaussian Process Dynamical Model” . . . . .	15
2.2	“Making the Coupled Gaussian Process Dynamical Model Modular and Scalable with Variational Approximations” . . . . .	15
2.3	“Predicting Perceived Naturalness of Human Animations Based on Generative Movement Primitive Models” . . . . .	16
2.4	“Evaluating Perceptual Predictions Based on Movement Primitive Models in VR- and Online-Experiments” . . . . .	17
<b>3</b>	<b>GENERAL DISCUSSION</b>	<b>19</b>
3.1	Limitations . . . . .	20
3.2	Future directions . . . . .	20
3.3	Conclusion . . . . .	22
	<b>REFERENCES</b>	<b>23</b>
	<b>APPENDIX A AUTHOR CONTRIBUTIONS</b>	<b>30</b>

A.1	“The Variational Coupled Gaussian Process Dynamical Model” . . . . .	30
A.2	“Making the Coupled Gaussian Process Dynamical Model Modular and Scalable with Variational Approximations” . . . . .	30
A.3	“Predicting Perceived Naturalness of Human Animations Based on Generative Movement Primitive Models” . . . . .	31
A.4	“Evaluating Perceptual Predictions Based on Movement Primitive Models in VR- and Online-Experiments” . . . . .	31
APPENDIX B THE VARIATIONAL COUPLED GAUSSIAN PROCESS DYNAMICAL MODEL		<b>32</b>
APPENDIX C MAKING THE COUPLED GAUSSIAN PROCESS DYNAMICAL MODEL MODULAR AND SCALABLE WITH VARIATIONAL APPROXIMATIONS		<b>42</b>
APPENDIX D PREDICTING PERCEIVED NATURALNESS OF HUMAN ANIMATIONS BASED ON GENERATIVE MOVEMENT PRIMITIVE MODELS		<b>68</b>
APPENDIX E EVALUATING PERCEPTUAL PREDICTIONS BASED ON MOVEMENT PRIMITIVE MODELS IN VR- AND ONLINE-EXPERIMENTS		<b>87</b>
APPENDIX F ZUSAMMENFASSUNG IN DEUTSCHER SPRACHE		<b>97</b>
APPENDIX G EIGENSTÄNDIGKEITSERKLÄRUNG		<b>99</b>

# Acknowledgments

I thank my supervisor, Prof. Dominik Endres for guidance, encouragement and inspiration. Many discussions on a wild variety of topics have been enjoyable and enlightening. Dominik ensured my freedom to explore, yet always pointing me in the right direction, so that I could stay focused.

Furthermore I thank Prof. Gunnar Blohm for inviting me to Kanada to join his extraordinary team and work on a very exciting project. He has proved his hospitality, cooking skills, and gave great career advice.

I thank Prof. Alexander Schütz and Prof. Frank Bremmer for kindly serving as member of my thesis committee. Alexander provided valuable input to my project. Frank's lectures led my way to neuroscience, and he has accompanied me on this path until today, as speaker of the IRTG 1901.

I thank all members of Dominik's lab, especially Dmytro Velychko. His critical remarks and insights have often been very valuable. I thank Olaf Haag for his tremendous effort in recording data. I thank Adrian Schütz and Michaela Vorn-dran for making the best of the pandemic in our care community during lockdown. Special thanks to Raphael Schween and Riccardo Scott for their constructive criticism.

Finally, I thank my parents and my sister, my family-in-law, my friends and band-mates. Most importantly, I would like to thank my wife Tabea for her encouragement, support and advice. Also, I want to thank my daughter Matilda, for giving me purpose and teaching me patience.

# 1

## Introduction

The main assumption of this thesis is a shared representation for action and perception. I will motivate this assumption by first reviewing accounts of perception and action converging towards a common representation for both. Then I will describe how the perceptual experiments carried out for this thesis can be used to draw conclusions about this assumption.

It is helpful to keep the levels of analysis in mind to understand the focus of my dissertation. They were proposed to guide the investigation of an enormously complex system like the brain (Marr 1982):

1. On the **computational level**, questions regarding the goal of the computation are addressed. This specifies *what* problem is solved by the system.
2. On the **algorithmic level**, the representations and necessary computations are defined to answer *how* to solve the problem.
3. On the **implementational level**, an exact description of an physical system representing the data and performing the computation.

My primary interest lies on the algorithmic level, focusing on representations used by a system towards solving problems posed by normative theories (Kording et al. 2020). The free-energy minimization principle provides a normative perspective which is useful with this regard, because it provides a unified account for action and perception (see Chapter 1.3.1). Other normative theories (Schmidhuber 2010) might propose different goals, but this most likely does not change the involved representations needed. With regards to the implementational level,

the CPU based computations are only lightly constrained by biological realism: Representations are only loosely attributed to different subpopulations of neurons in the central nervous system.

## 1.1 PERCEPTION

Our visual system is highly optimized: without effort we can distinguish objects and recognize and interpret human movement. In this chapter, I first review some computational models in Chapter 1.1.1 for early vision and object recognition, because it illustrates how specifying goals on a computational level together with representations and algorithms can explain and predict neuronal behaviour. This provides the perspective on biological movement perception behind the rationale of my psychophysical experiments for which I present some neurophysiological and computational work in Chapter 1.1.2.

### 1.1.1 COMPUTATIONAL MODELS OF THE VISUAL SYSTEM

Since the ground-breaking work of Hubel and Wiesel (1959), the complex hierarchy of the visual system has been investigated heavily (Felleman and Van Essen 1991). The wealth of neurophysiological evidence enabled computational neuroscientists to simulate neurons from the primary visual area (V1) up to the inferior temporal gyrus (area IT), using a model called HMAX (Hierarchical-MAX-pooling, Riesenhuber and Poggio 1999): This model could recognize objects using a manually tuned five-layered neural network with an architecture inspired by common assumptions of neurophysiologists. Even though the object depictions (stimuli) were simple, this was a major achievement before the advent of deep learning.

In parallel, investigation of the visual system in terms of its function was approached by the investigation of natural image statistics (Field 1987; Hyvärinen, Hurri, and Hoyer 2009). Based on the assumption of optimal adaptation to visual input, investigation of natural images enabled the specification of *what* the visual system does. The most prominent example for this approach is the explanation of V1 receptive field properties by the postulation of a sparse code of natural images (Olshausen and Field 1996). Rao and Ballard (1999) proposed a model of visual processing which minimizes prediction error in a hierarchical model: Higher level neurons try to predict lower level responses. Lower levels receive these predictions via feedback connections and return the error between this prediction and their actual response via feedforward connections. The lowest level corresponds to a natural image, while the highest level represents a latent representation of the “causes” of the image. The model is trained to minimize prediction error across all levels. The properties of the model neurons which emerge during training can

account for even more properties of V1 neurons than the sparse coding model.

These endeavours can be seen as precursors of the Bayesian Brain theory (Knill and Pouget 2004) for perception and the free-energy principle (Friston 2010) for action and perception. The data-centeredness of this statistical-ecological approach (Hyvärinen, Hurri, and Hoyer 2009) combined with deep architectures similar to HMAX revolutionized computer vision when the backpropagation algorithm was implemented efficiently on GPUs (Krizhevsky, Sutskever, and Hinton 2012): A convolutional neural network (CNN), when trained to classify images, learns parameters in its lowest level which resemble V1 receptive fields (Zeiler and Fergus 2014), as well as predicting IT cell responses (Yamins et al. 2014). Furthermore, CNNs surpass human classification performance on specific datasets (He et al. 2015). While this is an impressive result, human object recognition is far more robust (Geirhos et al. 2018), which is showcased by adversarial images: minimal changes to an image not noticeable to a human observer can be found which cause the model to misclassify (Szegedy et al. 2013).

In summary, we see that natural, ecologically valid input is important for addressing normative questions on the computational level of analysis. Much progress in object recognition has been made, initially inspired by neurophysiological findings. Yet, chasing after the highest accuracy score for object recognition has recently led away from biologically inspired models. One (of many) ways to make machine learning more human-like is to consider different goals than object classification, e.g. biological movement perception.

### 1.1.2 BIOLOGICAL MOVEMENT PERCEPTION

#### PSYCHOPHYSICS

Point-light stimuli were introduced by Johansson (1973) to investigate biological motion: He attached light bulbs to the main joints of an actor walking, running, or dancing in the dark. Human observers identify the underlying movement from the resulting motion patterns with ease, demonstrating the impressive capabilities of the perceptual system. Point-light stimuli gained huge influence in the investigation of human (Troje 2002; Johansson 1973) and animal (Troje and Westhoff 2006) movement perception (for reviews see: Troje and Chang 2013).

They enable investigation of biological motion irrespective of body form and easy to analyze. Still, body motion and form are both part of biological movement

perception (Giese and Poggio 2003; Theusner, de Lussanet, and Lappe 2011). From the perspective of natural image statistics, movement stimuli should match the visual statistics of real-life moving humans. This has yet to be achieved, but stick figures and volumetric avatars are approximations towards more ecologically valid stimuli. Stick figures visualize the connections between the joints, thus making the kinematic hierarchy explicit. In contrast to point-light stimuli, no movement is necessary to infer the body shape. Therefore, the visual system needs no extra frames to infer the shape and can use more information for the actual dynamics. Still, they are as simple to implement as point-light stimuli. Volumetric avatars, on the other hand, allow for a high degree of realism, but are harder to implement. Volumetric avatars have shown a slight increase of perception sensitivity compared to stick-figures (Hodgins, O'Brien, and Tumblin 1998).

#### MOVEMENT PERCEPTION IN SUPERIOR TEMPORAL SULCUS

Neurophysiological evidence suggests that the superior temporal sulcus (STS) is involved in biological movement perception (Perrett et al. 1985). Two more or less separate visual information pathways converge in STS after approximately a 1/10 seconds (Oram and Perrett 1996). Neurons in STS respond to biological motion (as well as to theory-of-mind-, face-recognition-, voice- and story-tasks: Deen et al. 2015), which are connected over the posterior parietal lobe to the premotor cortex, both responding to displays of action as well (Nelissen, Borra, et al. 2011; Nelissen, Luppino, et al. 2005).

Models that incorporate and explain some of the neurophysiological findings have been devised (Giese and Poggio 2003; Jhuang et al. 2007; Theusner, Lussanet, and Lappe 2014; Simonyan and Zisserman 2014): Giese and Poggio (2003) build upon the HMAX model described in Chapter 1.1.1 implementing form and motion pathways for robust activity recognition. The model exhibits many traits found in psychophysical experiments like the inversion effect (Troje and Westhoff 2006). Simonyan and Zisserman (2014) use a similar deep architecture, but use learned instead of hand-crafted features. While deep learning enables impressive results on action recognition, they can not handle point-light stimuli as humans can (Peng et al. 2021).

## MOVEMENT PERCEPTION IN MOTOR AREAS

Point-light walkers also evoke activity in sensory-motor and supplementary motor areas. In an EEG-Study Inuggi et al. (2018) found an event-related potential 435ms after stimulus onset, but only for walking movement associated with locomotion, in contrast to tread-mill walking movement. This result suggests, that meaningful (goal-directed) action is visually processed by motor areas (see also Thompson, Bird, and Catmur 2019).

## MOVEMENT PERCEPTION: CONCLUSION

One discrepancy with neurophysiology of all these models are missing recurrent connections (Spoerer et al. 2020), which might explain why artificial systems have worse generalization ability compared to human perception. Investigations of these issues is still at the very beginning (Serre 2019). Importantly, there exists neurophysiological evidence that motor areas are involved in perceptual processes, but until now, there are only a few computational models that account for this finding (see Chapter 1.3.1).

## 1.2 ACTION

In this chapter I present computational and neurophysiological work on motor production. The focus here is on embedding the movement primitive types I used for the psychophysical experiments of this thesis into the hierarchy of motor production, i.e. primitives proposed to simplify control on a distal level, and ones used for planning on a more proximal level. The primitives which form the basis of my psychophysical investigations are embedded in this context. Please refer to the appended papers for a more detailed explanation.

### 1.2.1 MUSCLE SYNERGIES

Movement primitives on the lowest level of motor production are commonly referred to as muscle synergies (Macpherson 1988). They describe muscle activity (measured by Electromyography, EMG) as linear combination of muscle co-activations:

$$\vec{x} = \sum_{q=1}^Q w_q \vec{y}_q \quad (1.1)$$

Here,  $\vec{x} \in \mathbb{R}^D$  is the vector of  $D$  EMG signals at a given time which is constructed by  $Q$  primitives  $\vec{y} \in \mathbb{R}^D$ . If  $Y = (\vec{y}_1, \dots, \vec{y}_Q)$  is chosen suitably, it provides a linear mapping from a low-dimensional representation, i.e. synergy activations  $\vec{w} = (w_1, \dots, w_Q)^T$ , to the observed naturally occurring muscle activity.

These synergies are activated at a given time, thus termed synchronous or spatial synergies (Tresch and Jarc 2009). In naturally behaving animals or humans, synergies are identified using blind source separating algorithms, most commonly non-negative matrix factorization (NMF, Lee and Seung 1999), but also Factor Analysis (FA), Principle Component Analysis (PCA), Independent Component Analysis (ICA) and Anechoic Mixture Models (AMM Omlor and Giese n.d.), for a comparison see Tresch, Cheung, and d’Avella (2006); Endres, Chiovetto, and Giese (2013). Synergies thus found could explain different forms of motor impairment after stroke (Cheung et al. 2012). Levine et al. (2014) found motor synergy encoder neurons in the spine of mice, providing an interface between motor cortex and motorneurons to reduce the number of DOFs.

Central pattern generators (CPGs) build a class of spinal control systems in the temporal domain: They are neural circuits capable of producing rhythmic output (Grillner 2006). They could serve as neural basis for the activation of muscle synergies (Mussa-Ivaldi and Solla 2004). Temporal and muscle synergies can be combined into time-varying (d’Avella et al. 2008).

The proposed neural origin of muscle synergies is debated. The low dimensionality observed in EMG muscle signals might also be a consequence of biomechanical- and task constraints (Kutch and Valero-Cuevas 2012). For the existence of perceptual MPs, however, the answer to this debate should not be fundamental, because they most likely reside on another hierarchy level.

### 1.2.2 KINEMATIC PRIMITIVES

On a higher, more abstract level, kinematic primitives have been proposed. Invariances, e.g. the two-thirds power law, in 2D and 3D end-effector trajectories (Viviani and Flash 1995; Endres, Meirovitch, et al. 2013) can be explained by piecewise constant polynomial trajectories, i.e. temporal primitives, encoded by via-points. Viviani and Stucci (1992) found that the power law holds in perception of movement as well. This suggests that movement perception might be based on MPs as well. **Temporal MPs** based on Gaussian processes (GPs, Rasmussen and Williams 2006), which have also been used as trajectory representation in robots (Clever et al. 2016), is one of the MP types I have studied for this dissertation.

### 1.2.3 DYNAMICAL PRIMITIVES

The class of Dynamical Primitives uses dynamical systems to represent the movement, in contrast to describing the movement signal (trajectories or activations) directly. Dynamical primitives are popular in robotics, due to their robustness against perturbation. Especially popular are **DMPs** proposed by Ijspeert et al. (2013), where they are typically used to provide kinematic motor plans, with subsequent controllers turning these plans into action. Perceptual capabilities of this model has also been demonstrated by handwritten character recognition. While they provide some flexibility in planning, they can not be composed to form new complex movement.

The Gaussian process dynamical model (GPDM, Wang, Fleet, and Hertzmann 2008) provides a Bayesian model for dynamical MPs, which has initially been

applied for computer graphics. Here, Gaussian process (GP) regression is used to learn a dynamics mapping in a low-dimensional latent space, i.e. predictions from previous latent states to the next one. GP regression is also used to learn the mapping from the latent states to the movement data, at each time-point. It is often used to model a time series of poses, but can be applied to model muscle activation as well, which can be interpreted in neurophysiological terms: the latent-state corresponds to the synergy activations, whose temporal evolution is governed by dynamics corresponding to CPGs, with a (now non-linear) latent-state to muscle-activation mapping corresponding to the muscle synergies. GPDMs do not provide a compact representation of movement, because each MP is parameterized by examples of whole-body movements, and it lacks modularity as is the case with DMPs.

The **coupled GPDM** (Velychko, Endres, et al. 2014), when used as a model for motor production, describes different body parts as individual GPDMs. Now, each GPDM predicts not only its own next latent state, but the latent states of all the other GPDMs as well, and combines all predictions, thus coupling all parts. This enables modular recombinations of body-part specific MPs. The introduction of sparse variational approximations make the representation compact (Velychko, Knopp, and Endres 2017), with controllable complexity (see Appendix C).

#### 1.2.4 COMPLEXITY ESTIMATION

Regardless of the MP type, the complexity of the resulting representation has to be specified a priori, e.g. the number of primitives. While many studies use an arbitrary value for the variance accounted for (VAF), it has been suggested to use the Bayesian model score, or approximations thereof, to determine this complexity parameter (Endres, Chiovetto, and Giese 2013). One goal of this thesis is to evaluate if model scores provide an useful account for the perceived naturalness of movement (see Chapter 1.4).

### 1.3 PERCEPTION AND ACTION

In the two previous chapters I have reviewed the motor- as well as the visual system of the central nervous system. There is a substantial overlap of brain areas involved in visual and motor processing.

#### 1.3.1 A SHARED REPRESENTATION FOR PERCEPTION AND ACTION

Di Pellegrino et al. (1992) first discovered neurons in premotor cortex which are activated by observing as well as performing an action. These were subsequently termed Mirror Neurons Gallese et al. (1996), and gained widespread public interest, but the meaning of their discovery has often been mis- and/or over-interpreted (Hickok 2009).

While empirical data is ambiguous, there are some theoretical accounts which provide explanations for firing pattern of mirror neurons.

Prinz (1997) introduced the common coding approach to perception and action: Instead of treating action planning and visual perception as separate, he argues that they share a representation. He furthermore suggests that actions are planned according to their perceived outcome, thereby action is represented in the same vein as perception. This idea can be traced back to William James' ideomotor theory (James 1890). Hommel et al. (2001) (see Hommel 2019, for an update) include the common coding approach into their theory of event coding (TEC), which additionally postulates feature-based coding of events, and emphasize the distal nature of the representation, i.e. codes are very different from muscle innervation patterns and retinal information. Instead, the required code should be a more abstract, maybe language like representation. Looking back at Chapter 1.2, we see the similarity of the required action-feature codes to MPs. One limitation of TEC is that experimental data is yet mostly limited to very simple button-press action tasks, with little ecological validity. Furthermore, the theory gives no explicit definition of the common code and gives a simplified two-level hierarchy.

Kilner, Friston, and Frith (2007) and Friston, Mattout, and Kilner (2011) provide an account of the mirror neuron system within the free-energy framework, which incorporates ideas of TEC, with some differences:

1. the common code approach is taken to the extreme, by eliminating distinctions between motor and sensory representation altogether ("The only difference between the motor cortex and visual cortex is that one predicts

retinotopic input, while the other predicts proprioceptive input from the motor plant” Friston, Mattout, and Kilner 2011).

2. They stress the importance of a deep hierarchy.

In Friston, Mattout, and Kilner (2011), this scheme is implemented (even though the toy model of handwriting only contains one layer) for the production and recognition of handwriting: A generative model predicts proprioceptive and visual sensory information from a hidden cause. Recognition is achieved by inverting the generative model: The hidden cause of sensory input is inferred with visual input fixed and without proprioceptive input by predictive error minimization. In action production, the hidden cause corresponds to an intention, and is fixed. The system then is forced to choose action that move its state to fulfill the predictions of the generative model. Even though these simulations demonstrate the viability of the general approach, the exact hierarchy remains unspecified and do not answer if the model can handle high-dimensional and ecologically valid data.

Similarly, neurophysiological experiments do not provide sufficient constraints for biologically plausible implementation of a common code. Mirror neurons have been found in premotor area F5 and the inferior parietal lobe, area PF (Fogassi et al. 2005; Nelissen, Luppino, et al. 2005). Area PF is also part of one path between STS and F5 (Nelissen, Borra, et al. 2011).

#### 1.4 RESEARCH RATIONALE

The previous chapters have shown converging neurophysiological and theoretical evidence for a shared representation of action and perception. Nevertheless, the exact nature of this representation is still undetermined.

In this thesis, I assume that MPs can serve as shared representation of action and perception. This addresses the issue of the unclear specification of mirror neuron system accounts described in this chapter. Chapter 1.2 has shown, that MPs provide a way to simplify control in real world applications. They furthermore provide a level of abstraction from the kinematic representation, which has also been found in the mirror neuron system. Yet, their applicability as shared representation of action and perception has not been investigated explicitly.

I use model comparison to evaluate the feasibility of MPs as perceptual representation. The marginal likelihood provides a principled way to compare models:

When used to compare a set of models containing the data generating model, this model will score the highest. It is given by:

$$p(\mathcal{D}|\mathcal{M}) = \int_{\Theta} p(\mathcal{D}|\Theta, \mathcal{M})p(\Theta|\mathcal{M})d\Theta. \quad (1.2)$$

This is the normalizing constant appearing in Bayes’ Formula (e.g. Koller and Friedman 2009; Bishop 2007; Murphy 2012), also known as *model evidence*, and describes the probability of the data  $\mathcal{D}$  irrespective of specific model parameters  $\Theta$  for a model  $\mathcal{M}$ . In our case, the  $\mathcal{D}$  is the set of movements. The model  $\mathcal{M}$  is specified by the MP type together with the complexity parameter, e.g. TMP with 5 primitives, or vCGPDM with 11 dynamics- and 17 pose-inducing points.  $\Theta$  are the specific model parameters, e.g. the weight and primitives for the TMP, or latent variables and inducing points for the vCGPDM.

While the marginal likelihood provides the theoretically best way to compare models, it is in many instances intractable, and therefore we have to find good approximations for it. In this thesis I consider two model scores as approximations to the marginal likelihood: the evidence lower bound (ELBO) and the crossvalidation mean-squared-error (MSE). The ELBO is obtained by introducing an approximate posterior which enables computational tractability and can be optimized to be close to the exact posterior. Thus, in variational models we learn the posterior of the parameters and simultaneously obtain a lower bound on the marginal likelihood, which is useful for model comparison. Yet, the approximations might be inappropriate. A computationally more expensive approximation can be obtained by using M-fold cross-validation: The data set is partitioned into  $M$  sets. Then  $M - 1$  sets are used to train the model, which is then used to predict the left out set. Thus, we can compute  $M$  MSE values (in case of Gaussian distributions), whose average can be used for model comparison.

Theoretically, if one of the evaluated models is describing how the CNS actually produces movement, it would have the highest marginal likelihood. If the perceptual system uses the same model, it would be the model which generates the perceptually most valid movement. Even though this is almost certainly not the case (Box 1976)\*, I assume that a close correspondence between model score

---

\*“Since all models are wrong the scientist cannot obtain a ”correct” one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena.”

and perceptual validity indicates if a model is suitable for a shared representation.

The perceptual validity is estimated in a two-alternative forced choice (2AFC) and two-interval forced choice (2IFC) paradigms (Fechner 1860), where a model generated movement and a baseline (mocap-based) movement are shown simultaneously (2AFC) or sequentially (2IFC). The participant then was forced to respond which movement she perceived as more natural. The number of trials where a model fools a participant divided by the number of all trials gives the confusion rate, which measures the perceived naturalness of the model.

The correspondence between model score and perceptual can then be estimated using logistic regression (see Appendix D 3.3).

$$p_i = \frac{1}{1 + \exp(-(\alpha + \beta \cdot \text{modelscore}_i))} \quad (1.3)$$

$$r_i \sim \text{Bernoulli}(p_i) \quad (1.4)$$

The generative perspective of logistic regression is as follows: The model score of each trial  $i$  is linearly transformed with intercept  $\alpha$  and scale  $\beta$  and mapped to the interval  $(0, 1)$ . The predicted confusion rate  $p_i$  is then used as parameter for the Bernoulli distribution to sample the response  $r_i$ .

Estimates/posterior values for the parameters  $\alpha$  and  $\beta$  are obtained by maximum-likelihood and MCMC sampling. To evaluate the predictiveness of the model score for the perceived naturalness, we can compute the logarithmic likelihood-ratio <sup>†</sup>.

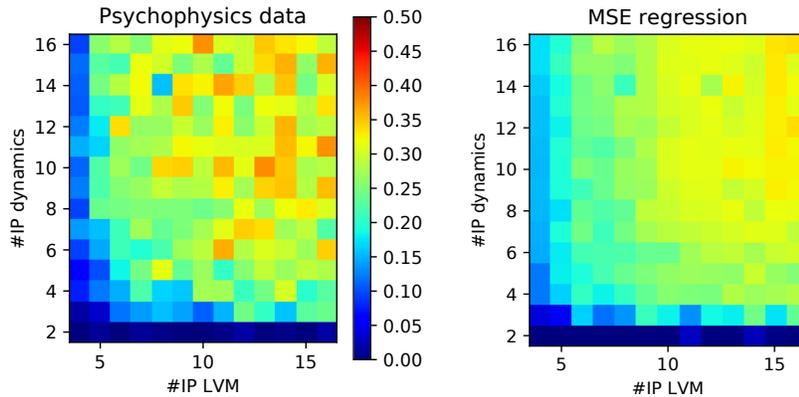
---

<sup>†</sup>the ratio between the likelihood of the data being generated by the model and the likelihood of the data being generated by a constant  $p$

# 2

## Summaries of Published Papers

In this Chapter, I provide brief summaries and selected figures from the publications. Please refer to the full papers in the Appendix.



**Figure 2.1:** Results of the psychophysical experiment, Appendix B, Fig. 2: (Left) Measured confusion rate and (Right) logistic regression for vCGPDM models parametrized by the number of dynamics- (#IP dynamics) and pose- (#IP LVM) inducing points.

## 2.1 “THE VARIATIONAL COUPLED GAUSSIAN PROCESS DYNAMICAL MODEL”

Dmytro Velychko, Benjamin Knopp, and Dominik Endres (2017). “The Variational Coupled Gaussian Process Dynamical Model”. In: *International Conference on Artificial Neural Networks*. Springer, pp. 291–299. DOI: [10.1007/978-3-319-68600-4\\_34](https://doi.org/10.1007/978-3-319-68600-4_34) Appendix B

In this paper, sparse variational approximations are introduced to the coupled Gaussian process dynamical model (cGPDM Velychko, Endres, et al. 2014). This results in a compact representation, which allows for modular recombinations of primitives.

I implemented and ran a psychophysical experiment to evaluate the resulting compact representation for walking movements, and analyzed the data. This compact representation is perceptually valid, and the perceptual validity can be predicted from the mean squared kinematics error (Fig. 2.1).

## 2.2 “MAKING THE COUPLED GAUSSIAN PROCESS DYNAMICAL MODEL MODULAR AND SCALABLE WITH VARIATIONAL APPROXIMATIONS”

Dmytro Velychko, Benjamin Knopp, and Dominik M. Endres (Oct. 2018). “Making the Coupled Gaussian Process Dynamical Model Modular and Scalable with Variational Approximations”. In: *Entropy* 20.10, p. 724. DOI: [10.3390/e20100724](https://doi.org/10.3390/e20100724) Appendix C

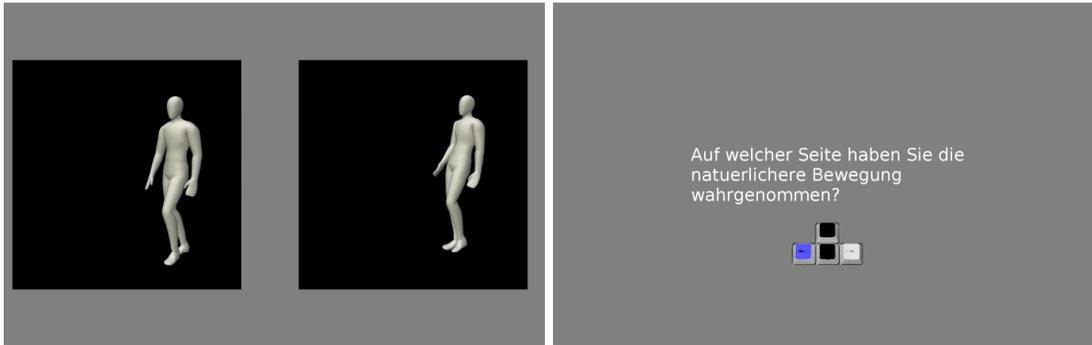


Figure 2.2: Screenshot of the experiment, Appendix B C Fig. 5

“Making the Coupled Gaussian Process Dynamical Model Modular and Scalable with Variational Approximations” extends “The Variational Coupled Gaussian Process Dynamical Model” by introducing synthetic and object-passing-movement datasets, and providing new analyses for the perceptual- and cross-validatory model comparisons. Fig. 2.2 depicts the volumetric avatar used in context of the experimental graphical user interface.

### 2.3 “PREDICTING PERCEIVED NATURALNESS OF HUMAN ANIMATIONS BASED ON GENERATIVE MOVEMENT PRIMITIVE MODELS”

Benjamin Knopp, Dmytro Velychko, Johannes Dreibrodt, and Dominik Endres (Sept. 2019). “Predicting Perceived Naturalness of Human Animations Based on Generative Movement Primitive Models”. In: *ACM Trans. Appl. Percept.* 16.3, 15:1–15:18. ISSN: 1544-3558. DOI: [10.1145/3355401](https://doi.org/10.1145/3355401) Appendix D

In this paper, I considerably extended the perceptual validation experiments. In addition to the vCGPDM, Temporal- and Dynamical MP models were concluded and experimentally validated. We found that for walking movements, temporal MPs achieve the highest perceptual validity (see Fig. 2.3). We demonstrated, that perceived naturalness of the generated movements can be predicted by model scores. In particular, because the dynamics and pose can be disentangled in the terms of the ELBO, we could show, in line with previous results, that faithful modelling of the movement dynamics is more important for perception than a flexible model of poses.

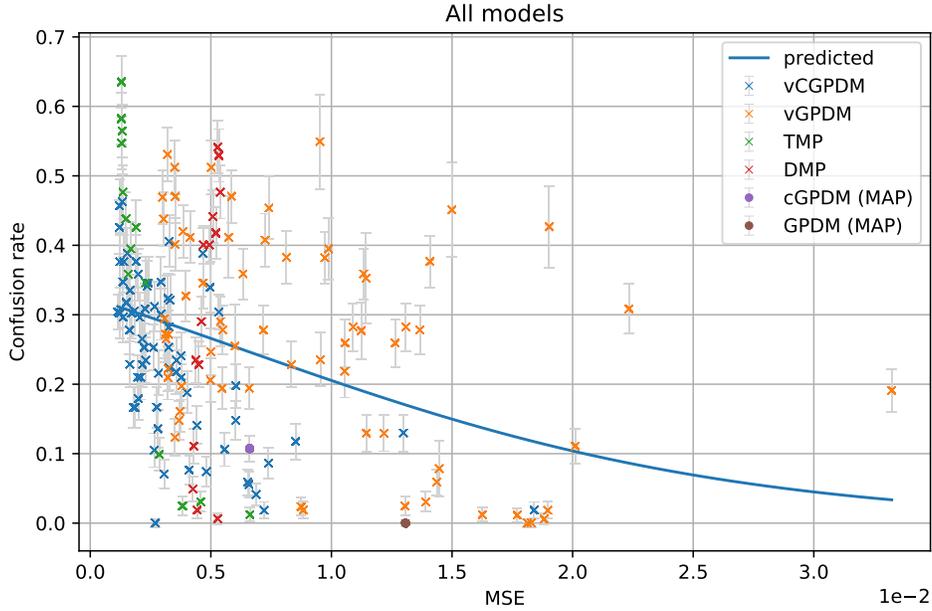
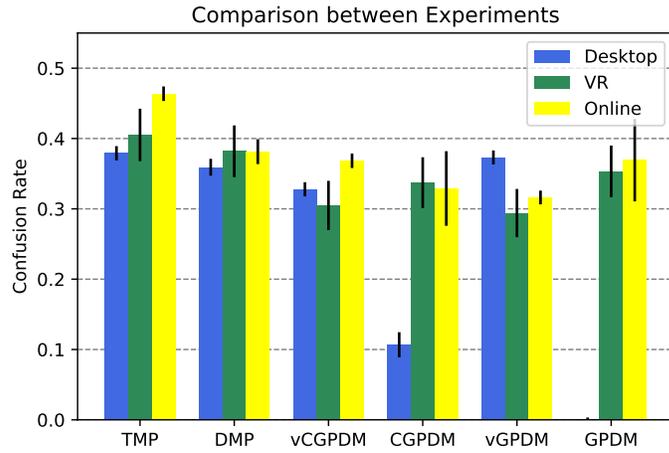


Figure 2.3: Measured confusion rate for all tested models, Appendix D, Fig. 7

## 2.4 “EVALUATING PERCEPTUAL PREDICTIONS BASED ON MOVEMENT PRIMITIVE MODELS IN VR- AND ONLINE-EXPERIMENTS”

Benjamin Knopp, Dmytro Velychko, Johannes Dreibrodt, Alexander C. Schütz, et al. (Sept. 12, 2020). “Evaluating Perceptual Predictions Based on Movement Primitive Models in VR- and Online-Experiments”. In: *ACM Symposium on Applied Perception 2020*. SAP ’20: ACM Symposium on Applied Perception 2020. Virtual Event USA: ACM, pp. 1–9. ISBN: 978-1-4503-7618-1. DOI: [10.1145/3385955.3407940](https://doi.org/10.1145/3385955.3407940) Appendix E

Predictions are essential for a common representation of action and perception if this should enable us to interact (Schütz-Bosbach and Prinz 2007): The latency between stimulus presentation and response in the STS is approximately 1/10s (Endres and Oram 2010). Therefore, a new paradigm to evaluate if predictions of different MP types match biological movement perception was introduced in Knopp, Velychko, Dreibrodt, Schütz, et al. (2020). For this we use object-passing movements and two experimental settings. We first conducted a VR experiment using volumetric avatars as stimulus, and then conducted the experiment in a web-browser based implementation using stick-figure stimuli. We found comparable



**Figure 2.4:** Average confusion rates of MP Types in three different paradigms, Appendix E, Fig. 3.

results in both settings, despite the different stimulus form. Furthermore, the perceived naturalness of the predictions is similar to the perceived naturalness of movements we have found in our previous (non-predictive) perception experiments (see Fig. 2.4).

# 3

## General Discussion

I have established a framework for evaluating the perceptual validity of generative models of movement production. In this framework I have evaluated perceived naturalness of movement and movement predictions generated by three distinct classes of movement primitives. To the best of my knowledge, my studies are the first to rigorously evaluate the complexity of movement primitive (MP) representations for whole-body movements required for perceptual believability. Walking and object-passing movements have been used for these evaluations. I have found that temporal movement primitives (TMP) achieved the highest perceptual validity of walking movements as well as object passing predictions.

The motivation for these studies stems from the assumption of a common code for action and perception, for which I have reviewed theories and neurophysiological findings in Chapter 1. While perceptual experiments alone are not sufficient to test this hypothesis, it is possible to collect evidence for it by testing derived consequences. Under this assumption, the model used to produce the movement can be linked to the perception of a movement. Here, I used logistic regression to link MP model scores to perceived naturalness of movements or movement predictions (see Chapter 1.4).

Of all models, the TMP are most predictive for perception. Together with the high perceptual scores, this makes them the most likely candidate for a shared representation.

Dynamical movement primitive (DMP) models also have proven their capabil-

ity of producing perceptually plausible movement. Yet, they lack predictiveness for the perceptual validity of movements, which limits their usefulness as perceptual representation in humans.

In comparison to TMP and DMP models, coupled and monolithic Gaussian process dynamical models ((c)GPDM) achieved on average lower perceptual validity, so they are probably less suitable to model perception. Nevertheless, since they allow for separate inspection of pose- and dynamics-parameters, it is possible to conclude that a faithful model of the dynamics is important for perception, while the pose model is less relevant.

### 3.1 LIMITATIONS

I chose a two alternative forced choice paradigm to measure the perceived naturalness of movement. This yields only one bit of information per stimulus presentation. Furthermore I have tested three MP Types with many different complexity parameter values. This requires a large amount of stimulus presentation to obtain reliable estimates, which is feasible if the collected data can be pooled over participants. However, the data indicates that inter-individual differences exist. Increasing the number of presentations per participant is infeasible due to participant fatigue.

### 3.2 FUTURE DIRECTIONS

One straight-forward direction of my investigations is to test if the advantage of TMP- compared to other models still holds for different movements. Furthermore, the inclusion of more generative models in the experimental framework would yield further insight of the perceptual representation of biological movement. The online version of the experiment promises to enable this large scale testing of many models and many movements.

Eye-tracking data might account for attentional effects which could potentially confound the results. This would, in principle, not prohibit crowd-sourcing the experiment, as a web-cam based setup might already be sufficient (Papoutsaki, Laskey, and Huang 2017). Furthermore, for the analysis of the data, multi-level models might account for some of the previously discussed inter-individual differences.

It would be ideal to devise a complete model for perception and action. One interesting approach towards this goal is to learn a latent representation of natural movement stimuli. In work not included in this thesis, I have shown that the conditional variational auto-encoder (Sohn, Lee, and Yan 2015) can be used to predict pose from images for simple two-dimensional objects. Including temporal dependencies in this model could provide the link to the kinematic level I have investigated in this thesis.

The model for pose tracking might provide an alternative application as stimulus generator: by conditioning the model on the movement kinematics, the corresponding visual stimulus can be sampled. This stimulus would be as random as possible, while still respecting natural image statistics, thus advancing the ecological validity of biological movement stimuli used so far.

### 3.3 CONCLUSION

In this thesis, the main assumption is a shared representation of action and perception. Of the three perceptually evaluated models of motor productions, temporal movement primitives emerge as the most likely candidate for such a representation. While the perceptual experiments can not draw definite conclusions on this assumption, I am confident the research I have provided will stimulate new experiments and refined models for action and perception.

# References

- Bishop, Christopher M. (2007). *Pattern Recognition and Machine Learning*. Springer.
- Box, George EP (1976). “Science and Statistics”. In: *Journal of the American Statistical Association* 71.356, pp. 791–799.
- Cheung, Vincent C. K., Andrea Turolla, Michela Agostini, Stefano Silvoni, Caoimhe Bennis, Patrick Kasi, Sabrina Paganoni, Paolo Bonato, and Emilio Bizzi (2012). “Muscle Synergy Patterns as Physiological Markers of Motor Cortical Damage”. In: *Proceedings of the National Academy of Sciences* 109.36, pp. 14652–14656. ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.1212056109](https://doi.org/10.1073/pnas.1212056109). pmid: [22908288](https://pubmed.ncbi.nlm.nih.gov/22908288/).
- Clever, Debora, Monika Harant, Henning Koch, Katja Mombaur, and Dominik Endres (2016). “A Novel Approach for the Generation of Complex Humanoid Walking Sequences Based on a Combination of Optimal Control and Learning of Movement Primitives”. In: *Robotics and Autonomous Systems* 83, pp. 287–298. ISSN: 0921-8890. DOI: [10.1016/j.robot.2016.06.001](https://doi.org/10.1016/j.robot.2016.06.001).
- D’Avella, Andrea, Laure Fernandez, Alessandro Portone, and Francesco Lacquaniti (2008). “Modulation of Phasic and Tonic Muscle Synergies with Reaching Direction and Speed”. In: *Journal of neurophysiology* 100.3, pp. 1433–1454.
- Deen, Ben, Kami Koldewyn, Nancy Kanwisher, and Rebecca Saxe (2015). “Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus”. In: *Cerebral Cortex (New York, NY)* 25.11, pp. 4596–4609. ISSN: 1047-3211. DOI: [10.1093/cercor/bhv111](https://doi.org/10.1093/cercor/bhv111). pmid: [26048954](https://pubmed.ncbi.nlm.nih.gov/26048954/).
- Di Pellegrino, Giuseppe, Luciano Fadiga, Leonardo Fogassi, Vittorio Gallese, and Giacomo Rizzolatti (1992). “Understanding Motor Events: A Neurophysiological Study”. In: *Experimental brain research* 91.1, pp. 176–180.
- Endres, Dominik and Mike Oram (2010). “Feature Extraction from Spike Trains with Bayesian Binning: ‘Latency Is Where the Signal Starts’”. In: *Journal of Computational Neuroscience* 29.1, pp. 149–169.
- Endres, Dominik M., Enrico Chiovetto, and Martin Giese (2013). “Model Selection for the Extraction of Movement Primitives”. In: *Frontiers in Computational Neuroscience* 7. ISSN: 1662-5188. DOI: [10.3389/fncom.2013.00185](https://doi.org/10.3389/fncom.2013.00185).
- Endres, Dominik M., Yaron Meirovitch, Tamar Flash, and Martin A. Giese (2013). “Segmenting Sign Language into Motor Primitives with Bayesian Binning”. In: *Frontiers in Computational Neuroscience* 7. ISSN: 1662-5188. DOI: [10.3389/fncom.2013.00068](https://doi.org/10.3389/fncom.2013.00068).
- Fechner, Gustav Theodor (1860). *Elemente Der Psychophysik*. Vol. 2. Breitkopf u. Härtel.

- Felleman, Daniel J. and David C. Van Essen (1991). “Distributed Hierarchical Processing in the Primate Cerebral Cortex.” In: *Cerebral cortex (New York, NY: 1991)* 1.1, pp. 1–47.
- Field, David J. (1987). “Relations between the Statistics of Natural Images and the Response Properties of Cortical Cells”. In: *Journal of the Optical Society of America A* 4.12, p. 2379. ISSN: 1084-7529, 1520-8532. DOI: [10.1364/JOSAA.4.002379](https://doi.org/10.1364/JOSAA.4.002379).
- Fogassi, Leonardo, Pier Francesco Ferrari, Benno Gesierich, Stefano Rozzi, Fabian Chersi, and Giacomo Rizzolatti (2005). “Parietal Lobe: From Action Organization to Intention Understanding”. In: *Science* 308.5722, pp. 662–667.
- Friston, Karl (2010). “The Free-Energy Principle: A Unified Brain Theory?” In: *Nature Reviews Neuroscience* 11.2, pp. 127–138. ISSN: 1471-0048. DOI: [10.1038/nrn2787](https://doi.org/10.1038/nrn2787).
- Friston, Karl, Jérémie Mattout, and James Kilner (2011). “Action Understanding and Active Inference”. In: *Biological Cybernetics* 104.1, pp. 137–160. ISSN: 1432-0770. DOI: [10.1007/s00422-011-0424-z](https://doi.org/10.1007/s00422-011-0424-z).
- Gallese, Vittorio, Luciano Fadiga, Leonardo Fogassi, and Giacomo Rizzolatti (1996). “Action Recognition in the Premotor Cortex”. In: *Brain* 119.2, pp. 593–609.
- Geirhos, Robert, Carlos R. Medina Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann (2018). *Generalisation in Humans and Deep Neural Networks*. arXiv: [1808.08750](https://arxiv.org/abs/1808.08750).
- Giese, Martin and Tomaso Poggio (2003). “Neural Mechanisms for the Recognition of Biological Movements: Cognitive Neuroscience”. In: *Nature Reviews Neuroscience* 4.3, pp. 179–192. ISSN: 1471-003X, 1471-0048. DOI: [10.1038/nrn1057](https://doi.org/10.1038/nrn1057).
- Grillner, Sten (2006). “Biological Pattern Generation: The Cellular and Computational Logic of Networks in Motion”. In: *Neuron* 52.5, pp. 751–766. ISSN: 0896-6273. DOI: [10.1016/j.neuron.2006.11.008](https://doi.org/10.1016/j.neuron.2006.11.008).
- He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2015). “Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification”. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034.
- Hickok, Gregory (2009). “Eight Problems for the Mirror Neuron Theory of Action Understanding in Monkeys and Humans”. In: *Journal of Cognitive Neuroscience* 21.7, pp. 1229–1243. ISSN: 0898-929X. DOI: [10.1162/jocn.2009.21189](https://doi.org/10.1162/jocn.2009.21189).
- Hodgins, Jessica K., James F. O’Brien, and Jack Tumblin (1998). “Perception of Human Motion with Different Geometric Models”. In: *IEEE Transactions on Visualization and Computer Graphics* 4.4, pp. 307–316. ISSN: 10772626. DOI: [10.1109/2945.765325](https://doi.org/10.1109/2945.765325).
- Hommel, Bernhard (2019). “Theory of Event Coding (TEC) V2.0: Representing and Controlling Perception and Action”. In: *Attention, Perception, & Psy-*

- chophysics* 81.7, pp. 2139–2154. ISSN: 1943-393X. DOI: [10.3758/s13414-019-01779-4](https://doi.org/10.3758/s13414-019-01779-4).
- Hommel, Bernhard, Jochen Müsseler, Gisa Aschersleben, and Wolfgang Prinz (2001). “The Theory of Event Coding (TEC): A Framework for Perception and Action Planning”. In: *Behavioral and Brain Sciences* 24.05, pp. 849–878. ISSN: 0140-525X. DOI: [10.1017/S0140525X01000103](https://doi.org/10.1017/S0140525X01000103).
- Hubel, David H. and Torsten N. Wiesel (1959). “Receptive Fields of Single Neurons in the Cat’s Striate Cortex”. In: *The Journal of physiology* 148.3, pp. 574–591.
- Hyvärinen, Aapo, Jarmo Hurri, and Patrick O. Hoyer (2009). *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*. Vol. 39. Springer Science & Business Media.
- Ijspeert, Auke Jan, Jun Nakanishi, Heiko Hoffmann, Peter Pastor, and Stefan Schaal (2013). “Dynamical Movement Primitives: Learning Attractor Models for Motor Behaviors”. In: *Neural Computation* 25.2, pp. 328–373. ISSN: 0899-7667, 1530-888X. DOI: [10.1162/NECO\\_a\\_00393](https://doi.org/10.1162/NECO_a_00393).
- Inuggi, Alberto, Claudio Campus, Roberta Vastano, Ghislain Saunier, Alejo Keuroghlanian, and Thierry Pozzo (2018). “Observation of Point-Light-Walker Locomotion Induces Motor Resonance When Explicitly Represented; An EEG Source Analysis Study”. In: *Frontiers in Psychology* 9. ISSN: 1664-1078. DOI: [10.3389/fpsyg.2018.00303](https://doi.org/10.3389/fpsyg.2018.00303).
- James, William (1890). “The Perception of Reality”. In: *Principles of psychology* 2, pp. 283–324.
- Jhuang, Hueihan, Thomas Serre, Lior Wolf, and Tomaso Poggio (2007). “A Biologically Inspired System for Action Recognition”. In: *2007 IEEE 11th International Conference on Computer Vision*. Ieee, pp. 1–8.
- Johansson, Gunnar (1973). “Visual Perception of Biological Motion and a Model for Its Analysis”. In: *Perception & Psychophysics* 14.2, pp. 201–211. ISSN: 0031-5117, 1532-5962. DOI: [10.3758/BF03212378](https://doi.org/10.3758/BF03212378).
- Kilner, James M., Karl J. Friston, and Chris D. Frith (2007). “Predictive Coding: An Account of the Mirror Neuron System”. In: *Cognitive Processing* 8.3, pp. 159–166. ISSN: 1612-4790. DOI: [10.1007/s10339-007-0170-2](https://doi.org/10.1007/s10339-007-0170-2).
- Knill, David C. and Alexandre Pouget (2004). “The Bayesian Brain: The Role of Uncertainty in Neural Coding and Computation”. In: *TRENDS in Neurosciences* 27.12, pp. 712–719.
- Knopp, Benjamin, Dmytro Velychko, Johannes Dreibrodt, and Dominik Endres (2019). “Predicting Perceived Naturalness of Human Animations Based on Generative Movement Primitive Models”. In: *ACM Trans. Appl. Percept.* 16.3, 15:1–15:18. ISSN: 1544-3558. DOI: [10.1145/3355401](https://doi.org/10.1145/3355401).
- Knopp, Benjamin, Dmytro Velychko, Johannes Dreibrodt, Alexander C. Schütz, and Dominik Endres (2020). “Evaluating Perceptual Predictions Based on Movement Primitive Models in VR- and Online-Experiments”. In: *ACM Symposium*

- on *Applied Perception 2020*. SAP '20: ACM Symposium on Applied Perception 2020. Virtual Event USA: ACM, pp. 1–9. ISBN: 978-1-4503-7618-1. DOI: [10.1145/3385955.3407940](https://doi.org/10.1145/3385955.3407940).
- Koller, Daphne and Nir Friedman (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT press.
- Kording, Konrad P., Gunnar Blohm, Paul Schrater, and Kendrick Kay (2020). *Appreciating the Variety of Goals in Computational Neuroscience*. arXiv: [2002.03211](https://arxiv.org/abs/2002.03211).
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (2012). “Imagenet Classification with Deep Convolutional Neural Networks”. In: *Advances in neural information processing systems* 25, pp. 1097–1105.
- Kutch, Jason J. and Francisco J. Valero-Cuevas (2012). “Challenges and New Approaches to Proving the Existence of Muscle Synergies of Neural Origin”. In: *PLoS Computational Biology* 8.5. ISSN: 1553-734X. DOI: [10.1371/journal.pcbi.1002434](https://doi.org/10.1371/journal.pcbi.1002434). pmid: [22570602](https://pubmed.ncbi.nlm.nih.gov/22570602/).
- Lee, Daniel D. and H. Sebastian Seung (1999). “Learning the Parts of Objects by Non-Negative Matrix Factorization”. In: *Nature* 401.6755 (6755), pp. 788–791. ISSN: 1476-4687. DOI: [10.1038/44565](https://doi.org/10.1038/44565).
- Levine, Ariel J., Christopher A. Hinckley, Kathryn L. Hilde, Shawn P. Driscoll, Tiffany H. Poon, Jessica M. Montgomery, and Samuel L. Pfaff (2014). “Identification of a Cellular Node for Motor Control Pathways”. In: *Nature Neuroscience* 17.4 (4), pp. 586–593. ISSN: 1546-1726. DOI: [10.1038/nn.3675](https://doi.org/10.1038/nn.3675).
- Macpherson, JANE M. (1988). “Strategies That Simplify the Control of Quadrupedal Stance. II. Electromyographic Activity”. In: *Journal of Neurophysiology* 60.1, pp. 218–231.
- Marr, David (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. USA: Henry Holt and Co., Inc. ISBN: 978-0-7167-1567-2.
- Murphy, Kevin P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT press.
- Mussa-Ivaldi, Ferdinando A. and Sara A. Solla (2004). “Neural Primitives for Motion Control”. In: *IEEE Journal of Oceanic Engineering* 29.3, pp. 640–650.
- Nelissen, Koen, Elena Borra, Marzio Gerbella, Stefano Rozzi, Giuseppe Luppino, Wim Vanduffel, Giacomo Rizzolatti, and Guy A. Orban (2011). “Action Observation Circuits in the Macaque Monkey Cortex”. In: *Journal of Neuroscience* 31.10, pp. 3743–3756. ISSN: 0270-6474, 1529-2401. DOI: [10.1523/JNEUROSCI.4803-10.2011](https://doi.org/10.1523/JNEUROSCI.4803-10.2011). pmid: [21389229](https://pubmed.ncbi.nlm.nih.gov/21389229/).
- Nelissen, Koen, Giuseppe Luppino, Wim Vanduffel, Giacomo Rizzolatti, and Guy A. Orban (2005). “Observing Others: Multiple Action Representation in the Frontal Lobe”. In: *Science* 310.5746, pp. 332–336. ISSN: 0036-8075, 1095-9203. DOI: [10.1126/science.1115593](https://doi.org/10.1126/science.1115593). pmid: [16224029](https://pubmed.ncbi.nlm.nih.gov/16224029/).

- Olshausen, Bruno A. and David J. Field (1996). “Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images”. In: *Nature* 381.6583, p. 607.
- Omlor, Lars and Martin A. Giese (n.d.). “Anechoic Blind Source Separation Using Wigner Marginals”. In: (), p. 38.
- Oram, M. W. and D. I. Perrett (1996). “Integration of Form and Motion in the Anterior Superior Temporal Polysensory Area (STPa) of the Macaque Monkey”. In: *Journal of neurophysiology* 76.1, pp. 109–129.
- Papoutsaki, Alexandra, James Laskey, and Jeff Huang (2017). “SearchGazer: Webcam Eye Tracking for Remote Studies of Web Search”. In: *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval. CHIIR '17*. New York, NY, USA: Association for Computing Machinery, pp. 17–26. ISBN: 978-1-4503-4677-1. DOI: [10.1145/3020165.3020170](https://doi.org/10.1145/3020165.3020170).
- Peng, Yujia, Hannah Lee, Tianmin Shu, and Hongjing Lu (2021). “Exploring Biological Motion Perception in Two-Stream Convolutional Neural Networks”. In: *Vision Research* 178, pp. 28–40. ISSN: 0042-6989. DOI: [10.1016/j.visres.2020.09.005](https://doi.org/10.1016/j.visres.2020.09.005).
- Perrett, D. I., P. A. J. Smith, A. J. Mistlin, A. J. Chitty, A. S. Head, D. D. Potter, R. Broennimann, A. D. Milner, and Ma A. Jeeves (1985). “Visual Analysis of Body Movements by Neurones in the Temporal Cortex of the Macaque Monkey: A Preliminary Report”. In: *Behavioural brain research* 16.2-3, pp. 153–170.
- Prinz, Wolfgang (1997). “Perception and Action Planning”. In: *European Journal of Cognitive Psychology* 9.2, pp. 129–154. ISSN: 0954-1446, 1464-0635. DOI: [10.1080/713752551](https://doi.org/10.1080/713752551).
- Rao, Rajesh P. and Dana H. Ballard (1999). “Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects”. In: *Nature Neuroscience* 2.1, pp. 79–87. ISSN: 1097-6256, 1546-1726. DOI: [10.1038/4580](https://doi.org/10.1038/4580).
- Rasmussen, Carl Edward and Christopher K. I. Williams (2006). *Gaussian Processes for Machine Learning*. MIT Press.
- Riesenhuber, Maximilian and Tomaso Poggio (1999). “Hierarchical Models of Object Recognition in Cortex”. In: *Nature Neuroscience* 2.11, pp. 1019–1025. ISSN: 1546-1726. DOI: [10.1038/14819](https://doi.org/10.1038/14819).
- Schmidhuber, Jürgen (2010). “Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010)”. In: *IEEE Transactions on Autonomous Mental Development* 2.3, pp. 230–247.
- Schütz-Bosbach, Simone and Wolfgang Prinz (2007). “Prospective Coding in Event Representation”. In: *Cognitive Processing* 8.2, pp. 93–102. ISSN: 1612-4790. DOI: [10.1007/s10339-007-0167-x](https://doi.org/10.1007/s10339-007-0167-x).
- Serre, Thomas (2019). “Deep Learning: The Good, the Bad, and the Ugly”. In: *Annual Review of Vision Science* 5.1, pp. 399–426. DOI: [10.1146/annurev-vision-091718-014951](https://doi.org/10.1146/annurev-vision-091718-014951). pmid: [31394043](https://pubmed.ncbi.nlm.nih.gov/31394043/).

- Simonyan, Karen and Andrew Zisserman (2014). *Two-Stream Convolutional Networks for Action Recognition in Videos*. arXiv: [1406.2199 \[cs\]](#).
- Sohn, Kihyuk, Honglak Lee, and Xinchun Yan (2015). “Learning Structured Output Representation Using Deep Conditional Generative Models”. In: p. 9.
- Spoerer, Courtney, Tim Kietzmann, Johannes Mehrer, Ian Charest, and Nikolaus Kriegeskorte (2020). “Recurrent Neural Networks Can Explain Flexible Trading of Speed and Accuracy in Biological Vision”. In: *PLOS Computational Biology* 16, e1008215. DOI: [10.1371/journal.pcbi.1008215](#).
- Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus (2013). *Intriguing Properties of Neural Networks*. arXiv: [1312.6199](#).
- Theusner, Stefanie, Marc H. E. de Lussanet, and Markus Lappe (2011). “Adaptation to Biological Motion Leads to a Motion and a Form Aftereffect”. In: *Attention, Perception, & Psychophysics* 73.6, pp. 1843–1855. ISSN: 1943-393X. DOI: [10.3758/s13414-011-0133-7](#).
- Theusner, Stefanie, Marc de Lussanet, and Markus Lappe (2014). “Action Recognition by Motion Detection in Posture Space”. In: *Journal of Neuroscience* 34.3, pp. 909–921. ISSN: 0270-6474, 1529-2401. DOI: [10.1523/JNEUROSCI.2900-13.2014](#). pmid: [24431449](#).
- Thompson, Emma L., Geoffrey Bird, and Caroline Catmur (2019). “Conceptualizing and Testing Action Understanding”. In: *Neuroscience & Biobehavioral Reviews* 105, pp. 106–114. ISSN: 0149-7634. DOI: [10.1016/j.neubiorev.2019.08.002](#).
- Tresch, Matthew C., Vincent CK Cheung, and Andrea d’Avella (2006). “Matrix Factorization Algorithms for the Identification of Muscle Synergies: Evaluation on Simulated and Experimental Data Sets”. In: *Journal of neurophysiology* 95.4, pp. 2199–2212.
- Tresch, Matthew C. and Anthony Jarc (2009). “The Case for and against Muscle Synergies”. In: *Current Opinion in Neurobiology*. Motor Systems • Neurology of Behaviour 19.6, pp. 601–607. ISSN: 0959-4388. DOI: [10.1016/j.conb.2009.09.002](#).
- Troje, Nikolaus F. (2002). “Decomposing Biological Motion: A Framework for Analysis and Synthesis of Human Gait Patterns”. In: *Journal of Vision* 2.5, pp. 2–2. ISSN: 1534-7362. DOI: [10.1167/2.5.2](#).
- Troje, Nikolaus F. and Dorita HF Chang (2013). “Shape-Independent Processing of Biological Motion”. In: *People watching: Social, perceptual, and neurophysiological studies of body perception*, pp. 82–100.
- Troje, Nikolaus F. and Cord Westhoff (2006). “The Inversion Effect in Biological Motion Perception: Evidence for a “Life Detector”?” In: *Current Biology* 16.8, pp. 821–824. ISSN: 0960-9822. DOI: [10.1016/j.cub.2006.03.022](#).
- Velychko, Dmytro, Dominik Endres, Nick Taubert, and Martin A. Giese (2014). “Coupling Gaussian Process Dynamical Models with Product-of-Experts Ker-

- nels.” In: *Proceedings of the 24th International Conference on Artificial Neural Networks, LNCS 8681*. Springer, pp. 603–610.
- Velychko, Dmytro, Benjamin Knopp, and Dominik Endres (2017). “The Variational Coupled Gaussian Process Dynamical Model”. In: *International Conference on Artificial Neural Networks*. Springer, pp. 291–299. DOI: [10.1007/978-3-319-68600-4\\_34](https://doi.org/10.1007/978-3-319-68600-4_34).
- Velychko, Dmytro, Benjamin Knopp, and Dominik M. Endres (2018). “Making the Coupled Gaussian Process Dynamical Model Modular and Scalable with Variational Approximations”. In: *Entropy* 20.10, p. 724. DOI: [10.3390/e20100724](https://doi.org/10.3390/e20100724).
- Viviani, P. and N. Stucci (1992). “Biological Movements Look Uniform: Evidence of Motor-Perceptual Interactions”. In: *Journal of Experimental Psychology: Human Perception and Performance* 18.3, pp. 603–623.
- Viviani, Paolo and Tamar Flash (1995). “Minimum-Jerk, Two-Thirds Power Law, and Isochrony: Converging Approaches to Movement Planning.” In: *Journal of Experimental Psychology: Human Perception and Performance* 21.1, p. 32.
- Wang, Jack Meng-Chieh, David J. Fleet, and Aaron Hertzmann (2008). “Gaussian Process Dynamical Models for Human Motion”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.2, pp. 283–298. ISSN: 0162-8828. DOI: [10.1109/TPAMI.2007.1167](https://doi.org/10.1109/TPAMI.2007.1167).
- Yamins, Daniel L. K., Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo (2014). “Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex”. In: *Proceedings of the National Academy of Sciences* 111.23, pp. 8619–8624. ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.1403112111](https://doi.org/10.1073/pnas.1403112111). pmid: [24812127](https://pubmed.ncbi.nlm.nih.gov/24812127/).
- Zeiler, Matthew D. and Rob Fergus (2014). “Visualizing and Understanding Convolutional Networks”. In: *European Conference on Computer Vision*. Springer, pp. 818–833.



## Author contributions

### A.1 “THE VARIATIONAL COUPLED GAUSSIAN PROCESS DYNAMICAL MODEL”

Dmytro Velychko, Benjamin Knopp, and Dominik Endres (2017). “The Variational Coupled Gaussian Process Dynamical Model”. In: *International Conference on Artificial Neural Networks*. Springer, pp. 291–299. DOI: [10.1007/978-3-319-68600-4\\_34](https://doi.org/10.1007/978-3-319-68600-4_34)

- Dmytro Velychko: designed research, derived and implemented the model, collected and processed motion-capture data
- Benjamin Knopp: implemented and ran experiment, analyzed experimental data, wrote the manuscript, **contribution: 30%**
- Dominik Endres: designed research, supervised, wrote the manuscript

### A.2 “MAKING THE COUPLED GAUSSIAN PROCESS DYNAMICAL MODEL MODULAR AND SCALABLE WITH VARIATIONAL APPROXIMATIONS”

Dmytro Velychko, Benjamin Knopp, and Dominik M. Endres (Oct. 2018). “Making the Coupled Gaussian Process Dynamical Model Modular and Scalable with Variational Approximations”. In: *Entropy* 20.10, p. 724. DOI: [10.3390/e20100724](https://doi.org/10.3390/e20100724)

- Dmytro Velychko: designed research, derived and implemented the model, collected and processed motion-capture data
- Benjamin Knopp: implemented and ran experiment, analyzed experimental data, wrote the manuscript, **contribution: 30%**
- Dominik Endres: designed research, supervised, wrote the manuscript

### A.3 “PREDICTING PERCEIVED NATURALNESS OF HUMAN ANIMATIONS BASED ON GENERATIVE MOVEMENT PRIMITIVE MODELS”

Benjamin Knopp, Dmytro Velychko, Johannes Dreibrodt, and Dominik Endres (Sept. 2019). “Predicting Perceived Naturalness of Human Animations Based on Generative Movement Primitive Models”. In: *ACM Trans. Appl. Percept.* 16.3, 15:1–15:18. ISSN: 1544-3558. DOI: [10.1145/3355401](https://doi.org/10.1145/3355401)

- Benjamin Knopp: designed research, analyzed experimental data, wrote the manuscript, **contribution:** 50%
- Dmytro Velychko: implemented the models
- Johannes Dreibrodt: implemented and ran experiment
- Dominik Endres: designed research, supervised, wrote the manuscript

### A.4 “EVALUATING PERCEPTUAL PREDICTIONS BASED ON MOVEMENT PRIMITIVE MODELS IN VR- AND ONLINE-EXPERIMENTS”

Benjamin Knopp, Dmytro Velychko, Johannes Dreibrodt, Alexander C. Schütz, et al. (Sept. 12, 2020). “Evaluating Perceptual Predictions Based on Movement Primitive Models in VR- and Online-Experiments”. In: *ACM Symposium on Applied Perception 2020*. SAP ’20: ACM Symposium on Applied Perception 2020. Virtual Event USA: ACM, pp. 1–9. ISBN: 978-1-4503-7618-1. DOI: [10.1145/3385955.3407940](https://doi.org/10.1145/3385955.3407940)

- Benjamin Knopp: designed research, implemented and ran online experiment, analyzed experimental data, wrote the manuscript, **contribution:** 50%
- Dmytro Velychko: implemented the models
- Johannes Dreibrodt: implemented and ran VR experiment
- Alexander C. Schütz: designed research, supervised, wrote the manuscript
- Dominik Endres: designed research, supervised, wrote the manuscript

# B

## The Variational Coupled Gaussian Process Dynamical Model

# The Variational Coupled Gaussian Process Dynamical Model

Dmytro Velychko, Benjamin Knopp, and Dominik Endres

University of Marburg, Department of Psychology,  
Gutenbergstr. 18, 35032 Marburg, Germany  
{dmytro.velychko, benjamin.knopp, dominik.endres}@uni-marburg.de

**Abstract.** We present a full variational treatment of the Coupled Gaussian Process Dynamical Model (CGPDM) with non-marginalized coupling mappings. The CGPDM generates high-dimensional trajectories from coupled low-dimensional latent dynamical models. The deterministic variational treatment obviates the need for sampling and facilitates the use of the CGPDM on larger data sets. The non-marginalized coupling mappings allow for a flexible exchange of the constituent dynamics models at run time. This exchange possibility is crucial for the construction of modular movement primitive models. We test the model against the marginalized CGPDM, dynamic movement primitives and temporal movement primitives, finding that the CGPDM generally outperforms the other models. Human observers can hardly distinguish CGPDM-generated movements from real human movements.

**Keywords:** Gaussian Process, Variational Methods, Movement Primitives, Modularity

This is a preprint of the article  
Velychko, D., Knopp B. and Endres, D.: The Variational Coupled Gaussian Process Dynamical Model. To be published in the Proceedings of the 26th International Conference on Artificial Neural Networks, 1-8 (2017).  
The final publication will be available at DOI: 10.1007/978-3-319-68600-4 or <http://www.springerlink.com>

## 1 Introduction and Related Work

Planning and execution of human full-body movements is a formidable control problem for the brain. Modular movement primitives (MP) have been suggested as a means to simplify this control problem while retaining a sufficient degree of control flexibility for a wide range of task, see [4] for a review. 'Modular' in this context usually refers to the existence of an operation which allows for the combination of (simple) primitives into (complex) movements.

Technical applications of modular MPs have also been devised. For example in computer graphics, especially combined with dynamics models [7] and robotics,

e.g. the dynamical MP (DMP) [9]. Each DMP is encoded by a canonical second order differential equation with guaranteeable stability properties and learnable parameters.

To lift the restriction of canonical dynamics, the Coupled Gaussian Process Dynamical Model (CGPDM) [17] learns both the dynamics mappings and their coupling for a given movement. The learning is accomplished in a Gaussian process framework. The Gaussian process (GP) is a machine learning staple for classification and regression tasks. It can be interpreted as an abstraction of a neural network with a large, possibly infinite, hidden layer. Its advantages include theoretical elegance, tractability and closed-form solutions for posterior densities. It affords high flexibility but has poor (cubic) runtime scaling in the data set size. We improve this scaling with deterministic, sparse variational approximations using small sets of inducing points (IPs) and associated values [16] for each MP, resulting in the 'variational CGPDM' (vCGPDM). This yields a linear run-time dependence on the number of data points.

The CGPDM builds on the Gaussian process dynamical model (GPDM) [18], where a latent dynamics model is mapped onto observations by functions drawn from a GP. The GPDM can model the variability of human movements [15]. Sparse variational approximations have been developed for GPDM-like architectures [6] and even deep extensions thereof [11]. However, with the exception of the CGPDM, all these approaches have a 'monolithic' latent space(s) and thus lack the modularity of MPs. While deriving a variational approximation is not trivial, we expect it to avoid overfitting and yield a good bound on the marginal likelihood [2].

Our target application here is human movement modeling, but the vCGPDM could be easily applied to other systems where modularized control is beneficial, e.g. humanoid robotics [5].

We introduce the vCGPDM in section 2. In section 3, we first benchmark the vCGPDM against other MP models. Second, we determine the degree of human-tolerable sparseness in a psychophysics experiment. In section 4 we propose future research.

## 2 The model

A CGPDM is basically a number of GPDMs (the 'parts') run in parallel, with coupling between the latent space dynamics. See [17] for a graphical model representation. The model operates in discrete time  $t = 0, \dots, T$ . For every part  $i = 1, \dots, M$  there is a  $Q^i$ -dimensional latent space with second-order autoregressive dynamics and inputs from the latent spaces of the other parts. Let  $\mathbf{x}_t^i \in \mathbb{R}^{Q^i}$  be the state of latent space  $i$  at time  $t$ . Then

$$\mathbf{x}_t^i = \mathbf{f}^i(\mathbf{x}_{t-2}^1, \mathbf{x}_{t-1}^1, \dots, \mathbf{x}_{t-2}^M, \mathbf{x}_{t-1}^M). \quad (1)$$

We chose a second-order model, because our target application is human movement modeling, and the literature indicates (e.g. [15]) that this is a good choice for this task. However, we note that this can be easily changed in the model. The

latent states  $\mathbf{x}_t^i$  give rise to  $D^i$ -dimensional observations  $\mathbf{y}_t^i \in \mathbb{R}^{D^i}$  via functions  $\mathbf{g}^i(\cdot)$  plus isotropic Gaussian noise  $\eta_t^i$

$$\mathbf{y}_t^i = \mathbf{g}^i(\mathbf{x}_t^i) + \eta_t^i \quad (2)$$

The functions  $\mathbf{g}^i(\cdot)$  are drawn from a GP prior with zero mean function and a suitable kernel. In a vCGPDM, the functions  $\mathbf{f}^i(\dots)$  are also drawn from a GP prior with zero mean function, and a kernel that is derived with product-of-experts (PoE, [8]) coupling between the latent spaces of the different parts, as described by [17]: each part generates a Gaussian prediction about every part (i.e. including itself). Let  $\mathbf{x}_t^{i,j} = \mathbf{f}^{i,j}(\mathbf{x}_{t-2}^i, \mathbf{x}_{t-1}^i)$  be the mean of the prediction of part  $i$  about part  $j$  at time index  $t$ , and  $\alpha^{i,j}$  its variance. Following the standard PoE construction of multiplying the densities of the individual predictions and re-normalizing, one finds

$$p(\mathbf{x}_t^j | \mathbf{x}_t^{i,j}, \alpha^{i,j}) = \frac{\exp \left[ -\frac{1}{2\alpha^j} \left( \mathbf{x}_t^j - \alpha^j \sum_i \frac{\mathbf{x}_t^{i,j}}{\alpha^{i,j}} \right)^2 \right]}{(2\pi\alpha^j)^{\frac{Q^j}{2}}} \propto \prod_i \mathcal{N} \left( \mathbf{x}_t^j | \mathbf{x}_t^{i,j}, \alpha^{i,j} \right) \quad (3)$$

where  $\alpha^j = \left( \sum_i \alpha_{i,j}^{-1} \right)^{-1}$ . It was shown in [17] that the individual predictions  $\mathbf{x}_t^{i,j}$  can be marginalized out in closed form. We will keep the individual predictions, because this allows us to couple a previously learned dynamics model for a part (including its predictions about the other parts) to any other dynamics model for the other parts, thus obtaining a modular MP model.

The form of eqn. 3 indicates the function of the coupling variances: the smaller a given variance, the more important the prediction of the generating part. When the  $\alpha^{i,j}$  are optimized during learning, the model is able to discover which couplings are important for predicting the data, and which ones are not, see [17]. Put differently, if an  $\alpha^{i,j}$  is small compared to  $\alpha^{i' \neq i, j}$ , then part  $i$  is able to make a prediction about part  $j$  with (relatively) high certainty. Furthermore, as demonstrated in [17], the  $\alpha^{i,j}$  can be modulated after learning to generate novel movements which were not in the training data.

The basic CGPDM exhibits the usual cubic run time scaling with the number of data points, which prohibits learning from large data sets. We therefore developed a sparse variational approximation, following the treatment in [16,11]. We augment the model with IPs  $\mathbf{r}^i$  and associated values  $\mathbf{v}^i$  such that  $g^i(\mathbf{r}^i) = \mathbf{v}^i$  for the latent-to-observed mappings  $g^i(X^i)$  (referred to as 'LVM IPs' in the following), and condition the probability density of the function values of  $g^i(X^i)$  on these points/values, which we assume to be a sufficient statistic. We apply the same augmentation strategy to reduce the computational effort for learning the dynamics mappings, which are induced by  $\mathbf{z}^{i,j}$  and  $\mathbf{u}^{i,j}$  (referred to as 'dynamics IP').

**Key assumption of the vCGPDM:** to obtain a tractable variational posterior distribution  $q$  over the latent states  $\mathbf{x}_t^i = (x_{t,1}^i, \dots, x_{t,Q^i}^i)$ , we choose a distribution that factorizes across time steps  $0, \dots, T$ , parts  $1, \dots, M$  and

dimensions  $1, \dots, Q^i$  within parts, and assume that the individual distributions are Gaussian:

$$q(\mathbf{x}_0^1, \dots, \mathbf{x}_T^M) = \prod_{t=0}^T \prod_{i=1}^M \prod_{q=1}^{Q^i} q(\mathbf{x}_{t,q}^i); \quad q(\mathbf{x}_{t,q}^i) = \mathcal{N}(\mu_{t,q}^i, \sigma_{t,q}^{2,i}). \quad (4)$$

This approximation assumption is clearly a gross simplification of the correct latent state posterior. However, it allows us to make analytical progress: a free-energy evidence lower bound, ELBO (see eqn. 8 of [16] and eqn. S20 in the online supplementary material<sup>1</sup>) can now be computed in closed form if we choose the right kernels for the GPs. We opt for an ARD (automatic relevance detection) squared exponential kernel [3] for every part- $i$ -to- $j$  prediction GP:

$$k^{i,j}(\mathbf{X}, \mathbf{X}') = \exp\left(-\frac{1}{2} \sum_q^{Q^i} \frac{(\mathbf{X}_q - \mathbf{X}'_q)^2}{\lambda_q^{i,j}}\right). \quad (5)$$

and a radial basis function kernel for the latent-to-observed mappings. The computations yielding the ELBO are lengthy (and error-prone) but straightforward. The details can be found in section 2 of the online supplementary material. Whether our simplistic approximation assumption (eqn. 4) is useful depends on the data, but at least for human movement it seems appropriate (see section 3).

### 3 Results

We implemented the model in `Python 2.7` using the machine-learning framework `Theano` [1] for automatic differentiation to enable gradient-based maximization of the ELBO with the `scipy.optimize.fmin_l_bfgs_b` routine [10]. Latent space trajectories were initialized with PCA.

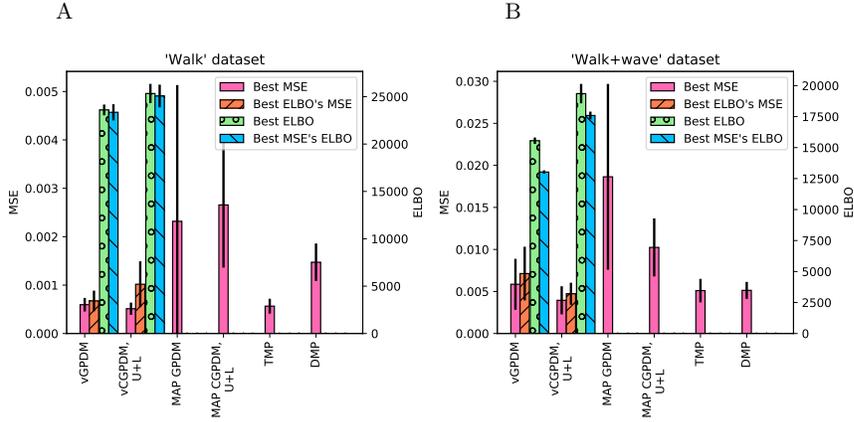
While the sparse approximations in the vCGPDM greatly reduce the memory consumption of the model, they might also introduce errors. Also, our fully factorized latent posterior approximation (eqn. 4) might be too simple. We tried to quantify these errors in a cross-validators model comparison, and in a human perception experiment.

#### 3.1 Human movement data

Comparisons were carried out on human movement data. We recorded these data with a 10-camera PhaseSpace Impulse motion capture system, mapped them onto a skeleton with 19 joints and computed joint angles in angle-axis representation, yielding a total of 60 degrees of freedom. The actors were instructed to walk straight with a natural arm swing, and to walk while waving both arms. Five walking-only and four walking+waving sequences each were used to train the models.

<sup>1</sup> available at <http://uni-marburg.de/wk8Vf>

### 3.2 MAP is worse than variational approximation



**Fig. 1.** Model comparison results. Shown is the average squared kinematics error on held-out data after dynamic time warping (MSE) and the variational lower bound on the model evidence (ELBO), where available. Error bars are standard errors of the mean. **A:** walking dataset. **B:** walking+waving dataset. For model descriptions and further details, see text.

To check how the predictive quality is affected by our sparse variational approximation, we conducted a comparison by five/four-fold cross-validation of the following models for walking/walking+waving. Our cross-validation score is the kinematics mean squared error (MSE), computed after dynamic time warping [14] of trajectories generated by initializing the model to the first two frames of a held-out trial onto the complete held-out trial: 1.) a GPDM with maximum-a-posteriori (MAP) estimation of the latent variables [18], called MAP GPDM in fig. 1. 2.) a fully marginalized two-part (upper/lower body) CGPDM with MAP estimation of the latent variables [17], called MAP CGPDM U+L. 3.) Their variational counterparts, vCGPDM U+L and vGPDM. We experimented with # LVM IPs= 4, . . . , 30, and # dynamics IPs= 2, . . . , 30. The MSE optima were near 10-15 IPs for both. All latent spaces were three-dimensional. 4.) Temporal movement primitives (instantaneous linear mixtures of functions of time) [5]. We used up to 10 primitives, the MSE optimum was located at  $\approx 6$ . 5.) Dynamical movement primitives (DMP) [9]. We used between 1-50 basis functions, the lowest MSE was found at  $\approx 15$ .

The results are plotted in fig. 1. Generally, all models perform better on the walking only dataset, than on walking+waving. This might be due to the latter being a more complex movement, as can be seen in the movie `modular_primitives.avi` in the online supplementary material. Of all tested models, the 2-part vCGPDM performs best in terms of MSE. It is significantly

better than the full-capacity (no IPs) MAP models, i.e. the development of a variational approximation which needs to store only  $\approx 10$  IPs rather than  $\approx 10^4$  data points was well worth the effort. Furthermore, note that the Best ELBO’s MSE (i.e. the MSE at the maximum of the ELBO w.r.t the #IPs) is a fairly good predictor of the best MSE, which indicates that our simple variational approximation is useful for model selection via ELBO. Further evidence for this is shown in fig. 1 of section 4 in the online supplementary material: we plotted MSE vs. ELBO for the vCGPDM U+L, symbols indicate different # LVM IPs. The negative correlation between MSE and ELBO is clearly visible. Furthermore, timing results for the vCGPDM can found in section 5 of the supplement, confirming the theoretical expectations of linear learning time scaling in the data set size for the vCGPDM.

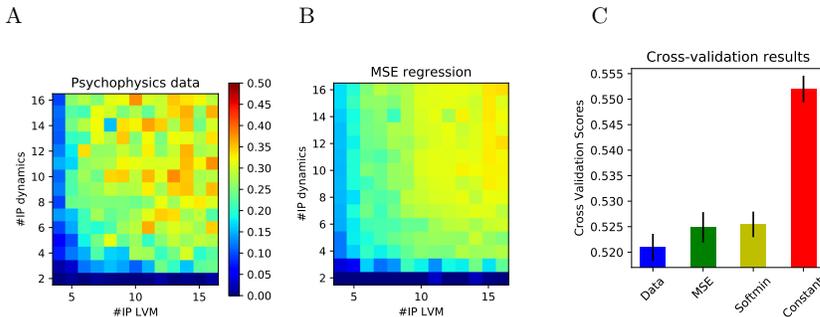
Note that the vCGPDM U+L outperforms the vGPDM particularly on the ‘walking+waving’ dataset. This shows the usefulness of having modular, coupled dynamics models when the (inter)acting (body)parts execute partially independent movements. A visual demonstration of that modularity can be found in the video `modular_primitives.avi` in the online supplementary material.

### 3.3 A small number of IPs is enough to fool human observers

Next, we investigated the number of inducing points needed for perceptually plausible movements with a psychophysical experiment: We showed human observers ( $n = 31$ , 10 male, mean age:  $23.8 \pm 3.5$ a) videos of natural and artificial movements side-by-side on a computer screen. The artificial movements were generated by the vCGPDM U+L. After presentation, the participants had to choose the movement which they perceived as more natural. Examples of stimuli are provided in the online supplementary material in the movie `example_stimuli.mov`. The walking sequences used for training and 9 additional walking sequences were used as natural stimuli. Each subject completed 1170 trials in randomized sequence, judging all artificial stimuli. We also tested for stimulus memorization effects via catch trials with previously unused natural movements in the last quarter of the experiment, finding none. All experimental procedures were approved by the local ethics commission.

Results are shown in fig. 2, A: we computed the frequency  $f_{gen}$  of choosing the vCGPDM-generated movement across all subjects as a function of the number of dynamics IPs and the number of LVM IPs. At best, we might expect  $f_{gen}$  to approach 0.5 when the generated movements are indistinguishable from the natural ones. We fitted those data with a logistic sigmoid  $\frac{1}{1+\exp(a \cdot r(\cdot)+c)}$  and a Bernoulli observation model, using two different regressor functions  $r(\cdot)$ : a soft-minimum between the number of IPs and the MSE. Panel B shows the fit of  $f_{gen}$  with MSE, panel C shows 107-fold crossvalidation results for the two regressors, using the average negative log-probability on the held-out data as score. Error bars are standard deviations. ‘Constant’ is the constant regressor, any other regressor should predict better. ‘Data’ uses the data mean of the individual #IP combinations as a predictor, and constitutes a lower bound on the cross validation score.

Clearly,  $f_{gen}$  increases with the number of IPs, approaching (but not quite reaching) 0.5 for a sufficiently large number of IPs, this is true for the MSE regression, too. Hence, MSE is a good predictor of perceptual performance. Furthermore, a rather small number of IPs is sufficient for modeling this data. This allows for compactly parametrized MPs.



**Fig. 2.** Perceived naturalness of the model, as a function of the number of inducing points (#IP) **A:** Rate of perceiving vCGPDM-generated stimulus as more natural than natural stimulus, averaged across all participants. **B:** Regression of data in panel A, MSE as regressor and logistic sigmoid as psychometric function. **C:** Regression model comparison with 107-fold cross-validation. Softmin and MSE perform comparably well. Both are close to optimal.

## 4 Conclusion

We developed a full variational approximation of the CGPDM, the vCGPDM, which obviates the need for sampling the latent space trajectories [6]. We demonstrated that the vCGPDM with a small number of IPs performs better than the full-capacity CGPDM with a MAP approximation to the latent states, and that the vCGPDM is also able to outperform other contemporary MP models, most likely due to its learnable dynamics. Next, we showed that it produces perceptually believable full-body movements. While perceptual evaluations of full and sparse GPDM-like models [15] have been done before, we are the first to investigate systematically the number of IPs of all model components required for perceptual plausibility. Furthermore, we showed that the MSE and the number of IPs can be used to predict average human classification performance almost optimally. This indicates that the model selection process on large databases of training movements for the model could possibly be automated.

We are now in a position to learn a large library of movements with a CGPDM, and study its compositionality. This is possible due to the compact representation of each MP. Instead of direct connections between parts in the vCGPDM, it is also conceivable to embed the parts into a hierarchical architecture, like [15].

While the vCGPDM is suitable when the number of parts is relatively small (computational complexity  $\mathcal{O}(T * M * (M * \#IP)^3)$  per optimization iteration), a hierarchical architecture might enable more computational savings for many parts. A further direction of future research are *sensorimotor* primitives, i.e. MPs that can be conditioned on sensory input [12,11,13] which we will implement by adding sensory predictions to the latent-to-observed mappings. **Acknowledgements** DFG-IRTG 1901 'The Brain in Action', DFG-SFB-TRR 135 project C06. We thank Olaf Haag for help with rendering the movies, and Björn Büdenbender for assistance with MoCap.

## References

1. Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I.J., Bergeron, A., Bouchard, N., Bengio, Y.: Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS Workshop (2012)
2. Bauer, M., van der Wilk, M., Rasmussen, C.: Understanding probabilistic sparse Gaussian process approximations. Tech. rep., arXiv:1606.04820 (2016)
3. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006)
4. Bizzi, E., Cheung, V., d'Avella, A., Saltiel, P., Tresch, M.: Combining modules for movement. *Brain Res. Rev.* 57(1), 125 – 133 (2008)
5. Clever, D., Harant, M., Koch, K.H., Mombaur, K., Endres, D.M.: A novel approach for the generation of complex humanoid walking sequences based on a combination of optimal control and learning of movement primitives. *Rob. Aut. Sys.* 83, 287–298 (2016), doi: 10.1016/j.robot.2016.06.001
6. Frigola, R., Chen, Y., Rasmussen, C.: Variational Gaussian process state-space models. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) *Advances in NIPS 27*, pp. 3680–3688 (2014)
7. Giese, M.A., Mukovskiy, A., Park, A.N., Omlor, L., Slotine, J.J.E.: Real-Time Synthesis of Body Movements Based on Learned Primitives. In: Cremers D., Rosenhahn B., Yuille A. L. (eds): *Statistical and Geometrical Approaches to Visual Motion Analysis, LNCS 5604*, 107–127 (2009)
8. Hinton, G.E.: Products of experts. In: *Proc. ICANN'99*. vol. 1, pp. 1–6 (1999)
9. Ijspeert, A.J., Nakanishi, J., Hoffmann, H., Pastor, P., Schaal, S.: Dynamical movement primitives: Learning attractor models for motor behaviors. *Neu. Comp.* 25(2), 328–373 (2013)
10. Jones, E., Oliphant, T., Peterson, P., et al.: SciPy: Open source scientific tools for Python (2001–), <http://www.scipy.org/>, [Online; accessed 2015-10-09]
11. Mattos, C.L.C., Dai, Z., Damianou, A., Forth, J., Barreto, G.A., Lawrence, N.D.: Recurrent Gaussian processes. Tech. rep., arXiv:1511.06644 (2016)
12. Paraschos, A., Daniel, C., Peters, J., Neumann, G.: Probabilistic movement primitives. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (eds.) *Advances in NIPS 26*, pp. 2616–2624 (2013)
13. Pastor, P., Kalakrishnan, M., Righetti, L., Schaal, S.: Towards associative skill memories. In: *IEEE-RAS Conf. Humanoids*. pp. 309–315 (2012)
14. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech Sig. Proc.* 26(1), 43–49 (Feb 1978)

15. Taubert, N., Christensen, A., Endres, D., Giese, M.: Online simulation of emotional interactive behaviors with hierarchical Gaussian process dynamical models. In: Proc. ACM SAP. pp. 25–32. ACM (2012)
16. Titsias, M.K., Lawrence, N.D.: Bayesian Gaussian process latent variable model. In: Proc. 13th AISTATS. pp. 844–851 (2010)
17. Velychko, D., Endres, D., Taubert, N., Giese, M.A.: Coupling Gaussian process dynamical models with product-of-experts kernels. In: Proc. 24th ICANN, LNCS 8681, pp. 603–610. Springer (2014)
18. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian process dynamical models for human motion. *IEEE Trans. Pattern Anal. Mach. Intell.* 30(2), 283–298 (2008)



Making the Coupled Gaussian Process  
Dynamical Model Modular and Scalable  
with Variational Approximations

Article

# Making the Coupled Gaussian Process Dynamical Model Modular and Scalable with Variational Approximations <sup>†</sup>

Dmytro Velychko \*, Benjamin Knopp and Dominik Endres \* 

Department of Psychology, University of Marburg, Gutenbergstr. 18, 35032 Marburg, Germany; benjamin.knopp@uni-marburg.de

\* Correspondence: dmytro.velychko@uni-marburg.de (D.V.); dominik.endres@uni-marburg.de (D.E.)

<sup>†</sup> This paper is an extended version of our paper published in the 26th International Conference on Artificial Neural Networks (ICANN 2017), Alghero, Italy, 11–14 September, 2017.

Received: 27 July 2018; Accepted: 20 September 2018; Published: 21 September 2018



**Abstract:** We describe a sparse, variational posterior approximation to the Coupled Gaussian Process Dynamical Model (CGPDM), which is a latent space coupled dynamical model in discrete time. The purpose of the approximation is threefold: first, to reduce training time of the model; second, to enable modular re-use of learned dynamics; and, third, to store these learned dynamics compactly. Our target applications here are human movement primitive (MP) models, where an MP is a reusable spatiotemporal component, or “module” of a human full-body movement. Besides re-usability of learned MPs, compactness is crucial, to allow for the storage of a large library of movements. We first derive the variational approximation, illustrate it on toy data, test its predictions against a range of other MP models and finally compare movements produced by the model against human perceptual expectations. We show that the variational CGPDM outperforms several other MP models on movement trajectory prediction. Furthermore, human observers find its movements nearly indistinguishable from replays of natural movement recordings for a very compact parameterization of the approximation.

**Keywords:** Gaussian processes; variational methods; movement primitives; modularity

---

## 1. Introduction

Two formidable problems that the human brain has to solve are planning and execution of movements of its body. As a means to simplify these problems while keeping a sufficient degree of control flexibility for a wide range of tasks, modular movement primitives (MP) have been suggested (see [1,2] for reviews). There is no universally accepted definition of the term “movement primitive”. For the purposes of this paper, an MP is a spatiotemporal component of a human (full-body) movement that may be produced by mapping a latent state onto observable variables, such as joint angles. The latent state can be generated by dynamical systems [3] or source functions [4,5]. “Modular” usually refers to the existence of an operation which allows for the spatial, temporal or spatiotemporal combination of (simple) primitives into (complex) movements.

Two prominent examples, where this operation is the linear combination of stereotypical time-courses or muscle-coactivations, are called temporal MP-models [6–9] or spatial MP-models [10,11]. While these models are inherently modular, the assumption of stereotyped MPs makes it difficult for a control system built out of these primitives to respond to perturbations. A type of MP which can be controlled on-line more easily is the dynamical MP (DMP) [3], which has been developed for robotics applications. In this approach, each primitive is encoded by a canonical second order differential

equation with guaranteeable stability properties and learnable parameters. A DMP can generate both discrete (e.g., reaching) and rhythmic (e.g., walking) movements and drives the trajectory of one degree of freedom, e.g., a joint angle. Modularity arises because of the latter property, which might be viewed as an “extreme” form of the modularization that we investigate here, where one movement module might affect several degrees of freedom. Similarly, recent extensions of the DMP framework allow for the reuse of a DMP across end-effectors via kinematical mappings [12] or across tasks [13].

We describe a model that learns MPs composed of coupled dynamical systems and associated kinematics mappings, where both components are learned, thus lifting the DMP’s restriction of canonical dynamics. We build on the Coupled Gaussian Process Dynamical Model (CGPDM) by [14], which combines the advantages of modularity and flexibility in the dynamics, at least theoretically. In a CGPDM, the temporal evolution functions for the latent dynamical systems are drawn out of a Gaussian process (GP) prior [15]. These dynamical systems are then coupled probabilistically, and the result is mapped onto observations by functions drawn from another GP. One drawback of the CGPDM is its fully non-parametric nature, which results in cubic scaling (with the dataset size) of learning complexity and quadratic scaling of MP storage size, i.e., the CGPDM can not be learned from large data sets, and its effective parameterization is not compact. We improve both scalability and compactness with deterministic, sparse variational approximations [16]. In this sparse variational CGPDM, each MP is parameterized by a small set of inducing points (IPs) and associated inducing values (IVs), leading to a compact representation with linear scaling of the training complexity in the number of data points, and constant storage requirements. This compactness is important for real-world applicability of the model, since there might be more primitives than muscles (or actuators) across tasks, as pointed out by Bizzi and Cheung [17]: the motor “code” might be sparse and overcomplete, similar to the sparse codes in early vision [18]. Table 1 provides an overview of the key MP models which we compare in this paper.

**Table 1.** Overview of movement primitive models compared in this paper. (v)CGPDM, (variational) coupled Gaussian process dynamical model; (v)GPDM, (variational) Gaussian process dynamical model; TMP, temporal movement primitives; DMP, dynamical movement primitives. Modular, learns reusable MPs. Scalable, below cubic learning complexity with respect to the data set size; Compact, size of the effective parameterization does not grow with the data set size; Canonical dynamics, dynamics model specified before learning; Learned dynamics, dynamics model is a free-form function.

	Modular	Scalable	Compact	Canonical Dynamics	Learned Dynamics
vCGPDM	✓	✓	✓	x	✓
CGPDM	x	x	x	x	✓
vGPDM	x	✓	✓	x	✓
GPDM	x	x	x	x	✓
TMP	✓	✓	✓	x	x
DMP	✓	✓	✓	✓	x

Our target application here is human movement modeling, but the vCGPDM could be easily applied to other systems where modularized control is beneficial, e.g., humanoid robotics [9].

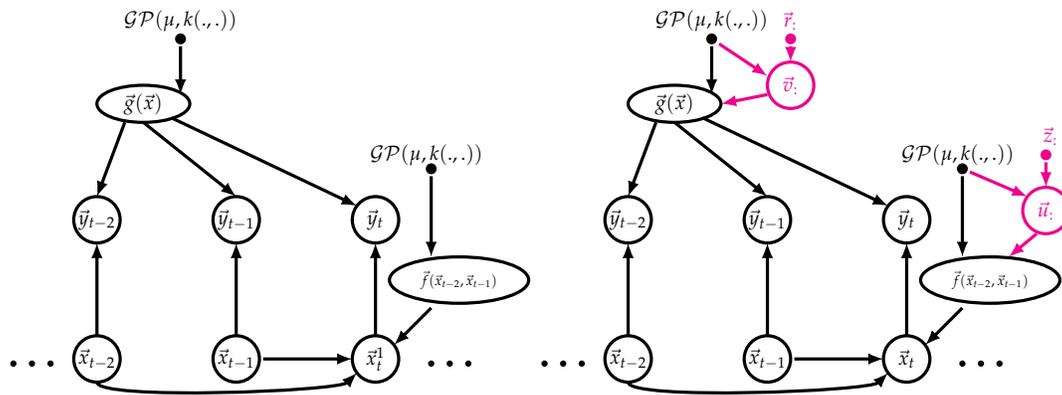
We briefly review related work in Section 2 and introduce the vCGPDM in Section 3. The derivation of the variational approximation is outlined in Section 4. In Section 5, we first illustrate the vCGPDM on artificial data. Second, we benchmark the vCGPDM against other MP models. Third, we perform an experiment to quantify the degree of human-tolerable sparseness in a psychophysics experiment. Fourth, we demonstrate modular movement composition with the vCGPDM. In Section 6, we propose future research directions based on our work.

This paper is a substantially extended version of our earlier conference publication [19].

## 2. Related Work

The Gaussian process (GP) is a machine learning staple for classification and regression tasks [15]. A GP is a prior on functions  $\mathbb{R}^Q \rightarrow \mathbb{R}$  from a  $Q$ -dimensional input space to one-dimensional output. By drawing  $D$  times from the GP, functions from  $\mathbb{R}^Q \rightarrow \mathbb{R}^D$  can be realized. Its advantages include theoretical elegance, tractability and closed-form solutions for posterior densities. Its main disadvantage is cubic runtime scaling with the number of data points. Several solutions have been proposed for this problem. Many of these involve a sparse representation of the posterior process via a small set of IPs, which may [20] or may not be a subset of the data points [21]. If the input space is unobserved, one obtains a GP latent variable model (GPLVM), for which sparse approximations have also been devised [22]. One problem with sparse GP approximations is their tendency to overfit [22], leading to incorrect variance predictions [23]. In that paper, it is also demonstrated that the problem can be alleviated by a variational approximation, which prompted us to develop a similar approach for the CGPDM: as in [24], we extend the sparse GPLVM in time, but we use an autoregressive dynamical system.

If the temporal evolution function of this dynamical system is also drawn from a GP, the resulting model is called Gaussian Process Dynamical Model (GPDM), which can be learned by maximum-a-posteriori approximation if the observed dimension  $D$  is greater than the latent dimension  $Q$  [25]. Figure 1 (left) shows a graphical model representation of the GPDM, and introduces the related notation which we use throughout the paper. Slices “:” indicate collections of variables along one integer index. Multiple slices refer to collections along multiple indices, e.g.,  $\vec{x}_:$  are the latent variables of all parts and time-steps. The GPDM can model the variability of human movements and has been used for computer animation with style control [26–28]. It has also been used with an additional switching prior on the dynamics for motion tracking and recognition [29] and deep variants have been devised [30]. However, with the exception of the coupled GPDM [14,19], all these approaches have a “monolithic” latent space and thus lack the modularity of MPs. One reason for this might be the fact that, for the maximum-a-posteriori approximation to work, the latent space has to be lower-dimensional than the observed space,  $Q \ll D$ . If, as explained above, we want a modular, possibly overcomplete (i.e., the effective  $Q > D$ ) set of MPs, we need learning approaches that are robust to overfitting. The works of Frigola et al. [31] indicate that such approaches may be obtained with variational approximations. In the following, we therefore introduce a variational approximation to CGPDM learning and inference based on an approach similar to Frigola et al. [32], but, as in [30], we aim to obviate the need for sampling altogether to allow for fast, repeatable trajectory generation. While deriving a variational approximation is not trivial, we expect it to avoid overfitting and yield a good bound on the marginal likelihood [33]. Figure 1 (right) shows the graphical model of the GPDM augmented by IPs and IVs. This augmentation yields a tractable variational approximation to the GPDM’s posterior [30].



Notation and Abbreviations (v)GPDM

Gaussian process dynamical model	GPDM	variational approximation to posterior of GPDM	vGPDM
discrete time index	$t = 1, \dots, T$	mean and kernel function	$\mu, k(\cdot, \cdot)$
latent space dimensionality	$Q$	latent states	$\vec{x}_t \in \mathbb{R}^Q$
observed space dimensionality	$D$	observable variables	$\vec{y}_t \in \mathbb{R}^D$
Gaussian process prior	$\mathcal{GP}(\mu, k(\cdot, \cdot))$	Inducing points/values	IP/IV
latent-to-observed function	$\vec{g}(\vec{x}_t)$	dynamics function	$\vec{f}(\vec{x}_{t-2}, \vec{x}_{t-1})$
IP of latent-to-observed function	$\vec{r}_\cdot = (\vec{r}_1, \vec{r}_2, \dots)$	IP of dynamics function	$\vec{z}_\cdot = (\vec{z}_1, \vec{z}_2, \dots)$
IV of latent-to-observed function	$\vec{v}_\cdot = (\vec{v}_1, \vec{v}_2, \dots)$	IV of dynamics function	$\vec{u}_\cdot = (\vec{u}_1, \vec{u}_2, \dots)$

**Figure 1.** Modular building blocks of the vCGPDM. **(Left)** The Gaussian process dynamical model (GPDM). A latent, second order dynamics model generates a time-series of vector-valued random variables  $\vec{x}_t$  which are drawn from a multivariate Gaussian distribution with mean function  $\vec{f}(\vec{x}_{t-2}, \vec{x}_{t-1})$ . The components of this mean function are drawn from a Gaussian Process  $\mathcal{GP}(\mu, k(\cdot, \cdot))$ . Each observable  $\vec{y}_t$  is drawn from multivariate Gaussian distribution with mean function  $\vec{g}(\vec{x}_t)$ , which have a Gaussian process prior, too. **(Right)** The GPDM augmented with inducing points and values for a sparse representation of the posterior process [23]. This enables faster variational Bayesian learning and inference, because the augmented GPs are effectively parameterized by these points (here,  $\vec{r}_\cdot, \vec{z}_\cdot$ ) and corresponding values (here,  $\vec{v}_\cdot, \vec{u}_\cdot$ ) rather than by the full dataset. They may be thought of as prototypical examples of the corresponding functions, e.g.,  $\vec{v}_k = \vec{g}(\vec{r}_k)$ . Slice notation “ $\cdot$ ” indicates collections of variables. For details, see text.

### 3. The Model

The basic building blocks, or “parts”, of the CGPDM are a number of GPDMs run in parallel. In the context of human movement modeling, e.g., they may be thought of as body parts. A part evolves in discrete time  $t = 0, \dots, T$  and is endowed with a  $Q$ -dimensional latent space, a  $D$ -dimensional observed space and second-order autoregressive dynamics described by a function  $\vec{f}(\vec{x}_{t-2}, \vec{x}_{t-1})$ . The component functions  $(\vec{f}(\vec{x}_{t-2}, \vec{x}_{t-1}))_q$  have a Gaussian process prior  $\mathcal{GP}(\mu, k(\cdot, \cdot))$  with mean function  $\mu$  and kernel  $k(\cdot, \cdot)$ . Second-order dynamics seem to be a good choice for our target application of human movement modeling [34], but the order can be easily altered simply by concatenating previous states into one larger vector. Let  $\vec{x}_t \in \mathbb{R}^Q$  the state of latent space of the part at time  $t$  (see Figure 1, left). This latent state produces observations  $\vec{y}_t \in \mathbb{R}^D$  via the function  $\vec{g}(\vec{x}_t)$  as well as isotropic Gaussian noise with variance  $\beta$ . The components  $(\vec{g}(\vec{x}_t))_d$  of this function are drawn from a Gaussian process prior, too. GPDMs can be learned from data via a combination of exact marginalization and maximum-a-posteriori learning of the latent dynamics [25].

While the GPDM is a very expressive model, it suffers from poor runtime and memory scaling with the data set size due to its non-parametric nature, which it inherits from the involved GPs. We remedied this problem by an approach pioneered in [23]: augmenting the GPs with inducing points (IPs, here:  $\vec{r}_\cdot, \vec{z}_\cdot$ ) and associated inducing values (IVs,  $\vec{v}_\cdot, \vec{u}_\cdot$ ) (see Figure 1, right). These IP/IV pairs might

be thought of as prototypical examples of the mappings represented by the corresponding functions, e.g.,  $\bar{v}_k^i = \bar{g}^i(\bar{r}_k^i)$ . Note that the IPs are not drawn from a prior, whereas the IVs are. Hence, the latter are model parameters, whereas the former are not: IPs are merely parameters of the approximation, or “variational parameters”. This augmentation allows for the derivation of a closed-form evidence lower bound (ELBO) on the marginal likelihood of the model.

In a CGPDM, the latent spaces of the parts are coupled to each other. We index parts by superscripts  $i = 1, \dots, M$ . The index notation in such models can be confusing for first-time readers, we provide a notation and index overview in the tables below Figure 2. In our target application, the coupling may reflect the influences which parts of an articulated body have on each other during the execution of a movement. The coupling is implemented by having the parts make Gaussian-distributed predictions  $\bar{x}_t^{i,j}$  about each other’s latent states with means generated by  $M \times M$  many mean coupling functions  $\bar{f}^{i,j}(\bar{x}_{t-2}^i, \bar{x}_{t-1}^i)$  and coupling variances  $\alpha^{i,j}$ .  $i$  indexes the origin part of a coupling, and  $j$  its target. Thus,  $\bar{f}^{i,i}$  refers to the dynamics function for part  $i$ . The components of  $\bar{f}^{i,j}(\bar{x}_{t-2}^i, \bar{x}_{t-1}^i)$  are drawn from GPs. As described in [14], these predictions are combined with a product-of-experts construction (PoE [35]), including the predictions which a part makes about its own future. A product-of-experts construction forces the experts to agree on one prediction (Equation (1), left), which amounts to multiplying the individual predictions (Equation (1), right) and renormalizing (Equation (1), middle):

$$p(\bar{x}_t^j | \bar{f}^{i,j}(\bar{x}_{t-2}^i, \bar{x}_{t-1}^i), \alpha^{i,j}) = \frac{\exp\left[-\frac{1}{2\alpha^j} \left(\bar{x}_t^j - \alpha^j \sum_i \frac{\bar{f}^{i,j}(\bar{x}_{t-2}^i, \bar{x}_{t-1}^i)}{\alpha^{i,j}}\right)^2\right]}{(2\pi\alpha^j)^{\frac{Q^j}{2}}} \propto \prod_i \mathcal{N}\left(\bar{x}_t^j | \bar{f}^{i,j}(\bar{x}_{t-2}^i, \bar{x}_{t-1}^i), \alpha^{i,j}\right) \quad (1)$$

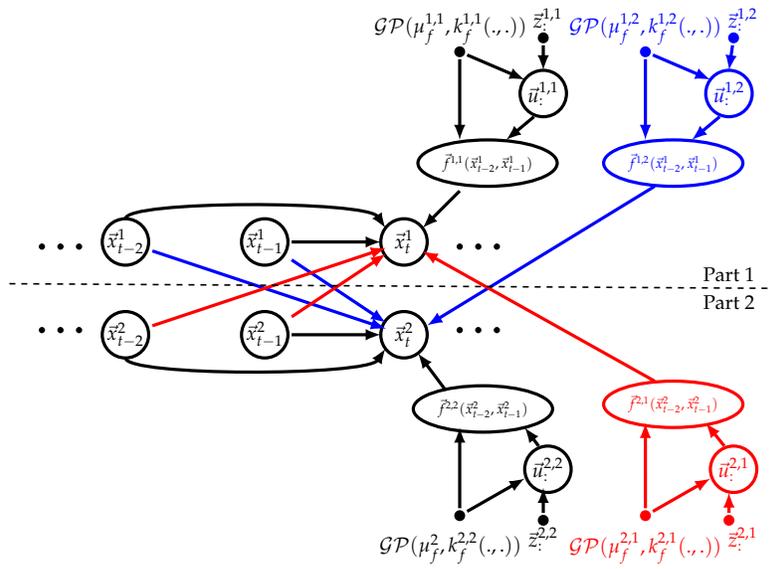
where  $\alpha^j = (\sum_i (\alpha^{i,j})^{-1})^{-1}$ .

To understand the function of the  $\alpha^{i,j}$ , consider the form of Equation (1): the smaller a given variance, the more important the prediction of the generating part. We optimize the  $\alpha^{i,j}$  during learning, letting the model discover which couplings are important for predicting the data. In other words, whenever an  $\alpha^{i,j}$  is small compared to  $\alpha^{i' \neq i, j}$ , then part  $i$  is able to make a prediction about part  $j$  with (relatively) high certainty. Furthermore, the  $\alpha^{i,j}$  can be modulated after learning to generate new movements, as shown below.

In the following, we denote all relevant timesteps before time  $t$  with subscript  $-t$ , e.g.,  $\bar{x}_{-t}^j = (\bar{x}_{t-2}^j, \bar{x}_{t-1}^j)$  for a second-order dynamics model. We showed in [14] that the individual predictions of part  $i$  about part  $j$ ,  $\bar{x}_t^{i,j}$  can be exactly marginalized, leading to a GPDM-like model for each part with a dynamics kernel given by the  $\alpha_{i,j}$ -weighted mean of the individual coupling kernels:

$$k_f^j(\bar{x}_{-t}^1, \bar{x}_{-t}^{1'}, \dots, \bar{x}_{-t}^M, \bar{x}_{-t}^{M'}) = \alpha^{j^2} \sum_{i=1}^M \frac{k_f^{i,j}(\bar{x}_{-t}^i, \bar{x}_{-t}^{i'}, \bar{x}_{-t}^j, \bar{x}_{-t}^{j'})}{\alpha^{i,j^2}} \quad (2)$$

However, doing so results in a model which lacks modularity: after learning, it is difficult to separate the parts from each other, and recombine them for the production of new movements that were not in the training data. We facilitate this modular recombination by restating CGPDM learning such that we can keep an explicit, sparse representation of the coupling functions. Another reason for a sparse representation is that the CGPDM exhibits cubic run time scaling with the data points, which it inherits from the composing GPDMs. To remedy these problems, we follow the treatment in [16,30]: we augment the model with IPs  $\bar{r}_k^i$  and associated IVs  $\bar{v}_k^i$  such that  $g^i(\bar{r}_k^i) = \bar{v}_k^i$  for the latent-to-observed mappings  $g^i(\cdot)$ . Then, we condition the probability density of the function values of  $g^i(\cdot)$  on these IPs/IVs, which we assume to be a sufficient statistic. Likewise, we reduce the computational effort for learning the dynamics and coupling mappings by inducing them through  $\bar{z}_t^{i,j}$  and  $\bar{u}_t^{i,j}$  (also known as “dynamics IPs/IVs”). See Figure 2 for a graphical representation of the augmented model.



Notation and Abbreviations (v)CGPDM

Coupled Gaussian process dynamical model	CGPDM	variational approximation to posterior of CGPDM	vCGPDM
latent states, part $i$ at time $t$	$\bar{x}_t^i \in \mathbb{R}^Q$	Inducing points/values	IP/IV
latent states, part $i$ before $t$	$\bar{x}_{t-1}^i$	latent predictions of part $j$ about part $i$ at time $t$	$\bar{x}_t^{j,i}$
observable variables, part $i$ at time $t$	$\bar{y}_t^i \in \mathbb{R}^D$	Gaussian process prior	$\mathcal{GP}(\mu, k(.,.))$
dynamics function, part $i$	$\bar{f}^{i,i}(\bar{x}_{t-2}^i, \bar{x}_{t-1}^i)$	latent-to-observed function, part $i$	$\bar{g}^i(\bar{x}_t^i)$
coupling function, part $j$ -to- $i$	$\bar{f}^{j,i}(\bar{x}_{t-2}^j, \bar{x}_{t-1}^j)$	coupling variance, part $j$ -to- $i$	$\alpha^{j,i}$
mean and kernel of $\mathcal{GP}$ prior on dynamics function, part $i$	$\mu_f^i, k_f^{i,i}(.,.)$	mean and kernel of $\mathcal{GP}$ prior on latent-to-observed function, part $i$	$\mu_g^i, k_g^{i,i}(.,.)$
IP of latent-to-observed function, part $i$	$\bar{r}^i = (\bar{r}_1^i, \bar{r}_2^i, \dots)$	IP of dynamics function, part $j$ -to- $i$	$\bar{z}_t^{j,i} = (\bar{z}_1^{j,i}, \bar{z}_2^{j,i}, \dots)$
IV of latent-to-observed function, part $i$	$\bar{v}^i = (\bar{v}_1^i, \bar{v}_2^i, \dots)$	IV of dynamics function, part $j$ -to- $i$	$\bar{u}_t^{j,i} = (\bar{u}_1^{j,i}, \bar{u}_2^{j,i}, \dots)$

Index Summary

$\bar{x}_{time}^{part}$	$\bar{u}_{IV-index}^{from-part,to-part}$	$\bar{f}_{time}^{from-part,to-part}$	$\bar{z}_{IP-index}^{from-part,to-part}$
$\bar{g}_{time}^{part}$	$\bar{v}_{IV-index}^{part}$	$\bar{r}_{IP-index}^{part}$	
$d$ -th component of vector $\bar{z}_3^{2,1} : (\bar{z}_3^{2,1})_d$			

**Figure 2.** (Top): Graphical model representation of the augmented Coupled Gaussian Process Dynamical Model (vCGPDM). Shown is a model with two parts, indicated by the superscripts  $i, j \in \{1, 2\}$ . Each part is a vGPDM (see Figure 1), augmented with inducing points  $\bar{z}^{j,i}$  and values  $\bar{u}^{j,i}$  for variational inference and learning, and modular re-composition of learned GPDM components. Observed variables  $\bar{y}_t^i$  and latent-to-observed mappings  $\bar{g}^i(\bar{x}_t^i)$  omitted for clarity. The vGPDMs interact by making predictions about each other’s latent space evolution via functions  $\bar{f}^{j,i}(\bar{x}_{t-2}^j, \bar{x}_{t-1}^j)$ , here  $\bar{f}^{1,2}()$  and  $\bar{f}^{2,1}()$ . Their predictions are product-of-experts combined with the predictions made by each GPDM’s dynamics model (functions  $\bar{f}^{i,i}(\bar{x}_{t-2}^i, \bar{x}_{t-1}^i)$ ). (Bottom): Notation and index summaries.

Besides introducing IPs, computing an ELBO requires a simplifying assumption about the latent state posterior, which is intractable. We choose a posterior distribution  $q$  over the latent states  $\vec{x}_t^i$  that factorizes across time steps  $0, \dots, T$ , parts  $1, \dots, M$  and latent dimensions  $1, \dots, Q^i$  within parts. Furthermore, we assume that the individual distributions are Gaussian:

$$q(\vec{x}_0^1, \dots, \vec{x}_T^M) = \prod_{t=0}^T \prod_{i=1}^M \prod_{q=1}^{Q^i} q((\vec{x}_t^i)_q) ; \quad q((\vec{x}_t^i)_q) = \mathcal{N}(\mu_{t,q}^i, \sigma_{t,q}^{2,i}). \tag{3}$$

While this approximation is clearly a gross simplification of the correct latent state posterior, with the right choice of kernels, an ELBO can now be computed. Our approximation assumption (Equation (3)) seems appropriate for human movement data, see Section 5. Whether it is also useful for other data remains to be determined.

As for a tractable kernel, we decided to use an ARD (automatic relevance determination) squared exponential kernel [36] for every part- $i$ -to- $j$  prediction GP:

$$k^{i,j}(\vec{x}_{-t}^i, \vec{x}_{-t}^{i'}) = \exp \left( -\frac{1}{2} \sum_{t \in -t} \sum_q^{Q^i} \frac{((\vec{x}_t^i)_q - (\vec{x}_t^{i'})_q)^2}{\lambda_q^{i,j,t}} \right). \tag{4}$$

and a radial basis function kernel for the latent-to-observed mappings. Next, we outline the key steps of the derivation of the ELBO.

#### 4. Computing an Evidence Lower Bound for the vCGPDM: An Overview

In this section, we provide an overview of the derivation of the evidence lower bound (ELBO) for the vCGPDM; for details, the reader is referred to Appendix C. We construct a sparse variational approximation by augmenting each of the  $M \times M$  dynamics and coupling mappings  $\vec{f}^{i,j}()$  with IPs and IVs. The variational distribution of the latent variables,  $q(\vec{x};) = q(\vec{x}_1^1, \dots, \vec{x}_T^M)$  factorizes according to Equation (3). We let  $q(\vec{u}^i;)$  and  $q(\vec{v}^i;)$  be unconstrained distributions, which will turn out to be multivariate Gaussians. In the following, we denote the coupling function values at  $t$  with  $\vec{f}_t^{i,j} = f^{i,j}(\vec{x}_{-t}^i)$  and likewise  $\vec{g}_t^i = g^i(\vec{x}_t^i)$ . The factor structure of the joint density of the augmented model follows from the graphical model (see Figures 1 and 2):

$$p(\vec{y};, \vec{g};, \vec{v};, \vec{x};, \vec{f};, \vec{u}; | \vec{z};, \vec{r};) = p(\vec{y}; | \vec{g};) p(\vec{g}; | \vec{x};, \vec{v};, \vec{r};) p(\vec{v}; | \vec{r};) p(\vec{x};, \vec{f}; | \vec{u};, \vec{z};) p(\vec{u}; | \vec{z};) \tag{5}$$

Note that we marginalized (most of) the functions  $f^{i,j}()$  here, keeping only their values at the latent points  $\vec{x};$  and at the IPs. Hence the dependence of  $\vec{g};$  on  $\vec{r};$ . Likewise,  $\vec{f};$  depends on  $\vec{z};$ . For easier notation, we omit spelling out the dependence of the IVs on the IPs in the following. Thus,

$$p(\vec{y}; | \vec{g};) = \prod_{i=1}^M \prod_{d=1}^{D_i} p((\vec{y}^i)_d | (\vec{g}^i)_d) \tag{6}$$

$$p(\vec{g}; | \vec{x};, \vec{v};) = \prod_{i=1}^M \prod_{d=1}^{D_i} p((\vec{g}^i)_d | \vec{x}_t^i, (\vec{v}^i)_d) \tag{7}$$

$$p(\vec{v};) = \prod_{i=1}^M p(\vec{v}^i); \quad p(\vec{u};) = \prod_{i=1}^M \prod_{j=1}^M p(\vec{u}^{i,j}). \tag{8}$$

where Equation (6) follows from the assumption of independent observation noise. Equation (7) is a consequence of the Gaussian process prior on the  $g^i()$ , which makes the components of  $\vec{g}_t^i$  independent. The density of the latent variables and the individual parts' predictions can be factorized as:

$$p(\vec{x};, \vec{f}; | \vec{u};) = p(\vec{x}; | \vec{f};, \vec{u};) p(\vec{f}; | \vec{u};) \tag{9}$$

with

$$p(\vec{x}^i | \vec{f}^i, \vec{u}^i) = \prod_{t=2}^T \prod_{i=1}^M p(\vec{x}_t^i | \vec{f}_t^i, \alpha^i) \tag{10}$$

$$p(\vec{f}^i | \vec{u}^i) = \prod_{t=2}^T \prod_{i=1}^M \prod_{j=1}^M p(\vec{f}_t^i | \vec{f}_{1:t-1}^i, \vec{x}_{0:t-1}^i, \vec{u}^i) \tag{11}$$

where Equation (10) follows from the graphical model and the product-of-experts construction (Equation (1)). An empty slice ( $t < 2$  for a second-order dynamics model) implies no conditioning. The first two latent states at  $t = 0, 1$  are drawn from independent Gaussians,  $\prod_{i=1}^M p(\vec{x}_0^i) p(\vec{x}_1^i)$ . Equation (11) is one possible way of factorizing the augmented Gaussian process prior on the coupling function values: when  $\vec{f}^i(\cdot)$  is marginalized, the function values at time  $t$  depend on all past function values and latent states. We use this particular factorization for analytical convenience. Note that the dependence of the right hand side of Equation (11) on  $\vec{x}_{0:t-1}^i$  does not contradict the factorization order of Equation (9), because it depends only on latent variables from timesteps prior to  $t$ . Furthermore, we choose the following proposal variational posterior:

$$q(\vec{g}^i, \vec{x}^i, \vec{v}^i, \vec{f}^i, \vec{u}^i) = p(\vec{g}^i | \vec{x}^i, \vec{v}^i) q(\vec{v}^i) p(\vec{f}^i | \vec{u}^i) q(\vec{x}^i) q(\vec{u}^i) \tag{12}$$

with  $p(\vec{g}^i | \vec{x}^i, \vec{v}^i)$  given by Equation (7),  $p(\vec{f}^i | \vec{u}^i)$  by Equation (11) and  $q(\vec{x}^i)$  by Equation (3). The densities  $q(\vec{v}^i)$  and  $q(\vec{u}^i)$  are unconstrained except for normalization. With these distributions, we derive the standard free-energy ELBO [36], denoting  $\Theta = (\vec{x}^i, \vec{u}^i, \vec{f}^i, \vec{v}^i, \vec{g}^i)$ :

$$\log p(\vec{y}^i) \geq \mathcal{L}(\Theta) = \int d\Theta q(\Theta) \log \left( \frac{p(\vec{y}^i, \Theta)}{q(\Theta)} \right) \tag{13}$$

exploiting the assumption that the IPs  $\vec{r}^i$  and IVs  $\vec{v}^i$  are sufficient statistics for the function values  $\vec{g}^i$ . As we explain in detail in Appendix C, after canceling common factors in the variational posterior (Equation (12)) and the joint model density (Equation (5), cf. [16]), we find that the ELBO can be decomposed into one summand per part that describes the quality of the kinematics mapping (latent-to-observed)  $\mathcal{L}_{kin}^i$ , and one summand for the dynamics  $\mathcal{L}_{dyn}$ :

$$\mathcal{L}(\Theta) = \sum_{i=1}^M \mathcal{L}_{kin}^i + \mathcal{L}_{dyn} \tag{14}$$

where

$$\mathcal{L}_{kin}^i = \sum_{d=1}^D \int d\vec{x}^i d(\vec{v}^i)_d d(\vec{g}^i)_d p((\vec{g}^i)_d | \vec{x}^i, (\vec{v}^i)_d) q(\vec{x}^i) q((\vec{v}^i)_d) \log \frac{p((\vec{y}^i)_d | (\vec{g}^i)_d)}{q((\vec{v}^i)_d)}. \tag{15}$$

is—up to the Shannon entropy of approximating posterior of the latent dynamics variables  $H(q(\vec{x}^i)) = - \int d\vec{x}^i q(\vec{x}^i) \log(q(\vec{x}^i))$ —equal to the Bayesian GPLVM ELBO of [16]. The remaining integral

$$\begin{aligned} \mathcal{L}_{dyn} = & \int d\vec{u}^i q(\vec{u}^i) \left[ \sum_{t=2}^T \int d\vec{x}_{1:t} q(\vec{x}_{1:t}) \left( \int d\vec{f}_t^i p(\vec{f}_t^i | \vec{f}_{2:t-1}^i, \vec{x}_{0:t-1}^i, \vec{u}^i) \log p(\vec{x}_t^i | \vec{f}_t^i, \alpha^i) \right) \right] \\ & + \int d\vec{u}^i q(\vec{u}^i) \log \frac{p(\vec{u}^i)}{q(\vec{u}^i)} + \int d\vec{x}_{0:1} q(\vec{x}_{0:1}) \log p(\vec{x}_{0:1}) + H(q(\vec{x}^i)) \end{aligned} \tag{16}$$

is derived in detail in Appendix C. Briefly, we use the assumption that the IPs and IVs  $\vec{z}^i$  and  $\vec{u}^i$  are sufficient statistics for the function values  $\vec{f}_t^i$ . Optimizing with respect to  $q(\vec{u}^i)$  can be carried out in closed form using variational calculus and yields

$$\mathcal{L}_{dyn}(\Theta) \geq \log \int p(\vec{u}^i) \exp(\mathcal{C}(\vec{u}^i)) d\vec{u}^i + H(q(\vec{x}^i)) \quad (17)$$

where  $\mathcal{C}(\vec{u}^i)$  is given by Equation (A31). The inequality is due to the sufficient statistics assumption, which introduces another approximation step that lower-bounds  $\mathcal{L}_{dyn}$ . We now have all the ingredients to compute the ELBO for the whole model, and learn it.

## 5. Results

We used the machine-learning framework Theano [37] for automatic differentiation in Python 2.7 (Python Software Foundation. Python Language Reference, version 2.7. Available at <http://www.python.org>) to implement the model, and learned via optimization of the ELBO with the `scipy.optimize.fmin_l_bfgs_b` routine [38]. Latent space trajectories were initialized with PCA. We obtained the best ELBOs by first optimizing all parameters jointly, followed by a blocked optimization procedure. We optimize three groups of parameters: latent points and variances, kernel parameters and couplings, and IPs. The number of iterations of the blocked procedure depended on the application; we provide details in the sections below.

The advantage of the sparse approximations in the vCGPDM is that memory consumption of the model is greatly reduced. However, this approximation might also introduce errors, along with the fully factorized latent posterior (Equation (3)). We tried to quantify these errors in a cross-validators model comparison, and in a human perception experiment.

### 5.1. Synthetic Data

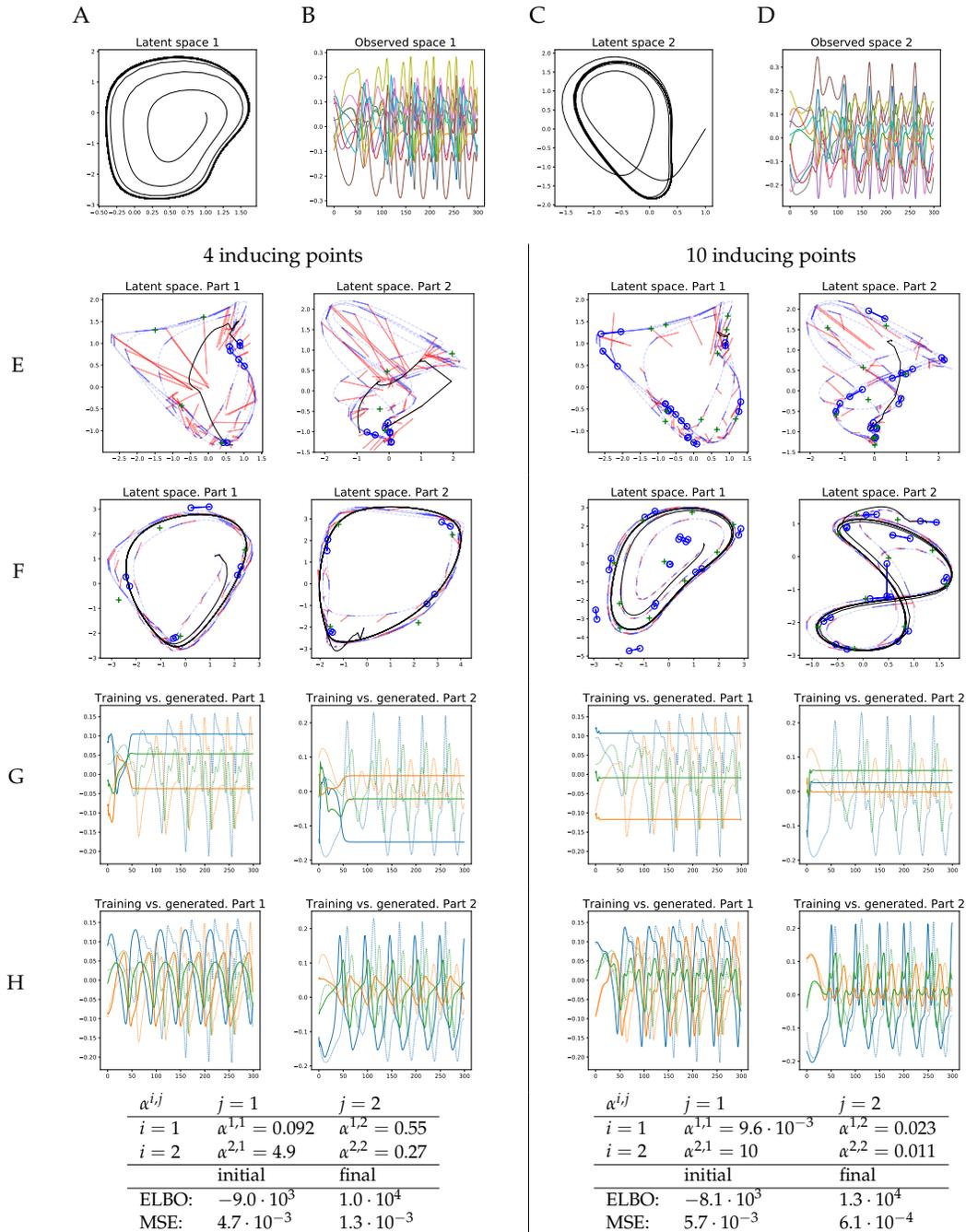
We demonstrate the learning of coupled dynamical systems on a synthetic dataset. First, we draw two dynamics transition functions  $g^1, g^2 \in \mathbb{R}^2 \rightarrow \mathbb{R}^2$  from a  $\mathcal{GP}$  with an RBF kernel, and then we generate latent trajectories according to:

$$\vec{x}_t^1 = g^1(\vec{x}_{t-1}^1) \quad (18)$$

$$\vec{x}_t^2 = 0.1g^1(\vec{x}_{t-1}^1) + 0.9g^2(\vec{x}_{t-1}^2) \quad (19)$$

at  $T = 300$  timepoints. Thus, we get two first-order, coupled latent dynamical systems, each of dimensionality 2. The trajectory in Latent Space 1 is independent of Latent Space 2, whereas Latent Space 2 is weakly coupled to Latent Space 1. Then, for each of the two parts, we draw 10 observed trajectories from another two RBF GPs with inputs on the latent trajectories. The latent trajectories are shown in Figure 3A,C. Figure 3B,D displays the corresponding observed trajectories. We learned a second-order vCGPDM from these data, iterating the blocked optimization until convergence of the ELBO to machine precision. We chose a second order system for this learning example, because the human movement models in the following are second-order vCGPDMs, too.

The results are shown in Figure 3E–H. Plots on the left half were generated with four IPs, plots in the right half with ten IPs. Figure 3E displays the initial positions of the dynamics IPs (blue circles) at the beginning of learning, connected circles form one second-order IP. Green crosses are the kinematics IPs (latent-to-observed mapping). Initial latent points (dashed blue lines) were obtained from the first two PCA components of the training data. Blue and red line segments are examples of the dynamics mapping: the end-points of the blue segments are the inputs, the distal endpoint of the red segment is the output. As one might expect, the initial conditions do not describe a dynamics which can produce trajectories resembling those in the training data: the black line is the mean latent trajectory, and Figure 3G shows the corresponding observable trajectories.



**Figure 3.** Synthetic training data. (A–D) Two-dimensional latent dynamics trajectories and corresponding observed 10-dimensional time series. Part 2 is weakly influenced by Part 1, Part 1 is not influenced by Part 2 (see Equation (18)). (E) Initial positions of second-order dynamics IPs (connected blue circles) and latent-to-observed IPs (green crosses). Line segments are examples of dynamics mapping inputs (endpoints of blue segments) and values (distal endpoints of red line segments). Black line: mean trajectory, generated by iterating the mean dynamics mapping from the same starting point as in (F) Latent space after learning. (G) Generated observable time series before learning (solid) and training data (dashed). (H) Generated time series (solid) after learning. Only three of the ten observable trajectories are presented for clarity. (Bottom) Learned couplings (see Equation (1)); ELBOs and MSEs rounded to two significant digits. Couplings  $\alpha^{i,j}$  reflect the dependency structure between parts: Part 1 is not driven by Part 2, but influences Part 2.

After learning, the latent trajectories appear to have a limit cycle (black line, Figure 3F), which is required to reproduce the training data. Furthermore, note that the inducing points align with that cycle, and the example mappings (blue and red line segments) indicate clearly how the latent trajectory is formed by iterating the GP mapping. Cross coupling GP mappings omitted for clarity. Note that the vCGPDM with ten IPs can model more complex latent dynamics manifolds than the four-IP vCGPDM. The observable trajectories (Figure 3H) look very similar to the training data up to a small phase shift, particularly for the 10 IP model. This observation is confirmed by the reduced mean squared trajectory error (MSE) between generated and training data after learning, which was evaluated after dynamic time warping [39] of the generated trajectories onto the training data. The MSEs are listed in the table at the bottom of Figure 3, where “final” indicates the values after learning, while “initial” indicates the values at the onset of learning after the latent space trajectories had been initialized to the first two PCA components of the training data. That learning was successful is also indicated by the increased final ELBO, which is higher for the 10 IP model.

We also provided the learned coupling  $\alpha$ s in this table. Recall that a low (high)  $\alpha$  means a large (small) influence of the corresponding part on the dynamics. The dependency structure between the latent spaces was correctly identified during learning:  $\alpha^{2,1} \gg \alpha^{1,1}$ , i.e., Part 2 has almost no influence on Part 1. In contrast,  $\alpha^{1,2} \approx 2\alpha^{2,2}$ , which indicates that Part 1 weakly controls Part 2.

### 5.2. Human Movement Data

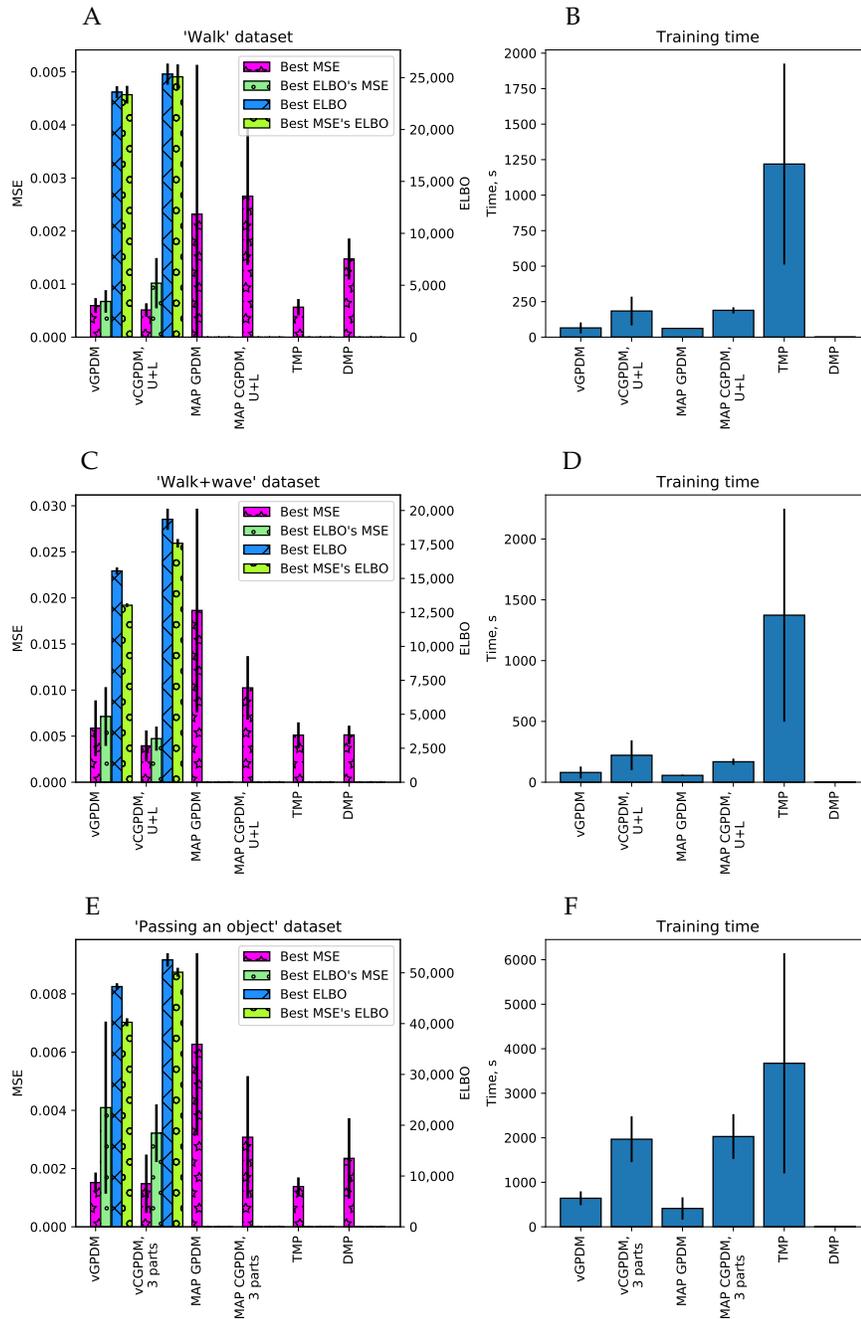
Model comparisons and psychophysical tests were carried out on human movement data. We employed a 10-camera PhaseSpace Impulse motion capture system, mapped the resulting position data onto a skeleton with 19 joints and computed joint angles in exponential-map representation, yielding a total of 60 degrees of freedom. Five walking-only and four walking + waving sequences each were used to train the models, as well as ten movements where the human participants were seated and passed a bottle from one hand to the other. Dynamical models were initialized with starting conditions taken from the training data. The blocked optimization was run for at most four iterations, which was enough to ensure convergence. It was terminated earlier if ELBO values did not change within machine precision between two subsequent iterations. Furthermore, we recorded another nine walking sequences for catch trials during the perception experiment, to rule out memorization effects. Generated and recorded sequences were rendered on a neutral avatar. Examples of stimuli, for different numbers of IPs, can be found in the movie `example_stimuli.mov` in the Supplementary Materials.

### 5.3. Variational Approximations are Better than MAP

We performed cross-validators model comparison on the following datasets: walking, walking + waving and passing-a-bottle. Examples of these data are shown in the movies in the Supplementary Materials: `S1_example_stimuli.mov` and `S4_pass_the_bottle.mkv`. We performed four-, five- and ten-fold crossvalidation, the number of folds was dictated by the dataset size. We were trying to determine how the sparsely parameterized vCGPDM performs in comparison to the full CGPDM, and several other MP models from the literature. Held-out data were always one complete trial. Models were trained on the remaining data and the generated trajectory was compared to the held-out one. Cross-validation score was the mean-squared error (MSE) of the kinematics after dynamic time warping [39] of trajectories generated by initializing the model to the first two frames of a held-out trial onto the complete held-out trial. We used dynamic time warping to compensate a slight phase difference in generated motions, which would otherwise lead to a much larger and uninformative MSE. We compared the following models:

- a GPDM with maximum-a-posteriori (MAP) estimation of the latent variables [25], called MAP GPDM in Figure 4;
- a fully marginalized two-part (upper/lower body) CGPDM with MAP estimation of the latent variables [14], called MAP CGPDM U+L;

- a three-part CGPDM model (left hand, right hand, and body) for the non-periodic “passing a bottle” dataset;
- their variational counterparts, vCGPDM 3-part, vCGPDM U+L and vGPDM;
- temporal MPs (TMP, instantaneous linear mixtures of functions of time) [9]; and
- DMPs [12].



**Figure 4.** Model comparison results. We plotted the average squared kinematics error on held-out data after dynamic time warping (MSE) and the variational lower bound on the model evidence (ELBO, Equation (A22)), where available, accompanied with corresponding model training time. Error bars are standard errors of the mean. (A,B) Walking dataset; (C,D) walking + waving dataset; and (E,F) “passing a bottle” dataset. Low MSE and high ELBO are better. For details, see text. Figure partially adapted from [19].

All latent spaces were three-dimensional. We tried 4–30 latent-to-observed IPs and 2–30 dynamics IPs. The MSE optima were near 10–15 IPs for both the walking and the walking + waving datasets, and near eight IPs for the “passing a bottle”. MAP GPDM and MAP CGPDM learning do not use any approximations or inducing points; they are the full GPs with covariance matrices  $\mathbf{K} \in \mathbb{R}^{T \times T}$ .

For the TMPs, we used up to 10 primitives; the MSE optimum was located at approximately six. For the DMPs, we used between 1 and 50 basis functions, and the lowest MSE was found around 15.

The results are plotted in Figure 4. Generally, the walking + waving movement is more difficult to reproduce for all models than walking only: the MSE of the latter is lower than that of the former, and the ELBO is higher. This indicates that the latter is a more complex movement, see also the movie `modular_primitives.avi` in the online Supplementary Materials. The two-part vCGPDM reaches the lowest MSE compared to all other models. Clearly, it is better than the full-capacity (no IPs) MAP models, which means that the extra effort of developing of a variational approximation which explicitly represents an approximation to the latent states’ posterior and needs to store only  $\approx 10$  IPs rather than  $\approx 10^4$  data points was well spent. In addition, the best ELBO’s MSE (that is, the MSE at the maximum of the ELBO) is a fairly good predictor of the best MSE, which justifies our simple variational approximation for model selection.

The vCGPDM U+L outperforms the vGPDM particularly on the “walking + waving” dataset. This shows the usefulness of having modular, coupled dynamics models when the (inter)acting (body)parts execute partially independent movements.

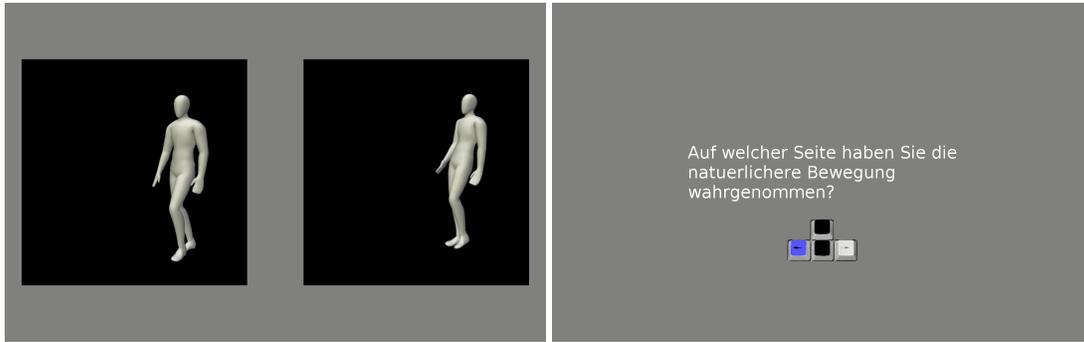
The vCGPDM with three parts for “passing a bottle” does not show a clear advantage over the monolithic model in the cross-validation test, and is on par with the TMP model. However, dynamics factorization did not affect the performance either. This may be indicative for the strong coupling necessary to successfully pass an object from one hand to the other. Such a strong coupling is parsimoniously expressed by having a single latent dynamical system drive all observable degrees of freedom.

The timing results show that the training times of the vCGPDM are usually less than 15 min. Error bars are standard deviations, estimated across all numbers of IPs and cross-validation splits. The rather large training time for the TMP model is due to the implementation from [9] which optimizes a rather large covariance matrix between all MPs.

#### 5.4. A Small Number of IPs Yields Perceptually Convincing Movements

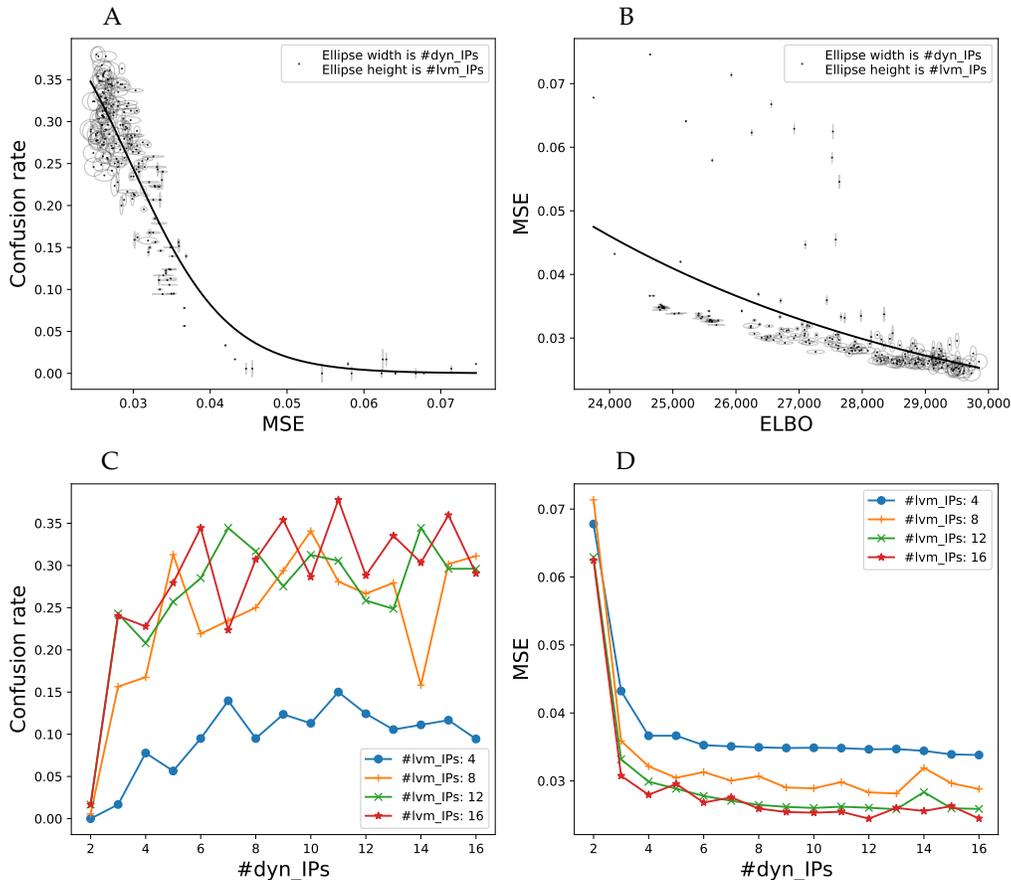
We conducted a psychophysical experiment to quantify the perceptual validity of the generated movements. More specifically, we investigated the model complexity required for perceptually convincing movements.

**Experiment:** Thirty-one human observers (10 male, mean age:  $23.8 \pm 3.5$ a) participated in a two-alternative forced-choice task to distinguish between natural and generated movements (see Figure 5 for an example of an experimental trial). Natural movements consisted of 15 walking movements. The artificial movements were generated by a two-part (upper/lower body) vCGPDM. We used 2–16 dynamics IPs and 4–16 latent-to-observed IPs. We chose these numbers based on pilot tests to span the range from clearly unnatural to very natural looking movements. To test whether participants simply memorized the 15 natural stimuli during the experiment, we added 10 catch trials in the last quarter of the experiment where previously unused natural movements were tested against the known natural stimuli. The trial sequence was randomized for every subject. All experimental procedures were approved by the local ethics commission.



**Figure 5.** Psychophysical Experiment. In each trial, a natural and a generated movement were simultaneously presented to participants (left). After presentation, they used the arrow keys to choose the movement perceived as more natural (right). There was no time limit on the response, but typically participants responded quickly (less than 1 s). After the response, people were asked to fixate a cross in the middle of the screen, which appeared for a second. The length of the stimuli was 1.8 s, with a total of 1170 presentations. A video of the experiment called `S2_experiment_demo.avi` is provided in the Supplementary Materials.

**Results:** We computed the confusion rate, i.e., the frequency of choosing the model-generated movement as more natural across all participants as a function of the number of IPs for the dynamics and latent-to-observed mappings. Optimally, we might expect this rate to approach  $\frac{1}{2}$  when the generated movements are indistinguishable from the natural ones. We investigated if the confusion rate approached this limit, how it depends on the mean-squared residual error on the training data, and how this error is connected to the ELBO. The results are plotted in Figure 6. We also fitted the confusion rate data with a logistic sigmoid  $\frac{0.5}{1+\exp(a \cdot \text{MSE}+c)}$  (solid line in Figure 6A), and the MSE with an exponential function (solid line in Figure 6, right). Each data point represents one combination of dynamics/latent-to-observed IP numbers, indicated by width and height of the ellipses. Clearly, confusion rate increases fairly monotonically with decreasing MSE, as indicated by the good logistic sigmoid fit. Furthermore, models with more IPs also tend to yield higher confusion rates. A sufficient number ( $>10$ ) dynamics IPs is more important than a large number of latent-to-observed IPs, which can be seen by the very narrow ellipses in the region with high MSE, and many wider ellipses in the lower MSE part of the figure. A similar observation can be made about the relationship between ELBO and MSE (Figure 6B). It indicates that ELBO is already a good predictor for the model performance. For a very small number of dynamics IPs, increasing the number of latent-to-observed IPs does not decrease the MSE as much as increasing the dynamics IPs does. Moreover, note that the relationship between MSE and ELBO becomes fairly monotonic when  $\text{ELBO} > 28,500$ , which is where human perceptual performance can be predicted from ELBO. While the confusion rate has not quite reached its theoretical maximum in our experiment, these results are evidence that human perceptual expectations can be nearly met with very compactly parameterized MP models (Figure 6C,D). Moreover, good dynamics models seem to outweigh precise kinematics. We found no evidence for stimulus memorization from the confusion rates of the catch trials.

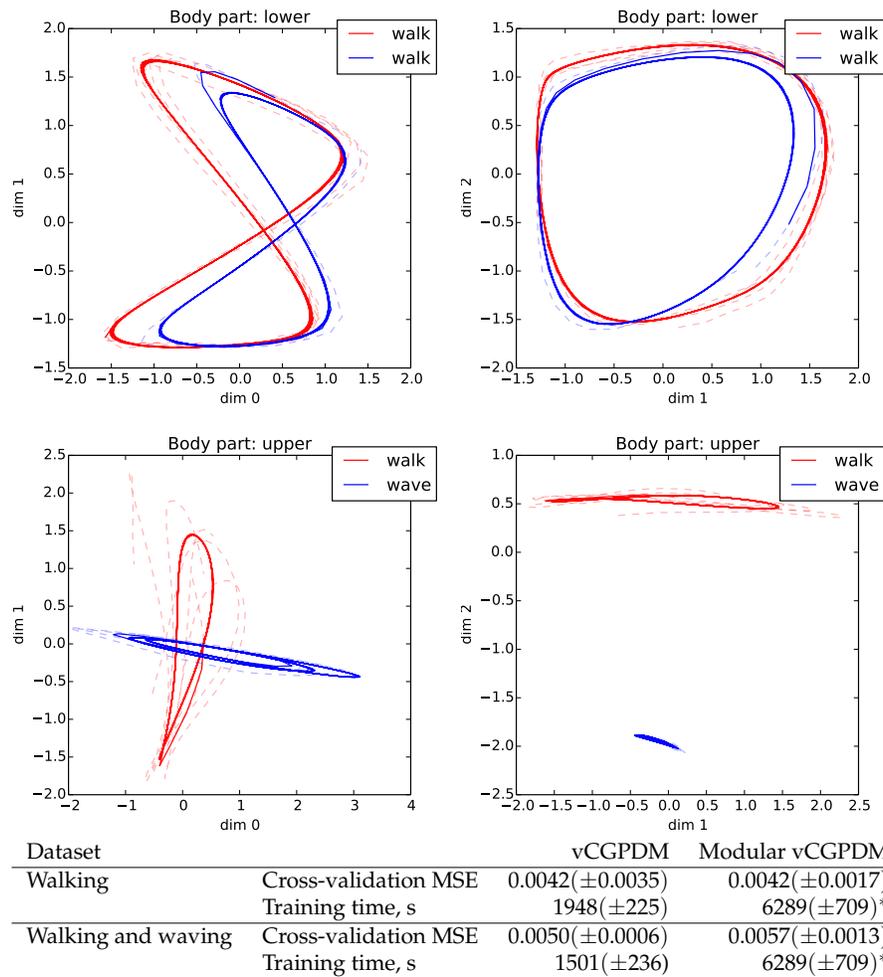


**Figure 6.** (A) Confusion rate between natural and vCGPDM-generated stimuli as a function of mean-squared residual error (MSE) on the training data, averaged across all participants. Each data point represents one combination between number of IPs/IVs for the latent-to-observed mapping (indicated by ellipse height) and number of IPs/IVs for the dynamics mappings (ellipse widths). A confusion rate of 0.5 indicates that human observers are not able to distinguish replays of real movements from model-generated counterparts. The vCGPDM is approaching this limit from below for a fairly small number of IPs/IVs. Solid line: fit with logistic sigmoid function. (B) Relationship between training MSE and ELBO. Solid line: fit with exponential function. Additional dynamics IPs contribute more to the reduction of the MSE than latent-to-observed IPs. MSE and therefore confusion rate can be predicted well from ELBO if ELBO > 28,500. (C,D) Influence of number of dynamics IPs on the confusion rate and MSE, respectively, for a selected number of latent-to-observed IPs. The confusion rate has a broad maximum around 8–12 dynamics IPs, whereas the MSE has a shallow minimum at that location.

### 5.5. Modularity Test

Next, we examined if the intended modularization of our model can be used to compose novel movements from previously learned parts. We trained a vCGPDM consisting of one part for the lower body (below and including pelvis), and a second part for the upper body. Twenty-five IPs for the latent-to-observed mapping of each part were shared across all movements. The walking MP, parameterized by 16 IPs for the lower-body dynamics and the lower-to-upper mappings, was also shared. We used a different set of 16 IPs for the upper body MPs between arm-swing and waving. Furthermore, the coupling  $\alpha^{j,i}$  were learned anew for each combination of upper/lower MPs. The resulting latent space trajectories are plotted in Figure 7. All generated trajectories (solid lines) are on average close to the training data (dashed lines). While the walking trajectories

for the lower body are very similar for the two movements, the upper body trajectories clearly differ. Movements generated from this model are very natural (see video S3\_modular\_primitives.mov in the Supplementary Materials). This is a first demonstration that the vCGPDM with non-marginalized couplings can be used to learn a library of compactly parameterized MPs, from which novel movements can be produced with little additional memory requirements (i.e., new coupling  $\alpha^{j,i}$  only).



**Figure 7.** (Top) Modularity example. Shown are 2D projections of generated 3D latent space trajectories (solid) and training data (dashed). Blue: walk + wave movements; red: walk + normal arm swing. Dynamics IPs re-used across movements for lower body. (Bottom) Cross-validation MSEs of non-modular and modular vCGPDM. Modular vCGPDM was trained on the combined dataset; training time (\*) is shared between both movement datasets.

For a quantitative evaluation, we looked at the leave-one-out cross-validation MSEs of the vCGPDM trained on datasets separately and modular vCGPDM trained on both datasets (see Figure 7, bottom). Within the standard errors, MSEs are equal, indicating that modular re-use of previously trained components does not necessarily sacrifice accuracy, while reducing storage requirements. Training time for the modular vCGPDM is larger due to the learning of the combined dataset and optimizing the couplings afterwards. This time would be amortized if more compositional movements were learned, where previously learned parts could be reused.

## 6. Conclusions

The vCGPDM, a full variational approximation to the CGPDM, allows for learning a deterministic approximation of latent space trajectories, and compactly parameterizing dynamics and kinematics mappings. First, we showed that the sparsely parameterized vCGPDM outperforms the full-capacity, monolithic CGPDM employing MAP to approximate the latent dynamics posterior. It also surpasses other current MP models; we speculate that this is accomplished by its learnable dynamics.

Second, we demonstrated that our compact representation of the latent space dynamics, and of the latent-to-observed mapping, enables the model to generate perceptually convincing full-body movements with a fairly small number of IPs. To our knowledge, a systematic investigation of the number of IPs needed for perceptual plausibility had not been done before, albeit more monolithic models were in the focus of earlier studies [27,34,40]. Moreover, we demonstrated that a high enough ELBO can be used to predict average human classification performance, which might allow for an automatic model selection process when training the model on large databases. Within the range of IPs which we tested, the ELBO was still increasing with their number. We chose that range because we wanted to see how few IPs would still lead to perceptually indistinguishable movements. Due to experimental time constraints, we did not investigate perceptual performance at the point where the ELBO begins to decrease with increasing IPs (i.e., the approximately optimal model), but we plan to do that in the future.

Third, we showed that the model can be employed in a modular fashion, using one lower-body dynamics model, and coupling it to two different models for the upper body. Note that the lower-to-upper coupling function was the same for the two upper-body models. Each of these models, including the coupling functions to the other model parts, may therefore be viewed as a modular MP that is parameterized compactly by a small number of IPs and values. This sparse parameterization allows us to infer modular MPs from a large collection of movements, and investigate their composition. To generate complex movement sequences, we will put a switching prior on top of the dynamical models, as in [29].

We are currently researching sensorimotor primitives, i.e., MPs that can be used to predict sensory input and be controlled by it via conditioning. This conditioning can take place on at least two timescales: a short one (while the MP is running), thus effectively turning the MPs into flexible control policies, such the probabilistic MPs described by Paraschos et al. [41], and a long timescale, i.e., the planning of the movement. This could be implemented by learning a mapping from goals and affordances onto the coupling weights, comparable to the DMPs with associative skill memories [42]. There is evidence that humans modulate the coupling between their MPs during the planning stage: whole-body posture changes have been observed in anticipation of reaching for a goal object in a known location, even if the object is currently invisible [43].

Lastly, we note that our CGPDM could be used as a flexible policy model for PILCO-style reinforcement learning (Probabilistic Inference for Learning Control, [44]). PILCO requires a dynamics model that can propagate uncertainties through time; the vCGPDM is able to do that. Thus, our model could be used as a lower dimensional dynamics model which can capture the dependencies between observable variables via latent space uncertainties.

**Supplementary Materials:** The following are available online <http://www.mdpi.com/1099-4300/20/10/724/s1>, Video S1: example\_stimuli, Video S2: experiment\_demo, Video S3: modular\_primitives, Video S4: pass\_the\_bottle.

**Author Contributions:** D.V. and D.E. conceived vCGPDM model. D.V. derived, implemented and tested models. B.K. and D.E. conceived psychophysical experiment. B.K. conducted experiment and evaluated data. All authors wrote paper.

**Funding:** This work was supported by the Deutsche Forschungsgemeinschaft under DFG-IRTG 1901 “The Brain in Action” and DFG-SFB-TRR 135 project C06.

**Acknowledgments:** We thank Olaf Haag for help with rendering the movies, and Björn Büdenbender for assistance with MoCap. Open access costs were paid by the University Library of Marburg and the DFG.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

MP	movement primitive
DMP	dynamical movement primitive
$\mathcal{GP}$	Gaussian process
GPDM	Gaussian process dynamical model
CGPDM	coupled Gaussian process dynamical model
vCGPDM	variational coupled Gaussian process dynamical model
IP	inducing point
IV	inducing value
MSE	mean squared error

## Appendix A. Exact Variational Optimization of Parts of the ELBO

While optimizing the full variational posterior in augmented Gaussian processes models with respect to the IVs, the following type of term appears several times in the ELBO equation:

$$\begin{aligned}\mathcal{R}(q(\vec{u}), r(\vec{v})) &= \int q(\vec{u}) \left( f(r(\vec{v}), \vec{u}) + \log \frac{p(\vec{u})}{q(\vec{u})} \right) d\vec{u} \\ &= \int q(\vec{u}) f(r(\vec{v}), \vec{u}) d\vec{u} + \int q(\vec{u}) \log p(\vec{u}) d\vec{u} - \int q(\vec{u}) \log q(\vec{u}) d\vec{u}\end{aligned}\quad (\text{A1})$$

To simplify the optimization of such terms, we would like to carry out the optimization with respect to the density  $q(\vec{u})$  analytically to remove the dependency on  $q(\vec{u})$ . Note that we allow only  $q(\vec{u}), r(\vec{v})$  to vary, while the functions  $f(r(\vec{v}), \vec{u})$  and  $p(\vec{u})$  are assumed to be fixed. To this end, we calculate for the optimal variational  $q^*(\vec{u})$  in the above equation. This approach was suggested in [23], however, it is not well described there. Here, we give an extended derivation. A necessary condition for maximality is a vanishing functional derivative under the constraint that the density  $q(\vec{u})$  is normalized to one:

$$\int q(\vec{u}) d\vec{u} - 1 = 0 \quad (\text{A2})$$

which is fulfilled at the stationary points of the Lagrangian

$$\mathcal{X}(q(\vec{u}), r(\vec{v})) = \mathcal{R}(q(\vec{u}), r(\vec{v})) + \lambda \left( \int q(\vec{u}) d\vec{u} - 1 \right) \quad (\text{A3})$$

where  $\lambda$  is chosen so that Equation (A2) holds. Taking the functional derivative of  $\mathcal{X}(q(\vec{u}), r(\vec{v}))$  and setting it to zero yields

$$\frac{\delta \mathcal{X}(q(\vec{u}), r(\vec{v}))}{\delta q(\vec{u})} = f(r(\vec{v}), \vec{u}) + \log p(\vec{u}) - \log q(\vec{u}) - 1 + \lambda = 0 \quad (\text{A4})$$

and therefore, denoting  $Z = \exp(-\lambda + 1)$

$$q^*(\vec{u}) = \exp(f(r(\vec{v}), \vec{u}) + \log p(\vec{u}) - 1 + \lambda) \quad (\text{A5})$$

$$q^*(\vec{u}) = \frac{1}{Z} p(\vec{u}) \exp(f(r(\vec{v}), \vec{u})) \quad (\text{A6})$$

$$Z = \exp(-\lambda + 1) = \int p(\vec{u}) \exp(f(r(\vec{v}), \vec{u})) d\vec{u} \quad (\text{A7})$$

Substituting the optimal  $q^*(\vec{u})$  into the original term, we get:

$$\begin{aligned}
 \mathcal{R}(r(\vec{v})) &= \int \frac{1}{Z} p(\vec{u}) \exp(f(r(\vec{v}), \vec{u})) \left( f(r(\vec{v}), \vec{u}) + \log \frac{p(\vec{u})}{\frac{1}{Z} p(\vec{u}) \exp(f(r(\vec{v}), \vec{u}))} \right) d\vec{u} \\
 &= \int \frac{1}{Z} p(\vec{u}) \exp(f(r(\vec{v}), \vec{u})) \left( \log \frac{p(\vec{u}) \exp(f(r(\vec{v}), \vec{u}))}{\frac{1}{Z} p(\vec{u}) \exp(f(r(\vec{v}), \vec{u}))} \right) d\vec{u} \\
 &= \log(Z) \frac{1}{Z} \int p(\vec{u}) \exp(f(r(\vec{v}), \vec{u})) d\vec{u} \\
 &= \log \int p(\vec{u}) \exp(f(r(\vec{v}), \vec{u})) d\vec{u}
 \end{aligned}
 \tag{A8}$$

This is the optimized version of Equation (A1), which depends only on  $r(\vec{v})$ .

**Appendix B. ARD RBF Kernel  $\Psi$  Statistics. Full Covariance Variational Parameters Case.**

During the computation of the ELBO, it is necessary to evaluate expected values of Gaussian process kernel functions under the variational posterior distributions. Here, we derive these expectations, referred to as  $\Psi$  statistics in the literature [16], for the type of kernel we used in this paper: an automatic relevance determination, squared exponential kernel (ARD RBF). The ARD RBF kernel is defined as:

$$k(\vec{x}, \vec{x}') = \sigma_f^2 \exp \left( -\frac{1}{2} \sum_{q=1}^Q \frac{((\vec{x})_q - (\vec{x}')_q)^2}{\lambda_q} \right)
 \tag{A9}$$

where  $\lambda_q$  are the ARD factors,  $\sigma_f^2$  is the variance of the kernel and  $Q$  is the dimensionality of  $\vec{x}$ . In matrix notation:

$$\lambda = \text{diag}(\lambda_1 \dots \lambda_Q)
 \tag{A10}$$

$$k(\vec{x}, \vec{x}') = \sigma_f^2 \exp \left( -\frac{1}{2} (\vec{x} - \vec{x}')^T \lambda^{-1} (\vec{x} - \vec{x}') \right)
 \tag{A11}$$

Let  $\vec{x}$  be a random variable drawn from multivariate Gaussian distributions with mean  $\vec{\mu}$  and covariance matrix  $S$ . Consider the following form of the approximate variational posterior distribution of  $\vec{x}$ :

$$q(\vec{x}) = \mathcal{N}(\vec{x}_n | \vec{\mu}, S)
 \tag{A12}$$

The  $\Psi_0$  statistic is the expectation of the kernel for two identical arguments, which is easy to calculate:

$$\begin{aligned}
 \Psi_0 &= \int k(\vec{x}_n, \vec{x}_n) \mathcal{N}(\vec{x}_n | \vec{\mu}_n, S_n) d\vec{x}_n \\
 &= \int \sigma_f^2 \mathcal{N}(\vec{x}_n | \vec{\mu}_n, S_n) d\vec{x}_n \\
 &= \sigma_f^2 \int \mathcal{N}(\vec{x}_n | \vec{\mu}_n, S_n) d\vec{x}_n \\
 &= \sigma_f^2
 \end{aligned}
 \tag{A13}$$

The  $\Psi_1$  statistic is the expectation with respect to one kernel argument, given that the other is constant:

$$\begin{aligned}
 \Psi_1 &= \int k(\vec{x}, \vec{z}) \mathcal{N}(\vec{x} | \vec{\mu}, S) d\vec{x} \\
 &= \int \sigma_f^2 \exp \left( -\frac{1}{2} (\vec{x} - \vec{z})^T \lambda^{-1} (\vec{x} - \vec{z}) \right) \mathcal{N}(\vec{x} | \vec{\mu}, S) d\vec{x}
 \end{aligned}
 \tag{A14}$$

To evaluate the integral, complete the ARD RBF kernel to a scaled Gaussian distribution with covariance matrix  $\lambda$  and mean  $\vec{z}$ :

$$\begin{aligned}
 \Psi_1 &= \sigma_f^2 \int \frac{Z(\lambda)}{Z(\lambda)} \exp \left( -\frac{1}{2} (\vec{x}_r - \vec{z})^T \lambda^{-1} (\vec{x}_r - \vec{z}) \right) \mathcal{N}(\vec{x}_n | \vec{\mu}_n, S_n) d\vec{x} \\
 &= \sigma_f^2 Z(\lambda) \int \mathcal{N}(\vec{x} | \vec{z}, \lambda) \mathcal{N}(\vec{x} | \vec{\mu}, S) d\vec{x}
 \end{aligned}
 \tag{A15}$$

with  $Z(\lambda) = (2\pi)^{Q/2} \sqrt{|\lambda|}$ . The integral over the product of two Gaussians can be carried out to yield (see [45], Identity 371):

$$\begin{aligned}
 \Psi_1 &= \sigma_f^2 Z(\lambda) \mathcal{N}(\bar{z} | \bar{\mu}, \lambda + S) \\
 &= \sigma_f^2 (2\pi)^{Q/2} \sqrt{|\lambda|} \frac{1}{(2\pi)^{Q/2} \sqrt{|\lambda + S|}} \exp\left(-\frac{1}{2}(\bar{z} - \bar{\mu})^T (\lambda + S)^{-1} (\bar{z} - \bar{\mu})\right) \\
 &= \sigma_f^2 \frac{\sqrt{|\lambda|}}{\sqrt{|\lambda + S|}} \exp\left(-\frac{1}{2}(\bar{z} - \bar{\mu})^T (\lambda + S)^{-1} (\bar{z} - \bar{\mu})\right) \\
 &= \sigma_f^2 \sqrt{\frac{\prod_{q=1}^Q \lambda_q}{|\lambda + S|}} \exp\left(-\frac{1}{2}(\bar{z} - \bar{\mu})^T (\lambda + S)^{-1} (\bar{z} - \bar{\mu})\right)
 \end{aligned} \tag{A16}$$

The  $\Psi_2$  statistic integral, which correlates two kernel function values at different points  $\bar{z}$  and  $\bar{z}'$ , can be solved in a similar manner: first, by collecting terms in the exponents and completing quadratic forms, and second, by the application of Identity 371 from [45]:

$$\begin{aligned}
 \Psi_2 &= \int k(\bar{x}, \bar{z}) k(\bar{z}', \bar{x}) \mathcal{N}(\bar{x} | \bar{\mu}, S) d\bar{x} \\
 &= (\sigma_f^2 Z(\lambda))^2 \int \mathcal{N}(\bar{x} | \bar{z}, \lambda) \mathcal{N}(\bar{x} | \bar{z}', \lambda) \mathcal{N}(\bar{x} | \bar{\mu}, S) d\bar{x} \\
 &= (\sigma_f^2 Z(\lambda))^2 \int \mathcal{N}(\bar{z}' | \bar{z}, 2\lambda) \mathcal{N}(\bar{x} | \frac{1}{2}(\bar{z}' + \bar{z}), \frac{1}{2}\lambda) \mathcal{N}(\bar{x} | \bar{\mu}, S) d\bar{x} \\
 &= (\sigma_f^2 Z(\lambda))^2 \mathcal{N}(\bar{z}' | \bar{z}, 2\lambda) \int \mathcal{N}(\bar{x} | \frac{1}{2}(\bar{z} + \bar{z}'), \frac{\lambda}{2}) \mathcal{N}(\bar{x} | \bar{\mu}, S) d\bar{x} \\
 &= (\sigma_f^2 Z(\lambda))^2 \mathcal{N}(\bar{z}' | \bar{z}, 2\lambda) \mathcal{N}(\bar{\mu} | \frac{1}{2}(\bar{z} + \bar{z}'), \frac{\lambda}{2} + S) \\
 &= \sigma_f^4 (2\pi)^Q \left(\prod_{q=1}^Q \lambda_q\right) \mathcal{N}(\bar{z}' | \bar{z}, 2\lambda) \mathcal{N}(\bar{\mu} | \frac{\bar{z} + \bar{z}'}{2}, \frac{\lambda}{2} + S)
 \end{aligned} \tag{A17}$$

For the case of a diagonal covariance matrix  $S$  the  $\Psi_2$  statistic can be simplified further [16].

### Appendix C. vCGPDM Dynamics ELBO Derivation

We now present a detailed derivation of the ELBO with a focus on the dynamics component. Assume we deal with  $M$  parts. We have  $M \times M$  latent dynamics mappings, which are combined into  $M$  mappings with product of experts—multiplying and renormalizing the distributions from all parts' predictions about each part. Each of the  $M \times M$  mappings  $f^{j,i}()$  is augmented with IPs  $\bar{z}^{j,i}$  and IVs  $\bar{u}^{j,i}$ , which are drawn out of the same  $\mathcal{GP}$  priors as the mappings. For clarity, we omit spelling out the dependence of the IVs on the IPs in the following, and we ask the reader to remember that any distribution over IVs is implicitly conditioned onto the corresponding IPs. The full augmented joint distribution of the model, which is derived in Section 4 (Equation (5)), is:

$$\begin{aligned}
 p(\bar{y}^i, \bar{g}^i, \bar{x}^i, \bar{v}^i, \bar{f}^{j,i}, \bar{u}^{j,i}) &= p(\bar{y}^i | \bar{g}^i) p(\bar{g}^i | \bar{x}^i, \bar{v}^i) p(\bar{v}^i) p(\bar{x}^i, \bar{f}^{j,i} | \bar{u}^{j,i}) p(\bar{u}^{j,i}) \\
 &= \left[ \prod_{i=1}^M p(\bar{y}^i | \bar{g}^i) p(\bar{g}^i | \bar{x}^i, \bar{v}^i) p(\bar{v}^i) \right] p(\bar{x}^i, \bar{f}^{j,i} | \bar{u}^{j,i}) p(\bar{u}^{j,i}) \\
 &= \left[ \prod_{i=1}^M \left[ \prod_{d=1}^{D_i} p((\bar{y}^i)_d | (\bar{g}^i)_d) p((\bar{g}^i)_d | \bar{x}^i, (\bar{v}^i)_d) p((\bar{v}^i)_d) \right] \right] \\
 &\times \left[ \prod_{t=1}^T \left[ \prod_{i=1}^M p(\bar{x}_t^i | \bar{f}_t^{j,i}, \alpha^{j,i}) \right] \left[ \prod_{i=1}^M \prod_{j=1}^M p(\bar{f}_t^{j,i} | \bar{f}_{1:t-1}^{j,i}, \bar{x}_{0:t-1}^{j,i}, \bar{u}^{j,i}) \right] \right] \\
 &\times \left[ \prod_{i=1}^M \prod_{j=1}^M p(\bar{u}^{j,i}) \right] \left[ \prod_{i=1}^M p(\bar{x}_0^i) \right]
 \end{aligned} \tag{A18}$$

The full proposal variational posterior is (Equation (12) in Section 4):

$$\begin{aligned}
 q(\vec{g}; \vec{x}, \vec{v}, \vec{f}; \vec{u}) &= p(\vec{g}; \vec{x}, \vec{v}) q(\vec{v}; \vec{x}) p(\vec{f}; \vec{x}, \vec{u}) q(\vec{x}; \vec{u}) \\
 &= p(\vec{g}; \vec{x}, \vec{v}) q(\vec{v}; \vec{x}) \left[ \prod_{t=1}^T \prod_{i=1}^M \prod_{j=1}^M p(\vec{f}_t^{j,i} | \vec{f}_{1:t-1}^{j,i}, \vec{x}_{0:t-1}^j, \vec{u}^{j,i}) \right] q(\vec{x}; \vec{u}) q(\vec{u}; \vec{v}) \tag{A19}
 \end{aligned}$$

Thus, the ELBO is given by:

$$\mathcal{L}(\Theta) = \int d\vec{g} d\vec{x} d\vec{v} d\vec{f} d\vec{u} q(\vec{g}; \vec{x}, \vec{v}, \vec{f}; \vec{u}) \log \left( \frac{p(\vec{y}; \vec{g}, \vec{x}, \vec{v}, \vec{f}; \vec{u})}{q(\vec{g}; \vec{x}, \vec{v}, \vec{f}; \vec{u})} \right) = \sum_{i=1}^M \mathcal{L}_{kin}^i + \mathcal{L}_{dyn} \tag{A20}$$

$$= \sum_{i=1}^M \sum_{d=1}^D \int d\vec{x}^i d\vec{v}^i d(\vec{g}^i)_d p((\vec{g}^i)_d | \vec{x}^i, \vec{v}^i) q(\vec{x}^i) q(\vec{v}^i) \log \frac{p((\vec{y}^i)_d | (\vec{g}^i)_d)}{q(\vec{v}^i)} \tag{A21}$$

$$\begin{aligned}
 &+ \int d\vec{u} q(\vec{u}; \vec{v}) \left[ \sum_{t=1}^T \int q(\vec{x}_t) q(\vec{x}_{-t}) \left( \int d\vec{f} \left[ \prod_{i=1}^M \prod_{j=1}^M p(\vec{f}_t^{j,i} | \vec{f}_{1:t-1}^{j,i}, \vec{x}_{0:t-1}^j, \vec{u}^{j,i}) \right] \log \prod_{i=1}^M p(\vec{x}_t^i | \vec{f}_t^i, \alpha^i) \right) \right] \\
 &+ \int d\vec{u} q(\vec{u}; \vec{v}) \log \frac{p(\vec{u}; \vec{v})}{q(\vec{u}; \vec{v})} + \sum_{t=0}^1 \int q(\vec{x}_t) \log p(\vec{x}_t) d\vec{x}_t + H(q(\vec{x})) \tag{A22}
 \end{aligned}$$

The term in Equation (A21), which we call  $\sum_{i=1}^M \mathcal{L}_{kin}^i$ , is the GPLVM ELBO up to  $H(q(\vec{x}))$  and is given in [16]. Next, we consider only the ELBO component which is relevant for the dynamics  $\mathcal{L}_{dyn}$  (last two lines of the right hand side of Equation (A22)) and apply the sufficient statistics assumption: knowing  $\vec{x}_{-t}^j$  and  $\vec{u}^{j,i}$  is sufficient for the  $\vec{f}_t^{j,i}$  distribution, i.e.,  $p(\vec{f}_t^{j,i} | \vec{f}_{1:t-1}^{j,i}, \vec{x}_{0:t-1}^j, \vec{u}^{j,i}) = p(\vec{f}_t^{j,i} | \vec{x}_{0:t-1}^j, \vec{u}^{j,i})$ . This assumption lower-bounds  $\mathcal{L}_{dyn}$ , because it constrains the variational posterior in Equation (A19) away from the correct solution. The sum over the initial latent points in the last line of Equation (A22) may be longer or shorter depending on the dynamics model order, here we use a second order model. The innermost integral can then be written as:

$$\begin{aligned}
 \mathcal{A} &= \int \left[ \prod_{i=1}^M \prod_{j=1}^M p(\vec{f}_t^{j,i} | \vec{x}_{-t}^j, \vec{u}^{j,i}) \right] \log \prod_{i=1}^M p(\vec{x}_t^i | \vec{f}_t^i, \alpha^i) d\vec{f}_t^i \\
 &= \sum_{i=1}^M \int \left[ \prod_{j=1}^M p(\vec{f}_t^{j,i} | \vec{x}_{-t}^j, \vec{u}^{j,i}) \right] \log p(\vec{x}_t^i | \vec{f}_t^i, \alpha^i) d\vec{f}_t^i \\
 &= \sum_{i=1}^M \int \left[ \prod_{j=1}^M \mathcal{N}(\vec{f}_t^{j,i} | \vec{\mu}_{\vec{f}_t^{j,i}}, \mathbf{S}_{\vec{f}_t^{j,i}}) \right] \log \mathcal{N}(\vec{x}_t^i | \alpha^i \sum_{j=1}^M (\alpha^{j,i})^{-1} \vec{f}_t^{j,i}, \mathbf{I} \alpha^i) d\vec{f}_t^i \\
 &= \sum_{i=1}^M \left[ -\frac{1}{2} \text{tr} \left[ \alpha^i \sum_{j=1}^M (\alpha^{j,i})^{-2} \mathbf{S}_{\vec{f}_t^{j,i}} \right] + \log \mathcal{N} \left( \vec{x}_t^i | \alpha^i \sum_{j=1}^M (\alpha^{j,i})^{-1} \vec{\mu}_{\vec{f}_t^{j,i}}, \mathbf{I} \alpha^i \right) \right] \tag{A23}
 \end{aligned}$$

$$\vec{\mu}_{\vec{f}_t^{j,i}} = \mathbf{K}_{\vec{x}_{-t}^j, \vec{z}_{-t}^{j,i}}^{j,i} \left( \mathbf{K}_{\vec{z}_{-t}^{j,i}, \vec{z}_{-t}^{j,i}}^{j,i} \right)^{-1} \vec{u}^{j,i} \tag{A24}$$

$$\mathbf{S}_{\vec{f}_t^{j,i}} = \mathbf{K}_{\vec{x}_{-t}^j, \vec{x}_{-t}^j}^{j,i} - \mathbf{K}_{\vec{x}_{-t}^j, \vec{z}_{-t}^{j,i}}^{j,i} \left( \mathbf{K}_{\vec{z}_{-t}^{j,i}, \vec{z}_{-t}^{j,i}}^{j,i} \right)^{-1} \mathbf{K}_{\vec{z}_{-t}^{j,i}, \vec{x}_{-t}^j}^{j,i} \tag{A25}$$

$$\alpha^i = \left( \sum_{j=1}^M (\alpha^{j,i})^{-1} \right)^{-1} \tag{A26}$$

where  $\mathbf{K}$  are kernel matrices, obtained by evaluating the kernel function at the pairs of points indicated by the subscripts. Equations (A24) and (A25) follow from the standard formulas for conditional Gaussians (see, e.g., [45], Identities 352–254). Equation (A26) follows from the product-of-experts construction (see Section 3, Equation (1)).

The next step in the evaluation of Equation (A22) is integrating  $\mathcal{A}$  over  $d\vec{x}_t^i$  and  $\vec{x}_{-t}^i$ , which requires averaging kernel matrices. The results of this averaging are denoted by  $\Psi$  (see Appendix B and [16] for a derivation). We denote  $\Psi_0^{j,i}(\vec{x}_{-t}^j) = \int d\vec{x}_{-t}^j q(\vec{x}_{-t}^j) \mathbf{K}_{\vec{x}_{-t}^j, \vec{x}_{-t}^j}^{j,i}$ , etc.:

$$\begin{aligned} \mathcal{B} &= \int q(\vec{x}_t^i) q(\vec{x}_{-t}^i) \mathcal{A} d\vec{x}_t^i d\vec{x}_{-t}^i \\ &= \sum_{i=1}^M \left( -\frac{1}{2} \alpha^i \sum_{j=1}^M (\alpha^{j,i})^{-2} \text{tr} \left[ \Psi_0^{j,i}(\vec{x}_{-t}^j) - \left( \mathbf{K}_{\vec{z}_t^j, \vec{z}_t^j}^{j,i} \right)^{-1} \Psi_2^{j,i}(\vec{x}_{-t}^j) \right] \right. \\ &\quad - Q_i \log \sqrt{2\pi \alpha^i} - \frac{1}{2} (\alpha^i)^{-1} \left[ \text{tr}(\mathbf{S}_{\vec{x}_t^i}) + \vec{\mu}_{\vec{x}_t^i}^T \vec{\mu}_{\vec{x}_t^i} \right] \\ &\quad + \vec{\mu}_{\vec{x}_t^i}^T \left( \sum_{j=1}^M (\alpha^{j,i})^{-1} \Psi_1^{j,i}(\vec{x}_{-t}^j) \right) \left( \mathbf{K}_{\vec{z}_t^j, \vec{z}_t^j}^{j,i} \right)^{-1} \vec{u}^{j,i} \\ &\quad \left. - \frac{1}{2} \alpha^i \sum_{j=1}^M \sum_{k=1}^M (\alpha^{j,i})^{-1} (\alpha^{k,i})^{-1} \vec{u}^{j,i T} \left( \mathbf{K}_{\vec{z}_t^j, \vec{z}_t^j}^{j,i} \right)^{-1} \Psi_2^{j,k,i}(\vec{x}_{-t}^j, \vec{x}_{-t}^k) \left( \mathbf{K}_{\vec{z}_t^k, \vec{z}_t^k}^{k,i} \right)^{-1} \vec{u}^{k,i} \right) \end{aligned} \tag{A27}$$

Next, we sum this expression over time points:

$$\begin{aligned} \mathcal{C} &= \sum_{t=1}^T \mathcal{B} \\ &= \sum_{i=1}^M \left( \sum_{t=1}^T \left[ -\frac{1}{2} \alpha^i \sum_{j=1}^M (\alpha^{j,i})^{-2} \text{tr} \left[ \Psi_0^{j,i}(\vec{x}_{-t}^j) - \left( \mathbf{K}_{\vec{z}_t^j, \vec{z}_t^j}^{j,i} \right)^{-1} \Psi_2^{j,i}(\vec{x}_{-t}^j) \right] \right. \right. \\ &\quad \left. - \log \sqrt{2\pi \alpha^i} - \frac{1}{2} (\alpha^i)^{-1} \left[ \text{tr}(\mathbf{S}_{\vec{x}_t^i}) + \vec{\mu}_{\vec{x}_t^i}^T \vec{\mu}_{\vec{x}_t^i} \right] \right] \\ &\quad + \sum_{j=1}^M (\alpha^{j,i})^{-1} \left[ \sum_{t=1}^T \vec{\mu}_{\vec{x}_t^i}^T \Psi_1^{j,i}(\vec{x}_{-t}^j) \right] \left( \mathbf{K}_{\vec{z}_t^j, \vec{z}_t^j}^{j,i} \right)^{-1} \vec{u}^{j,i} \\ &\quad \left. - \frac{1}{2} \alpha^i \sum_{j=1}^M \sum_{k=1}^M (\alpha^{j,i})^{-1} (\alpha^{k,i})^{-1} \vec{u}^{j,i T} \left( \mathbf{K}_{\vec{z}_t^j, \vec{z}_t^j}^{j,i} \right)^{-1} \left[ \sum_{t=1}^T \Psi_2^{j,k,i}(\vec{x}_{-t}^j, \vec{x}_{-t}^k) \right] \left( \mathbf{K}_{\vec{z}_t^k, \vec{z}_t^k}^{k,i} \right)^{-1} \vec{u}^{k,i} \right) \end{aligned} \tag{A28}$$

For every part  $i \in 1 \dots M$ , we stack up the  $\vec{u}^{j,i}$  into  $\vec{u}^i$  (first by IV-index  $k = 1, \dots, K$ , and then by part index such that  $\vec{u}_k^{j,i} = (\vec{u}^i)^{Q^i \cdot (K \cdot j + k) + q^i + 1}$ ) and construct a large block matrices  $\mathcal{F}^i$  and stacked vector  $\mathcal{G}^i$  with block elements

$$\mathcal{F}_{j,k}^i = \alpha^i \alpha_{j,i}^{-1} \alpha_{k,i}^{-1} \left( \mathbf{K}_{\vec{z}_t^j, \vec{z}_t^j}^{j,i} \right)^{-1} \left[ \sum_{t=1}^T \Psi_2^{j,k,i}(\vec{x}_{-t}^j, \vec{x}_{-t}^k) \right] \left( \mathbf{K}_{\vec{z}_t^k, \vec{z}_t^k}^{k,i} \right)^{-1} \tag{A29}$$

$$\mathcal{G}_j^i = \left( \alpha_{j,i}^{-1} \left[ \sum_{t=1}^T \vec{\mu}_{\vec{x}_t^i}^T \Psi_1^{j,i}(\vec{x}_{-t}^j) \right] \left( \mathbf{K}_{\vec{z}_t^j, \vec{z}_t^j}^{j,i} \right)^{-1} \right)^T \tag{A30}$$

For  $j \neq k$ :  $\Psi_2^{j,k,i}(\vec{x}_{-t}^j, \vec{x}_{-t}^k) = \Psi_1^{j,i}(\vec{x}_{-t}^j) \Psi_1^{k,i}(\vec{x}_{-t}^k)$ . Otherwise,  $\Psi_2^{j,j,i}(\vec{x}_{-t}^j, \vec{x}_{-t}^j) = \Psi_2^{j,i}(\vec{x}_{-t}^j)$ . We rewrite  $\mathcal{C}$  as a quadratic form in the stacked augmenting IVs  $\vec{u}^i$  to facilitate closed-form optimization of the dynamics ELB in Equation (A22) with respect to the stacked IV density  $q(\vec{u}^i)$  using variational calculus, as described in Appendix A:

$$\begin{aligned} \mathcal{C} &= \sum_{i=1}^M \left[ -\frac{1}{2} \bar{u}^{iT} \mathcal{F}^i \bar{u}^i + \bar{u}^{iT} \mathcal{G}^i + \mathcal{H}^i \right] \\ &= \sum_{i=1}^M \left[ -\frac{1}{2} (\bar{u}^i - \mathcal{F}^{i-1} \mathcal{G}^i)^T \mathcal{F}^i (\bar{u}^i - \mathcal{F}^{i-1} \mathcal{G}^i) + \frac{1}{2} \mathcal{G}^{iT} \mathcal{F}^{i-1} \mathcal{G}^i + \mathcal{H}^i \right] \end{aligned} \tag{A31}$$

$$\mathcal{C} = \sum_{i=1}^M \mathcal{C}^i \tag{A32}$$

$$\mathcal{C}^i = -\frac{1}{2} (\bar{u}^i - \mathcal{F}^{i-1} \mathcal{G}^i)^T \mathcal{F}^i (\bar{u}^i - \mathcal{F}^{i-1} \mathcal{G}^i) + \frac{1}{2} \mathcal{G}^{iT} \mathcal{F}^{i-1} \mathcal{G}^i + \mathcal{H}^i \tag{A33}$$

$$\begin{aligned} \mathcal{H}^i &= \sum_{t=1}^T \left[ -\frac{1}{2} \alpha_i \sum_{j=1}^M (\alpha_{j,i}^{-1})^2 \text{tr} \left[ \Psi_0^{j,i}(\bar{x}_{-t}^j) - \left( \mathbf{K}_{\bar{z}_t^j, \bar{z}_t^j}^{j,i} \right)^{-1} \Psi_2^{j,i}(\bar{x}_{-t}^j) \right] \right. \\ &\quad \left. - \log \sqrt{2\pi \alpha^i} - \frac{1}{2} \alpha_i^{-1} \left[ \text{tr}(\mathbf{S}_{\bar{x}_t^i}) + \bar{\mu}_{\bar{x}_t^i}^T \bar{\mu}_{\bar{x}_t^i} \right] \right] \end{aligned} \tag{A34}$$

After this optimization, we can write the dynamics ELBO using  $p(\bar{u}^i) = \prod_{j=1}^M p(\bar{u}^{j,i}) = \prod_{j=1}^M \mathcal{N}(\bar{u}^{j,i} | 0, \mathbf{K}_{\bar{z}_t^j, \bar{z}_t^j}^{j,i}) = \mathcal{N}(\bar{u}^i | 0, \mathbf{K}_{\bar{z}_t^i, \bar{z}_t^i}^i)$  where  $\mathbf{K}_{\bar{z}_t^i, \bar{z}_t^i}^i$  is a block-diagonal covariance matrix with the blocks given by the individual  $\mathbf{K}_{\bar{z}_t^j, \bar{z}_t^j}^{j,i}$ :

$$\begin{aligned} \mathcal{L}_{dyn}(\Theta) &\geq \log \int p(\bar{u}^i) \exp(\mathcal{C}) d\bar{u}^i + H(q(\bar{x}^i)) \\ &= \log \prod_{i=1}^M \int p(\bar{u}^i) \exp(\mathcal{C}^i) d\bar{u}^i + H(q(\bar{x}^i)) \\ &= \sum_{i=1}^M \left[ \log \int p(\bar{u}^i) \exp\left(-\frac{1}{2} (\bar{u}^i - \mathcal{F}^{i-1} \mathcal{G}^i)^T \mathcal{F}^i (\bar{u}^i - \mathcal{F}^{i-1} \mathcal{G}^i) d\bar{u}^i + \frac{1}{2} \mathcal{G}^{iT} \mathcal{F}^{i-1} \mathcal{G}^i + \mathcal{H}^i \right) \right] + H(q(\bar{x}^i)) \tag{A35} \\ &= \sum_{i=1}^M \left[ -\log \sqrt{(2\pi)^{\dim(\mathcal{F}^{i-1})} |\mathcal{F}^{i-1} + \mathbf{K}_{\bar{z}_t^i, \bar{z}_t^i}^i|} - \frac{1}{2} \mathcal{G}^{iT} \mathcal{F}^{i-1} (\mathcal{F}^{i-1} + \mathbf{K}_{\bar{z}_t^i, \bar{z}_t^i}^i)^{-1} \mathcal{F}^{i-1} \mathcal{G}^i + \log Z(\mathcal{F}^{i-1}) \right] \\ &\quad + \sum_{i=1}^M \left[ \frac{1}{2} \mathcal{G}^{iT} \mathcal{F}^{i-1} \mathcal{G}^i + \mathcal{H}^i \right] + H(q(\bar{x}^i)) \end{aligned}$$

This is the expression which we optimize with respect to  $q(\bar{x}^i)$  and  $q(\bar{u}^i)$ . Since the stacked dynamics IVs  $\bar{u}^i$  do not interact across parts in this expression (Line 2), it follows that density  $q(\bar{u}^i)$  factorizes across parts. Their optimal density for each part is given by (cf. Equation (A6),  $Z$  is the normalization constant of the multivariate Gaussian):

$$\begin{aligned} q(\bar{u}^i) &= \frac{1}{Z} p(\bar{u}^i) \exp(\mathcal{C}^i) \\ &= \frac{1}{Z} \mathcal{N}(\bar{u}^i | 0, \mathbf{K}_{\bar{z}_t^i, \bar{z}_t^i}^i) \exp\left(-\frac{1}{2} (\bar{u}^i - \mathcal{F}^{i-1} \mathcal{G}^i)^T \mathcal{F}^i (\bar{u}^i - \mathcal{F}^{i-1} \mathcal{G}^i)\right) \tag{A36} \\ &= \mathcal{N}(\bar{u}^i | (\mathbf{K}_{\bar{z}_t^i, \bar{z}_t^i}^{-1} + \mathcal{F}^i)^{-1} \mathcal{G}^i, (\mathbf{K}_{\bar{z}_t^i, \bar{z}_t^i}^{-1} + \mathcal{F}^i)^{-1}) \end{aligned}$$

**References**

1. Bizzi, E.; Cheung, V.; d’Avella, A.; Saltiel, P.; Tresch, M. Combining modules for movement. *Brain Res. Rev.* **2008**, *57*, 125–133. [CrossRef] [PubMed]
2. Endres, D.; Chiovetto, E.; Giese, M. Model selection for the extraction of movement primitives. *Front. Comput. Neurosci.* **2013**, *7*, 185. [CrossRef] [PubMed]
3. Schaal, S. Dynamic Movement Primitives -A Framework for Motor Control in Humans and Humanoid Robotics. In *Adaptive Motion of Animals and Machines*; Kimura, H., Tsuchiya, K., Ishiguro, A., Witte, H., Eds.; Springer: Tokyo, Japan, 2006; pp. 261–280. [CrossRef]

4. Giese, M.A.; Mukovskiy, A.; Park, A.N.; Omlor, L.; Slotine, J.J.E. Real-Time Synthesis of Body Movements Based on Learned Primitives. In *Statistical and Geometrical Approaches to Visual Motion Analysis*; Cremers, D., Rosenhahn, B., Yuille, A.L., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; Volume 5604, pp. 107–127.
5. Koch, K.H.; Clever, D.; Mombaur, K.; Endres, D.M. Learning Movement Primitives from Optimal and Dynamically Feasible Trajectories for Humanoid Walking. In Proceedings of the IEEE/RAS International Conference on Humanoid Robots (Humanoids 2015), Seoul, Korea, 3–5 November 2015; pp. 866–873.
6. Chiovetto, E.; Berret, B.; Pozzo, T. Tri-dimensional and triphasic muscle organization of whole-body pointing movements. *Neuroscience* **2010**, *170*, 1223–1238. [[CrossRef](#)] [[PubMed](#)]
7. Omlor, L.; Giese, M.A. Anechoic Blind Source Separation using Wigner Marginals. *J. Mach. Learn. Res.* **2011**, *12*, 1111–1148.
8. Chiovetto, E.; Giese, M.A. Kinematics of the coordination of pointing during locomotion. *PLoS ONE* **2013**, *8*, e79555. [[CrossRef](#)] [[PubMed](#)]
9. Clever, D.; Harant, M.; Koch, K.H.; Mombaur, K.; Endres, D.M. A novel approach for the generation of complex humanoid walking sequences based on a combination of optimal control and learning of movement primitives. *Robot. Autom. Syst.* **2016**, *83*, 287–298. [[CrossRef](#)]
10. Mussa-Ivaldi, F.A.; Solla, S.A. Neural Primitives for Motion Control. *IEEE J. Ocean. Eng.* **2004**, *29*, 640–650. [[CrossRef](#)]
11. Hart, C.B.; Giszter, S. Distinguishing Synchronous and Time Varying Synergies using Point Process Interval Statistics: Motor Primitives in Frog and Rat. *Front. Comput. Neurosci.* **2013**, *7*, 52. [[CrossRef](#)] [[PubMed](#)]
12. Ijspeert, A.J.; Nakanishi, J.; Hoffmann, H.; Pastor, P.; Schaal, S. Dynamical Movement Primitives: Learning Attractor Models for Motor Behaviors. *Neural Comput.* **2013**, *25*, 328–373. [[CrossRef](#)] [[PubMed](#)]
13. Rückert, E.; d’Avella, A. Learned Parameterized Dynamic Movement Primitives with Shared Synergies for Controlling Robotic and Musculoskeletal Systems. *Front. Comput. Neurosci.* **2013**, *7*, 138. [[CrossRef](#)] [[PubMed](#)]
14. Velychko, D.; Endres, D.; Taubert, N.; Giese, M.A. Coupling Gaussian Process Dynamical Models with Product-of-Experts Kernels. In *International Conference on Artificial Neural Networks*; Springer: Cham, Switzerland, 2014; pp. 603–610. [[CrossRef](#)]
15. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*; The MIT Press: Cambridge, MA, USA, 2005.
16. Titsias, M.K.; Lawrence, N.D. Bayesian Gaussian Process Latent Variable Model. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS), Sardinia, Italy, 13–15 May 2010; pp. 844–851.
17. Bizzi, E.; Cheung, V.C. The Neural Origin of Muscle Synergies. *Front. Comput. Neurosci.* **2013**, *7*, 51. [[CrossRef](#)] [[PubMed](#)]
18. Földiák, P.; Endres, D. Sparse coding. *Scholarpedia* **2008**, *3*, 2984. [[CrossRef](#)]
19. Velychko, D.; Knopp, B.; Endres, D. The variational coupled Gaussian Process Dynamical Model. In *International Conference on Artificial Neural Networks*; Springer: Cham, Switzerland, 2017; pp. 291–299. [[CrossRef](#)]
20. Candela, J.Q.; Rasmussen, C.E. A Unifying View of Sparse Approximate Gaussian Process Regression. *J. Mach. Learn. Res.* **2005**, *6*, 1939–1959.
21. Snelson, E.; Ghahramani, Z. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems 18*; MIT Press: Cambridge, MA, USA, 2006; pp. 1257–1264.
22. Lawrence, N.D. Learning for Larger Datasets with the Gaussian Process Latent Variable Model. *Artif. Intell. Stat.* **2007**, *2*, 243–250.
23. Titsias, M.K. Variational Learning of Inducing Variables in Sparse Gaussian Processes. *Artif. Intell. Stat.* **2009**, *5*, 567–574.
24. Damianou, A.C.; Titsias, M.; Lawrence, N.D. Variational Gaussian Process Dynamical Systems. In *Advances in Neural Information Processing Systems 24*; Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K., Eds.; MIT Press: Cambridge, MA, USA, 2011; pp. 2510–2518.
25. Wang, J.M.; Fleet, D.J.; Hertzmann, A. Gaussian Process Dynamical Models for Human Motion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 283–298. [[CrossRef](#)] [[PubMed](#)]
26. Urtasun, R.; Fleet, D.J.; Lawrence, N.D. Modeling Human Locomotion with Topologically Constrained Latent Variable Models. In *Human Motion—Understanding, Modeling, Capture and Animation*; Springer: Berlin/Heidelberg, Germany, 2007; Volume 4814, pp. 104–118.

27. Taubert, N.; Endres, D.; Christensen, A.; Giese, M.A. Shaking Hands in Latent Space. In *Annual Conference on Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2011; Volume 7006, pp. 330–334.
28. Levine, S.; Wang, J.M.; Haraux, A.; Popović, Z.; Koltun, V. Continuous Character Control with Low-Dimensional Embeddings. *ACM Trans. Graph.* **2012**, *31*, 28. [[CrossRef](#)]
29. Chen, J.; Kim, M.; Wang, Y.; Ji, Q. Switching Gaussian Process Dynamic Models for simultaneous composite motion tracking and recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2009 (CVPR 2009), Miami, FL, USA, 20–25 June 2009; pp. 2655–2662. [[CrossRef](#)]
30. Mattos, C.L.C.; Dai, Z.; Damianou, A.; Forth, J.; Barreto, G.A.; Lawrence, N.D. Recurrent Gaussian Processes. *arXiv* **2015**, arxiv:1511.06644.
31. Frigola, R.; Lindsten, F.; Schön, T.B.; Rasmussen, C. Bayesian Inference and Learning in Gaussian Process State-Space Models with Particle MCMC. In *Advances in Neural Information Processing Systems 26*; Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2013; pp. 3156–3164.
32. Frigola, R.; Chen, Y.; Rasmussen, C. Variational Gaussian Process State-Space Models. In *Advances in NIPS 27*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K., Eds.; NIPS Foundation: Montreal, QC, Canada, 2014; pp. 3680–3688.
33. Bauer, M.; van der Wilk, M.; Rasmussen, C. Understanding Probabilistic Sparse Gaussian Process Approximations. *arXiv* **2016**, arxiv:1606.04820.
34. Taubert, N.; Christensen, A.; Endres, D.; Giese, M. Online simulation of emotional interactive behaviors with hierarchical Gaussian process dynamical models. In Proceedings of the ACM Symposium on Applied Perception (SAP '12), Los Angeles, CA, USA, 3–4 August 2012; pp. 25–32. [[CrossRef](#)]
35. Hinton, G.E. Products of Experts. In Proceedings of the 9th International Conference on Artificial Neural Networks (ICANN'99), Edinburgh, UK, 7–10 September 1999; Volume 1, pp. 1–6.
36. Bishop, C.M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*; Springer: Secaucus, NJ, USA, 2006.
37. Bastien, F.; Lamblin, P.; Pascanu, R.; Bergstra, J.; Goodfellow, I.J.; Bergeron, A.; Bouchard, N.; Bengio, Y. Theano: New features and speed improvements. *arXiv* **2012**, arxiv:1211.5590.
38. Jones, E.; Oliphant, T.; Peterson, P. SciPy: Open Source Scientific Tools for Python. Available online: <https://www.scipy.org/> (accessed on 9 October 2015).
39. Sakoe, H.; Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech Signal Process.* **1978**, *26*, 43–49. [[CrossRef](#)]
40. Taubert, N.; Löffler, M.; Ludolph, N.; Christensen, A.; Endres, D.; Giese, M. A virtual reality setup for controllable, stylized real-time interactions between humans and avatars with sparse Gaussian process dynamical models. In Proceedings of the ACM Symposium on Applied Perception (SAP '13), Dublin, Ireland, 22–23 August 2013; pp. 41–44. [[CrossRef](#)]
41. Paraschos, A.; Daniel, C.; Peters, J.; Neumann, G. Probabilistic Movement Primitives. In *Advances in NIPS 26*; Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K., Eds.; NIPS Foundation: Montreal, QC, Canada, 2013; pp. 2616–2624.
42. Pastor, P.; Kalakrishnan, M.; Righetti, L.; Schaal, S. Towards Associative Skill Memories. In Proceedings of the 2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids), Osaka, Japan, 29 November–1 December 2012; pp. 309–315.
43. Land, W.M.; Rosenbaum, D.A.; Seegelke, C.; Schack, T. Whole-body posture planning in anticipation of a manual prehension task: Prospective and retrospective effects. *Acta Psychol.* **2013**, *144*, 298–307. [[CrossRef](#)] [[PubMed](#)]
44. Deisenroth, M.; Fox, D.; Rasmussen, C. Gaussian Processes for Data-Efficient Learning in Robotics and Control. *Pattern Anal. Mach. Intell. IEEE Trans.* **2015**, *37*, 408–423. [[CrossRef](#)] [[PubMed](#)]
45. Petersen, K.B.; Pedersen, M.S. *The Matrix Cookbook*; Version 20121115; Technical University of Denmark: Lyngby, Denmark, 2012.



D

Predicting Perceived Naturalness of  
Human Animations Based on Generative  
Movement Primitive Models

# Predicting Perceived Naturalness of Human Animations Based on Generative Movement Primitive Models

BENJAMIN KNOPP, DMYTRO VELYCHKO, JOHANNES DREIBRODT, and  
DOMINIK ENDRES, University of Marburg

We compared the perceptual validity of human avatar walking animations driven by six different representations of human movement using a graphics Turing test. All six representations are based on movement primitives (MPs), which are predictive models of full-body movement that differ in their complexity and prediction mechanism. Assuming that humans are experts at perceiving biological movement from noisy sensory signals, it follows that these percepts should be describable by a suitably constructed Bayesian ideal observer model. We build such models from MPs and investigate if the perceived naturalness of human animations are predictable from approximate Bayesian model scores of the MPs. We found that certain MP-based representations are capable of producing movements that are perceptually indistinguishable from natural movements. Furthermore, approximate Bayesian model scores of these representations can be used to predict perceived naturalness. In particular, we could show that movement dynamics are more important for perceived naturalness of human animations than single frame poses. This indicates that perception of human animations is highly sensitive to their temporal coherence. More generally, our results add evidence for a shared MP-representation of action and perception. Even though the motivation of our work is primarily drawn from neuroscience, we expect that our results will be applicable in virtual and augmented reality settings, when perceptually plausible human avatar movements are required.

CCS Concepts: • **Computing methodologies** → **Perception**; *Animation*; *Motion processing*; • **Theory of computation** → *Gaussian processes*;

Additional Key Words and Phrases: Human animation, movement primitives, perception, dynamical systems, psychophysics, Gaussian process dynamical model, dynamical movement primitives

## ACM Reference format:

Benjamin Knopp, Dmytro Velychko, Johannes Dreibrodt, and Dominik Endres. 2019. Predicting Perceived Naturalness of Human Animations Based on Generative Movement Primitive Models. *ACM Trans. Appl. Percept.* 16, 3, Article 15 (September 2019), 18 pages.

<https://doi.org/10.1145/3355401>

This work was funded by DFG, IRTG1901 - The brain in action, and SFB-TRR 135 - Cardinal mechanisms of perception.

Authors' addresses: B. Knopp, D. Velychko, J. Dreibrodt, and D. Endres, Department of Psychology, University of Marburg, Gutenbergstraße 18, 35039 Marburg; emails: {benjamin.knopp, dmytro.velychko}@uni-marburg.de, dreibrod@students.uni-marburg.de, dominik.endres@uni-marburg.de.



This work is licensed under a [Creative Commons Attribution International 4.0 License](https://creativecommons.org/licenses/by/4.0/).

© 2019 Copyright held by the owner/author(s).

1544-3558/2019/09-ART15

<https://doi.org/10.1145/3355401>

## 1 INTRODUCTION

The perception of biological movement<sup>1</sup> is of paramount importance for humans: in many situations, in real life as well as in virtual reality, it is necessary to predict internal states and goals of other actors from observed body movements. Such predictions are facilitated by a model of relevant degrees of freedom (DOF), and the abstraction of redundant ones. Strong evidence for the existence of such a model from a neuroscientific perspective is provided by the point-light walker experiments of Johansson (1994): just a few dots resembling the human body’s spatial configuration and dynamics are enough for robust detection of activities like walking, dancing, and the like. Practical evidence is given by the everlasting struggle of animators to produce perceptually valid human animations (without relying on motion captured data).

A related abstraction problem must be solved in motor production: our bodies have many more DOFs than needed for any given movement (Bernstein 1967); hence, the redundant DOFs need to be bound or remain uncontrolled. One way to bind these DOFs is via *movement primitives* (MPs) or synergies, as predicted by optimal control feedback theory (Todorov and Jordan 2003).

This relationship between movement perception and production suggests that a shared representation might be used to address them both, as proposed by the *common coding* hypothesis and the theory of event coding (Friston 2010; Hommel et al. 2001; Prinz 1997; Shin et al. 2010; Wolpert et al. 2003). However, this hypothesis does not specify the level of representation on which the common coding happens. We therefore investigate whether MPs are candidates for such a shared representation. Their suitability for complex movement production has already been demonstrated (Clever et al. 2017; Giszter 2015; Ijspeert et al. 2013; Omlor and Giese 2011), we would like to determine how close human perceptual performance is to an “ideal observer” comprised of MPs.

The “ideal observer” assumption is motivated by the apparent ease with which we perceive and interpret our fellow humans’ movements: we hypothesize that movement perception is another instance where we behave nearly Bayes-optimally (Knill and Pouget 2004). Hence, human perceptual expectations should be predictable by Bayesian model comparison between MP models. To test this hypothesis, we trained generative MP models on kinematic data of walking movements, and compared movements based on these MPs in a Graphics Turing Test. We are also interested in determining the model scores which are most predictive of human expectations.

## 2 RELATED WORK

Biological motion perception induced by point-light-stimuli is a related, and heavily investigated research topic (for an overview, see Troje (2013)); point-light stimuli, first introduced to demonstrate the perceptual binding of different points to one “Gestalt” (Johansson 1994), they have been used to study the perception of movement isolated from body shape and other cues (Bertenthal and Pinto 1994; Casile and Giese 2005; Troje 2002; Troje et al. 2005).

We are not concerned with the shape inference process from point-light-displays or stick figures, therefore we use 3D avatars, which are closer to natural stimuli. It has been shown that human observers have a higher sensitivity for detecting differences in movement when using 3D avatars compared to stick figures (Hodgins et al. 1998).

Motivation to use MPs as perceptual representations of movement is given by an action-perception coupling on the neural level (Dayan et al. 2007): the famous “2/3 power law”, an observed invariant in curved drawing movements, seems to have a perceptual representation in the brain. Parabolic MPs can simultaneously obey the 2/3 power law and minimize jerk, which has been proposed as a control principle for arm movements (Polyakov et al. 2009). Perceptual experiments investigating the segmentation of taekwondo solo forms imply that higher order polynomial MPs might be more appropriate perceptual descriptors for full-body movement (Endres et al. 2011).

<sup>1</sup>The term “biological motion” has been used to denote a point-light display of (biological) movement. We use the term ‘human animation’ for a 3D-rendered display of movement.

In an experiment similar to ours, it has been shown that hierarchical Gaussian process dynamical models can synthesize hand shake movements indistinguishable from natural ones (Taubert et al. 2012). Furthermore, the perception of emotion based on spatio-temporal MPs has been investigated by Roether et al. (2009) and Chiovetto et al. (2018). In our study, we are interested in comparing different MP types in a unified Bayesian framework (Endres et al. 2013) with respect to the perception of naturalness.

### 3 MODELS AND EXPERIMENTAL METHODS

In this section, we first introduce the investigated MP models, which are used to generate the stimuli for graphics Turing test (McGuigan 2006). Next, we describe our experiment designed to determine the perceived naturalness of the generated walking movements. Finally, we explain the data analysis methods used to predict the perceived naturalness from approximate Bayesian model scores.

#### 3.1 Movement Primitives

MPs refer to building blocks of complex movements, but there is little consensus on an exact definition. Consequently, many different types of MPs have been proposed in literature (Endres et al. 2013). These types can be classified as spatial (Giszter et al. 1992; Tresch et al. 1999), temporal (Clever et al. 2016; Endres et al. 2013), spatio-temporal (d’Avella et al. 2003; Omlor and Giese 2011) and dynamical MPs (Ijspeert et al. 2013).

We focus on dynamical and temporal MPs in this study, as we are interested in finding a higher level representation suitable for modeling perception, as opposed to spatial MPs, which have been used to model muscle synergies in the spinal chord (Giszter 2015). Anechoic mixture models have been proposed to enable phase shifted combinations of MPs (Chiovetto et al. 2018; Omlor and Giese 2011). We do not explicitly test this type of MP here, since the relative phase shifts the walking movements we studied are negligible.

We perceptually validate 6 generative MP models: Temporal MPs, Dynamical MPs and 4 flavors of the Gaussian Process Dynamical Model (GPDM) (Velychko et al. 2018; Wang et al. 2008): GPDM, variational GPDM, coupled GPDM, and variational coupled GPDM.

In this section, we can only provide a rough overview, just enough to enable readers from different backgrounds to understand parameters of the stimuli for the psychophysical experiment. Please refer to the cited papers for detailed information. Velychko et al. (2018) also provide graphical model representations and summarize the features of the MP models presented in this chapter.

**3.1.1 Temporal Movement Primitives (TMP) (Clever et al. 2016).** Temporal MPs describe the stereotyped temporal patterns of movement parameters (for example EMG, but also joint trajectories as well as endpoint trajectories). A possible biological implementation of temporal MPs might be central pattern generators (CPGs) (Ivanenko et al. 2004) combined with cortical top-down control. Temporal MPs incorporate a temporal predictive mechanism: the complete time-course of the movement is determined at its onset. This type of MPs allows for simple concatenation and temporal scaling.

The trajectory  $x_k(t)$  of a DOF  $X_k$ , e.g., a joint angle, is a weighted sum of  $Q$  MPs  $Y_q$ , which are functions of time  $y_q(t)$ .  $\varepsilon_i(t) \sim \mathcal{N}(0, \sigma_i)$  is Gaussian observation noise:

$$x_k(t) = \sum_{q=1}^Q w_{k,q} y_q(t) + \varepsilon_i(t). \quad (1)$$

We treat the number of MPs as ideal observer model parameter to be determined. In general, more MPs allow for more fine-grained temporal structure of the movement, but might lead to over-fitting. To determine the MPs and their number, we follow the approach of Clever et al. (2016): weights  $w$  and MPs  $Y_q$  have a Gaussian Process (GP) prior and are learned from the training data by maximizing a variational lower bound on the Bayesian model evidence (ELBO, evidence lower bound). The ELBO is equal to the negative free energy (Friston 2010). In

keeping with the free energy/Bayesian brain theory, one would therefore expect that the ELBO should be useful for selecting the appropriate number of MPs  $Q$  for the generation of perceptually valid movements.

**3.1.2 Dynamic Movement Primitives (DMP) (Ijspeert et al. 2013).** While temporal MPs directly model the movement parameters (e.g., trajectories or muscle activations), DMPs describe the stereotyped elements of movement as attractors of a dynamical system, thus enabling the prediction of the next state from the previous ones. Building on the hypothesis of separate brain areas for rhythmic and discrete movements, two kinds of dynamical systems are common: cyclic oscillators and point attractors (Schaal 2006).

More formally: DMP models represent a movement trajectory  $x_k(t)$  obeying a differential equation. They rely on a damped spring system which forces  $x_k(t)$  to contract to the specified goal  $g_k$ , if the dampening factor is high enough. Through the non-linear forcing function  $f_k$  (Equation (2)) the trajectories can be modified. This function is modeled as weighted sum of Gaussian basis functions  $\Psi_i(\tau)$  (Equation (4)). Time is replaced by  $\tau$ , which decays exponentially to zero (Equation (3)). DMPs are learned from training data by setting the weights  $w_i$  such that the training mean-squared error (MSE) is minimal.

$$\tau \ddot{x}_k = \alpha_z(\beta_z(g_k - x_k) - \dot{x}_k) + f_k(\tau) \quad (2)$$

$$\dot{\tau} \propto -\tau \quad (3)$$

$$f_k(\tau) = \frac{\sum_{i=1}^N \Psi_i(\tau) w_{k,i}}{\sum_{i=1}^N \Psi_i(\tau)} \tau (g_k - x_k(0)). \quad (4)$$

The number of basis functions  $N$  is the ideal observer model complexity parameter. It serves a similar role as the number of MPs in the TMP model: more basis functions allow for more complicated forcing functions, which enable richer temporal dynamics. The number can, e.g., be selected by cross-validation, we investigate if  $N$  reflects the perceived naturalness.

**3.1.3 Gaussian Process Dynamical Model (GPDM) (Wang et al. 2008).** Learnable dynamical systems for movement representation have been proposed in the context of computer graphics: the GPDM is a state-space model, which learns a dynamical mapping in a latent space of the whole-body movement. Such a model is also physiologically attractive, because it is able to reflect the dynamic nature of the environment and the body itself, without explicit assumptions of their form (Shenoy et al. 2013; Sussillo et al. 2015).

In contrast to DMPs, GPDMs learn a full dynamical model of latent variables  $Y$  in discrete time, which are mapped onto the observed DOFs  $X_k$ . Both the dynamics mapping  $f()$  (Equation (5)), as well as the mapping from latent to observed space  $g()$  (Equation (6)) are drawn from Gaussian process priors, hence the name.  $dt$  denotes the time discretization step-size:

$$y(t) = f(y(t - dt)) + \varepsilon_{y,t}, \quad (5)$$

$$x_k(t) = g_k(y(t)) + \varepsilon_{x,t}. \quad (6)$$

There are two main drawbacks which make the GPDM unlikely as a perceptual MP model: (1) there is no (obvious) way of a recombination operation that would make GPDMs modular. Modularity here refers to the possibility of generating a large repertoire of movements from the recombination of a small number of MPs. (2) Due to the non-parametric GPs prior, the movements *are* the movement representation, which is not compact.

A further consequence of this non-parametric prior is no explicit ideal observer model complexity parameter. Therefore, we compare the GPDM estimated by maximum *a-posteriori* inference (MAP) with the other movement primitive representations. The GPDM can also be trained by variational inference, giving rise to the vGPDM. This is a special case of the variational coupled GPDM described in 3.1.5.

**3.1.4 Coupled Gaussian Process Dynamical Model (cGPDM) (Velychko et al. 2014).** The cGPDM was proposed to make GPDMs modular. Here, one learns different dynamical models for different body parts. Each body part is described by a GPDM, where the latent variables predict not only the next time-step of their associated body part,

but also the temporal evolution of other body parts via coupling functions. This way, flexible coupling between body parts is possible. The vCGPDM can be regarded as a middle ground between DMPs encoding single DOFs, and the monolithic GPDM. The latent dynamical systems can thus be thought of as flexibly coupled CPGs routing commands to the muscles.

As with the MAP-trained GPDM introduced in the previous section, there is no explicit ideal observer model complexity parameter in the MAP-trained cGPDM.

**3.1.5 Variational (Coupled) Gaussian Process Dynamical Model (v(C)GPDM) (Velychko et al. 2018).** The vCGPDM compresses the movement representation of cGPDMs by introducing sparse variational approximations with a deterministic learning scheme. Here, each MP is parameterized by a small set of inducing points (IPs) and associated inducing values (IVs), leading to a compact representation with constant storage requirements. Flexible recombination of these IPs/IVs for each body part enables the required modularity. The initial choice of IPs/IVs is the only remaining source of stochasticity in the training process. It may have measurable effects, as we will show below. We use IPs for both mappings, serving as ideal observer model parameters: “dynamics” IPs for the dynamical model mapping, and “pose” IPs for the latent-to-observed variable mapping. More dynamics IPs allow for richer dynamics (similar to the parameters of DMP and TMP), while more pose IPs will allow for more (spatial) variability of poses.

An IP/IV pair might be thought of as a prototypical example for the mappings drawn from their associated Gaussian process. They thus provide some abstraction from the observed movement and might be implemented by small neuronal populations. Similar to the TMP, the vCGPDM is trained by maximizing an ELBO. The ELBO can be decomposed into one summand per part that describes the quality of the latent-to-observed mapping (“pose ELBO”) and one summand for the dynamics mapping (“dynamics ELBO”).

In our experiments, we set the number of body parts to  $M = 2$  with one part corresponding to the upper body and one to the lower. By setting  $M = 1$ , we recover a variational version of the GPDM, denoted vGPDM.

## 3.2 Experiment

Our experiment was split in two parts, with the second part’s parameter choices based on the results of the first part. Next, we describe the participants, the generation of stimuli, and then we detail the experimental paradigm.

**3.2.1 Participants.** We invited 31 participants to participate in the first part of the experiment via our participant management system (SONA System) and the university’s mailing list. Due to technical problems, we excluded one participant from the analysis. The remaining 21 female and 9 male participants were between 19 and 44 years old ( $\mu = 24.7a$ ,  $\sigma = 5.8a$ ). Based on the results of this first part, we invited 26 participants to perform the second part of the experiment (19 female, age between 19 and 37 years,  $\mu = 23.9a$ ,  $\sigma = 4.2a$ ). All participants had normal or corrected-to-normal vision and received course credit or financial compensation (8€/h) for participation. The experimental procedures were approved by the local ethics committee and the study was conducted in accordance with the Declaration of Helsinki. Informed written consent was given by all participants prior to the experiment.

**3.2.2 Stimuli.** We employed a 10-camera PhaseSpace Impulse motion capture system to capture walking movements of an actor, and used our skeleton estimation software (Velychko and Endres 2017) to estimate a skeleton geometry with 18 joints, pose (Euler angles of each bone relative to the corresponding parental bone) and position and rotation of the pelvis bone. The results were stored in the Biovision Hierarchical Data format (bvh). From these data, we selected 49 sequences containing 3 gait cycles.

We used all 49 walking sequences to render the natural stimuli. Using the trained models, we generated 1,758 movement sequences (see next subsection), which served as artificial stimuli. Given the natural and generated bvh-files, we used Autodesk MotionBuilder to animate a gray avatar (see Figure 1) with body size and shape similar to the actor. We then rendered these animations into the videos used as stimuli. All resulting stimuli have

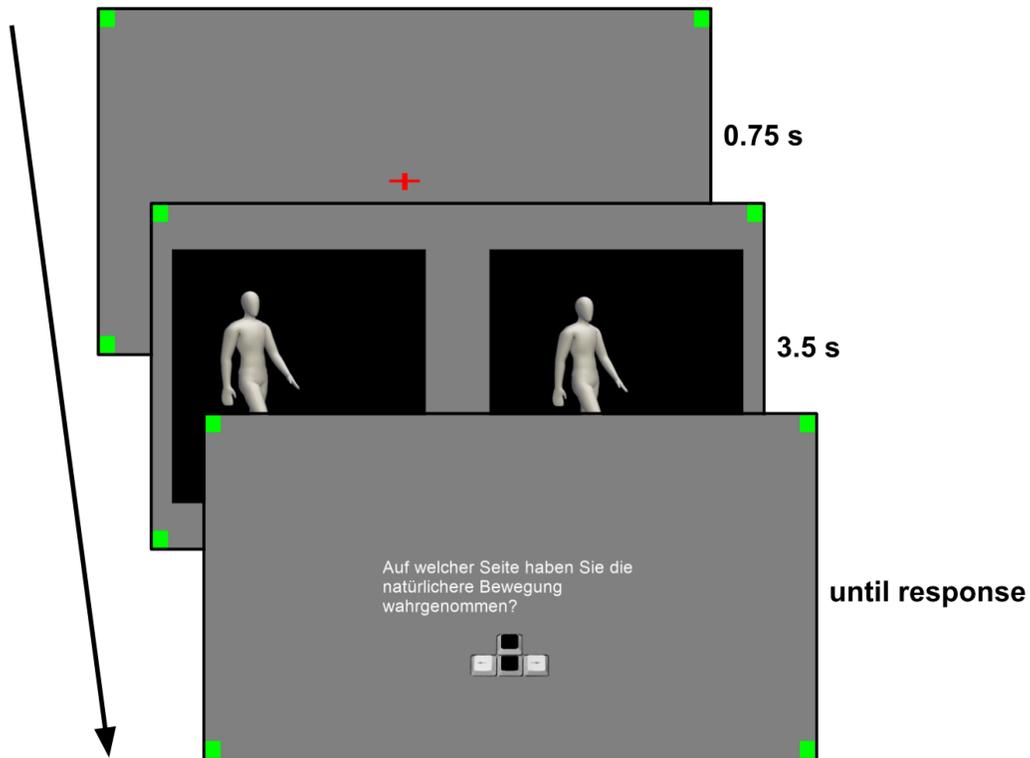


Fig. 1. Illustration of experimental procedure. Each trial begun with a fixation period of 0.75s. Then, participants watched simultaneous replays of natural and generated movements for 3.5s. After the presentation the participants were asked “On which side did you perceive the more natural movement?” and responded using the arrow keys of an keyboard.

a length of 3.5s with 60 frames per second. We supplied a demo video of some example trials in the supplementary material to give the reader a good impression of the stimuli and the task.

**3.2.3 Stimulus Generation.** We trained each MP model on nine gait sequences, and used the trained model to predict a tenth sequence. This enabled us to compute a leave-one-out cross-validation score for each model. Furthermore, the predicted sequence of joint angles was used for stimulus generation, as described above. Dynamical models were initialized with starting conditions taken from the training data. Sometimes the training procedure failed, because it is dependent on random initial values of the optimization algorithm. We hand-labeled obvious failures (e.g., sliding, limping, jerking, (see suppl. mat. first trial for an example)), excluding them from the data analysis, but retaining them to enable us to check the attention of the participants. Tables 1 and 2 summarize the tested models and ideal observer parameters. A more detailed description of the training procedure can be found in Velychko et al. (2018). We trained each model until the training target (ELBO or training MSE) did not change within machine precision anymore, but at most for one day. Most models were done training in a much shorter time.

**3.2.4 Procedure.** Participants were asked to distinguish between natural and generated movements in a two-alternative forced-choice task. For this, we designed an experiment using PsychoPy (Peirce 2009). During the experiment, participants were sitting in front of a 24-inch computer screen. After reading the written instructions, each trial proceeded as follows: (1) a fixation cross appeared for 0.75s, (2) followed by simultaneous side-by-side

presentation of generated and natural stimuli for 3.5s, and (3) finally collecting the participant’s response, indicating on which side the more natural stimulus was perceived. Participants were instructed to use the arrow keys of a standard computer keyboard to submit their answer. They used the left index finger for the left arrow key, and the right index finger for the right arrow key. Both avatars were walking in the same direction, which was drawn randomly for each trial (see Figure 1).

Each participant of the first part of the experiment carried out 643 trials in four blocks, which took approximately 90 minutes. With these 643 trials, 119 models were evaluated: each participant rated 1 to 10 artificial stimuli randomly drawn from the total set of 10 artificial stimuli for each model. These were tested against a randomized repetition of 44 natural stimuli. To test whether participants simply memorized the natural stimuli during the experiment, we added 6 catch trials in the last quarter of the experiment where previously unused natural movements were tested against the known natural stimuli.

For the second part of the experiment, we split the total number of 629 trials into two conditions with 314 and 315 trials, allowing the participants to participate in one or both at their convenience. Participants were distributed equally among both conditions. Each condition was split into 7 blocks, with 30s pauses in between. After the first part of the experiment, we determined that memorization effects could be disregarded. Hence, we decided not to use catch trials in the second part. Sixty-seven models were tested in each condition. The available artificial stimuli for each model were distributed equally between conditions, and presented randomized for each participant.

### 3.3 Data Analysis

The rationale of the experiment is as follows: after simultaneous presentation of artificial and natural (motion-capture-based) human animations, the participant is forced to choose the one perceived as more natural. The answer is communicated via key press. In each trial  $i$ , we compute a random variable  $R_i$  from the key press, which assumes the value  $r_i = 1$  if the participant was fooled by the artificially generated stimulus, and  $r_i = 0$  otherwise. Thus,  $R_i$  is a Bernoulli distributed random variable. We assume the *confusion rate*  $p_i$  to be dependent on only the ideal observer parameters of the generated stimulus, such as number of basis-functions/MPs/IPs or model scores (see Section 3.1):

$$p(R_i = r_i) = p_i^{r_i} (1 - p_i)^{1-r_i} \quad (7)$$

We assume a conjugate p(oste)rior on the confusion rate  $p_i$ , i.e., a beta distribution, and compute error bars on  $p_i$  under this assumption. Please note that we decided to report the confusion rate as “success”-measure from the perspective of the model, which we want to evaluate, instead of reporting the discrimination ability of the participant  $1 - p$  that is frequently used in the psychophysics literature.

*Power Analysis.* We would like to determine if the confusion rate of an artificial stimulus with a natural stimulus is less than chance. More precisely, denote hypothesis  $H_0: p_i \in [0.45, 0.55]$  and  $H_1: p_i \notin [0.45, 0.55]$ . We choose the number of trials such that the falsehood of  $H_0$  is discovered with power 0.8 when  $H_1$  is true, i.e.,  $1 - P(H_0|H_1) = 0.8$ . This yields a number of  $N = 158$  trials for each parameter combination. Considering this number and our goal to test a wide range of parameter combinations (120 in total), the resulting number of trials is too large for a single participant. We therefore distribute the necessary trials across participants, excluding the possibility of inter-participant comparisons.

*Logistic Regression.* Each stimulus parameter combination is associated with scores  $S_i$  measuring the quality of the generated movement after training: the predictive mean squared error (MSE) for all models, ELBO for TMP, and v(c)GPDM models and dynamics- and pose-ELBO only for the v(c)GPDM models. We use logistic regression to find the relation between these model scores and the confusion rate:

$$p_i = \frac{c}{1 + \exp(w_0 + w_1 S_i)}, \quad (8)$$

where  $c \in [0, 0.5]$  reflects our assumption that the confusion rate can at best approach chance level. Assuming independence across  $N$  trials, we can compute the log-likelihood of all trials:

$$p(r_1, \dots, r_N | w_0, w_1) = \log \left( \prod_{i=1}^N p(r_i) \right) \quad (9)$$

$$= \sum_{i=1}^N r_i \log(p_i) + \sum_{i=1}^N (1 - r_i) \log(1 - p_i). \quad (10)$$

We now learn the weights  $(w_0^*, w_1^*)$  by maximizing the log-likelihood function using the `scipy.optimize.fmin_l_bfgs_b` routine (Jones et al. 2001). The gradients required for this optimizer are computed with `autograd` in Python 3.6.

*Cross-Validation.* We test the predictive capabilities of the different regressors  $S_i$  using  $n$ -fold cross-validation: the data set is split into  $n$  blocks, then weights are learned using  $n-1$  blocks, and the log-likelihood of the left-out block is computed. This procedure is repeated  $n$  times, and the average left-out log-likelihood is used as score.

*Logarithmic Likelihood-Ratio.* We compare the predictive power of the different regressors against the null hypothesis:  $p_i$  is independent of  $S_i$ . We can now compute the cross-validatory log(likelihood-ratio) to evaluate the evidence for the statement “Model score  $S_i$  is more predictive of perceived naturalness than the best constant  $p_i$ ”.

## 4 RESULTS

We present the following results: participant evaluation, estimation of interesting parameter regimes, and finally comparison of model scores regarding their predictive power.

### 4.1 Evaluation of Participants

*Attention Checks.* During all parts of the experiment, we presented participants with attention check trials, where different, clearly unnatural stimuli had to be detected. We measured the detection rate of these stimuli. There were 17 attention check trials in the first part of the experiment and 15/14 in the second part’s conditions. Over all trials, the detection rate was 98.0%. Three participants of the experiment had a detection rate of under 85%. These were excluded from further data analysis.

*Catch Trials.* During the first part of the experiment, we collected data from 162 catch-trials. 72 responses specified the previously unknown stimulus as more natural (44.4%). The probability that these responses are random, i.e. that they were generated by a Bernoulli process with  $p = 0.5$  vs.  $p \neq 0.5$  ( $p \sim \text{beta}(1, 1)$ ) is  $\approx 0.8$ . We are therefore fairly certain that the participants did *not* use memorization strategies for their response.

### 4.2 Estimating Regions of Interest in Parameter Space

We evaluated the perceived naturalness of 103 models using 976 stimuli during the first experiment (see Table 1). We collected 16902 trial responses from 27 participants in the first part of the experiment. Each participant completed 620 trials to estimate the confusion rate of models after exclusion of catch trials and attention checks. Across all trials, the confusion rate was 0.228. Please check the supplementary material to find a video with some example trials (with simulated random answers) to get an impression of the visual consequences for different models.

We used the results of this first part of the experiment to estimate more models of interest. For the TMP models, we decided after inspection of the confusion rate (Figure 2, left) to increase the number of MPs up to 15. Interestingly, the confusion rate seems to converge in the slightly hyper-realistic regime at  $p \approx 0.55$ . For the DMP models, we decided on testing numbers of basis function ranging from 50 to 100 (Figure 2, right). The confusion rate peaks at 80 basis functions. This does not coincide with the minimal predictive MSE, which is reached with 25 basis functions and increases from there on.

Table 1. Overview of Generated Trials for Each MP Model Type, Number of Attention Check Trials, and Number of Tested Parameter Combinations (After Excluding Attention Check Trials) in the First Part of the Experiment

MP model type	# Trials	# Att. checks	# Parameters combinations
vCGPDM	7,290	108	45
vGPDM	6,156	297	38
TMP	1,458	0	9
DMP	1,296	54	8
cGPDM (MAP)	270	0	1
GPDM (MAP)	270	0	1
Total	16,740	459	102

The confusion rates of the vGPDM models peak at (35, 10), (30, 20), (20, 20), (25, 35) (#IP Dynamics, #IPs pose) parameter combinations. These four parameter combinations are indistinguishable from natural stimuli (Figure 3, left). We estimated, by visual inspection, the location of the maximal confusion rate assuming that the confusion rate is described by a concave function of the parameters with additional noise. This yielded (25, 25) as the location of the global maximum.

The measured confusion rates of the vCGPDM models are equal at (20, 15) and (20, 20). We estimated (25, 20) to be a global maximum for the vCGPDM, in the same manner as for the vGPDM. Based on our power analysis and time budget, we decided on testing 67 parameter combinations for vGPDM and vCGPDM each. This way, we ended up testing 629 additional stimuli for the second part of the experiment (see Figure 4).

We also included GPDM and CGPDM models trained by MAP (maximum *a-posteriori*) instead of the ELBO. We measured confusion rates of  $0.000 \pm 0.004$  for the MAP-GPDM, and  $0.11 \pm 0.02$  for the MAP-CGPDM. These models were not tested again in the second part of the experiment. All resulting models are summarized in Table 2.

### 4.3 Predicting Perceived Naturalness

Using data from both experimental parts, we predicted the confusion rate from model scores via logistic regression. The results are shown in Figure 5 for TMP and DMP models and in Figure 6 for vGPDM and vCGPDM models. Depicted are the measured and predicted confusion rates for the tested models (columns), and different scores (rows). Furthermore, cross-validation results are summarized as log likelihood-ratio “ln K” of the prediction of the respective regressors versus the constant prediction (null-) hypothesis above each graph. Each “X” represents the confusion rate achieved by a unique parameter combination. The regression yields best results for the TMP models. MSE and ELBO of TMP models have similar predictive capabilities, as they are highly correlated in the investigated parameter regime. While the MSE also has predictive power for the v(C)GPDM models, the ELBO is not a suitable regressor. Inspection of the pose and dynamic terms of the ELBO reveals that this is due to the low score of the pose ELBO:  $\ln K \approx -0.7$ . The dynamic ELBO on the other hand even surpasses the MSE for the vCGPDM (Figure 6, left). Visual inspection of the logistic regression result for the DMP models shows that there is no simple sigmoidal relation between the perceptual validity and the DMPs MSE. This corresponds to the mismatch between MSE and confusion rate reported in Figure 2.

### 4.4 Comparing Best Models of Each MP-class

We plotted the confusion rate of all MP-models over the MSE in Figure 7. Even though a small MSE indicates better perceptual performance of the models, the relationship between MSE and confusion rate differs between the MP-model classes. For example, the vGPDM achieves high confusion rates even with high MSE.

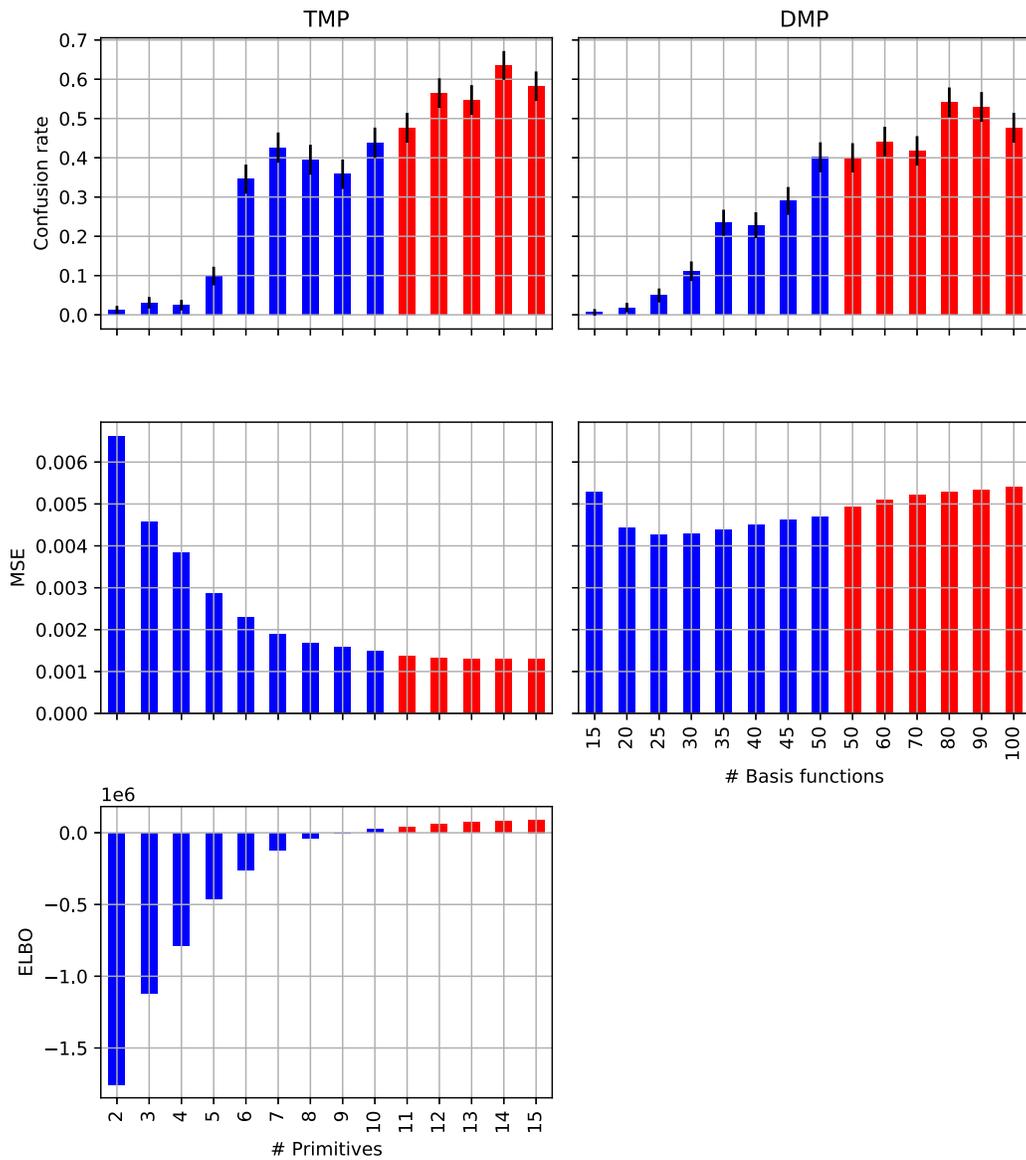


Fig. 2. Confusion rate, MSE, ELBO (from top to bottom) of TMP (left) and DMP (right) models for investigated model parameters. Data of first part of the experiment is colored blue, data of the second part is colored red.

For comparison of model performance we chose the best performing model of each MP-class, and computed the probabilities of all  $6! = 720$  many possible orderings of the models by confusion rate. We assumed  $\text{beta}(1,1)$  priors on the rate and a Bernoulli observation model, as before. The most probable ordering is  $\text{TMP} > \text{vGPDM} > \text{DMP} > \text{vCGPDM} > \text{CGPDM}(\text{MAP}) > \text{GPDM}(\text{MAP})$  with a probability of 0.36. We computed marginal confusion rates and marginal pairwise ordering probabilities, see Figure 8. TMP, vGPDM, and DMP are comparable, while all other models are clearly worse. We used the same statistical model to test if the TMP’s confusion rate is above 0.5, i.e., whether human participants perceive the model-generated stimulus as more natural than the natural one. Given our data, we are  $\approx 0.99$  sure of that.

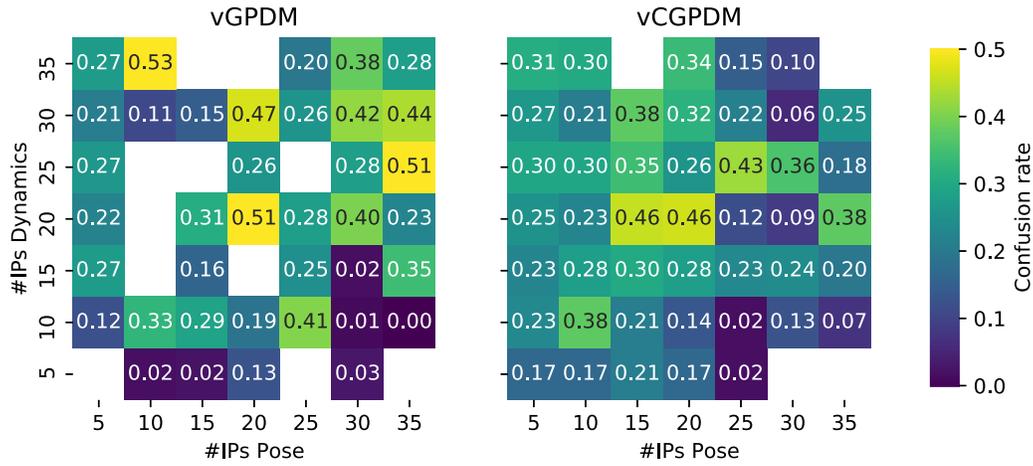


Fig. 3. Confusion rate of v(c)GPDM models in first part of experiment: Number of inducing points for the pose mapping on the x-axis, and for the dynamics mapping on the y-axis. The attention check parameter combinations are indicated by the white squares, where the model training procedure converged to obviously unnatural movements. Numbers on tiles are the measured confusion rates.

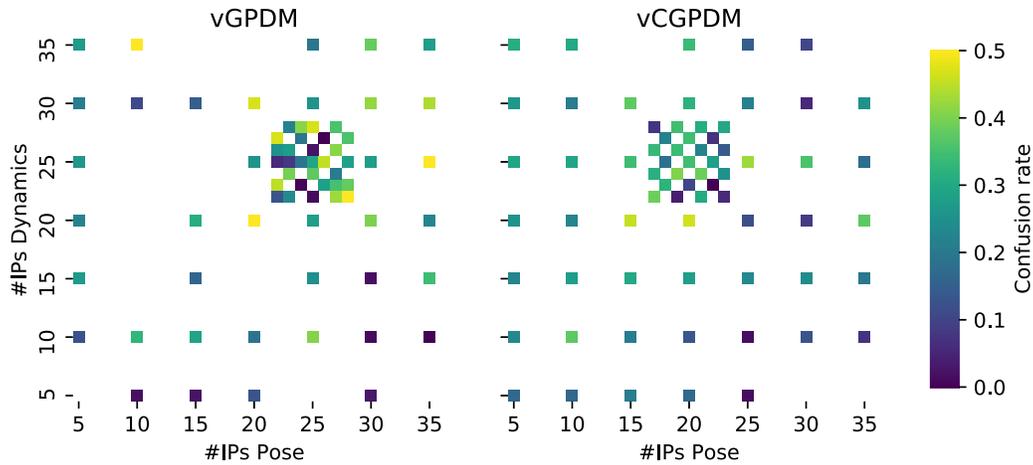


Fig. 4. Confusion rate of v(c)GPDM of first and second part of the experiment: Data of second part of the experiment are clustered around (25, 25) for vGPDM and (25, 20) for vCGPDM. Confusion rates are indicated by the same color-map as in Figure 3.

## 5 DISCUSSION

The tested MP models incorporate different (perceptual) predictive mechanisms: While TMPs determine the complete time course, the dynamical models make predictions for each next time-point from previous ones. The dynamical models therefore have advantages in feedback control applications where perturbations must be expected. TMPs, on the other hand, make perceptual predictions, as well as planning, easy, as there is no roll-out necessary to access the end-state of a movement.

The perceptually most valid, even hyper-realistic model is the variationally trained TMP. The shared representation between perception and production may therefore be more abstract: one dynamics model paired with

Table 2. Overview of Generated Trials for Each MP Model Type, Number of Attention Check Trials, and Number of Tested Parameter Combinations in the Second Part of the Experiment

MP model type	# Trials	# Att. checks	# Parameters combinations
vCGPDM	4,233	17	25
vGPDM	4,097	476	31
TMP	850	0	5
DMP	1,020	0	6
Total	10,200	493	67

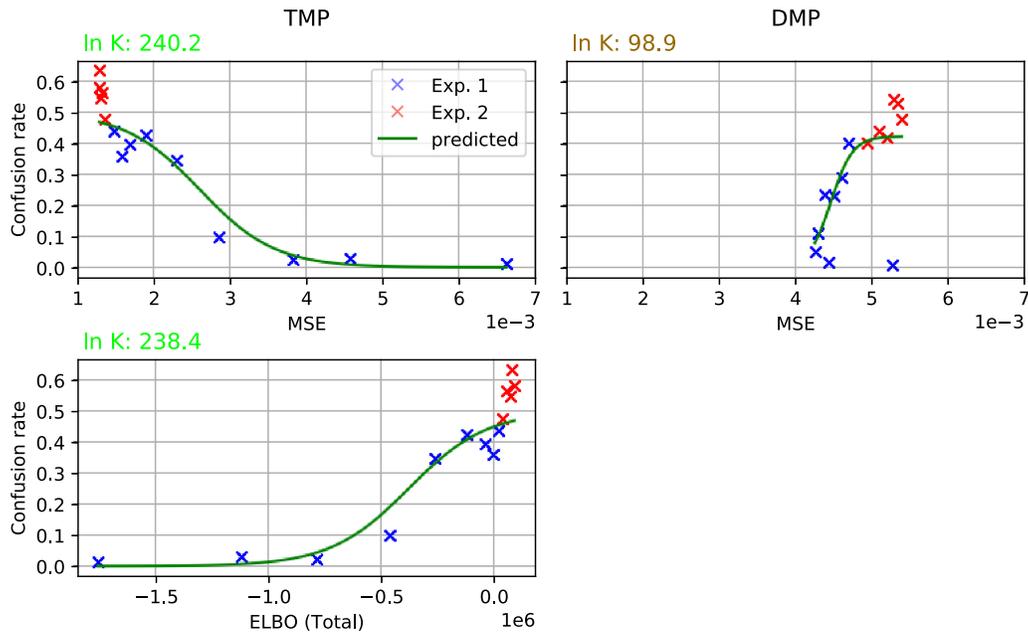


Fig. 5. Confusion rate of TMP (left) and DMP (right) models as function of model scores: MSE (top) and ELBO (bottom). Blue and red "X"s show confusion rates for model-parameters measured during experiment one and two. Green lines are predictions of the confusion rate (perceived naturalness) from the logistic regression using the regressor corresponding the abscissa label. Results of the cross-validation are summarized as log likelihood-ratio  $\ln K$  in the top left corner of each plot, with the text color visualizing low (red) to strong (green) evidence in favour of the regressor being a good predictor of naturalness perception. See 3.3 for more detail.

a corresponding TMP model that encodes typical (unperturbed) solutions of the dynamics model, for fast perceptual predictions (Giese and Poggio 2000). Currently, we are preparing an experiment to compare TMP and dynamical MP models regarding their specific predictive mechanism employed in movement perception.

The vGPDM is still comparable to the TMP and the DMP, but that might change with more data. All other models are clearly worse. However, we are almost certain that the variationally approximated models are better than their MAP counterparts, which highlights the advantages of sparse variational posterior parametrizations.

We showed that approximate Bayesian model scores (ELBO, held-out MSE) can be used to predict the perceived naturalness of human animations. Assuming that humans are experts (i.e., nearly ideal observers) at perceiving their conspecifics' movements from noisy sensory input, it follows that their movement recognition performance

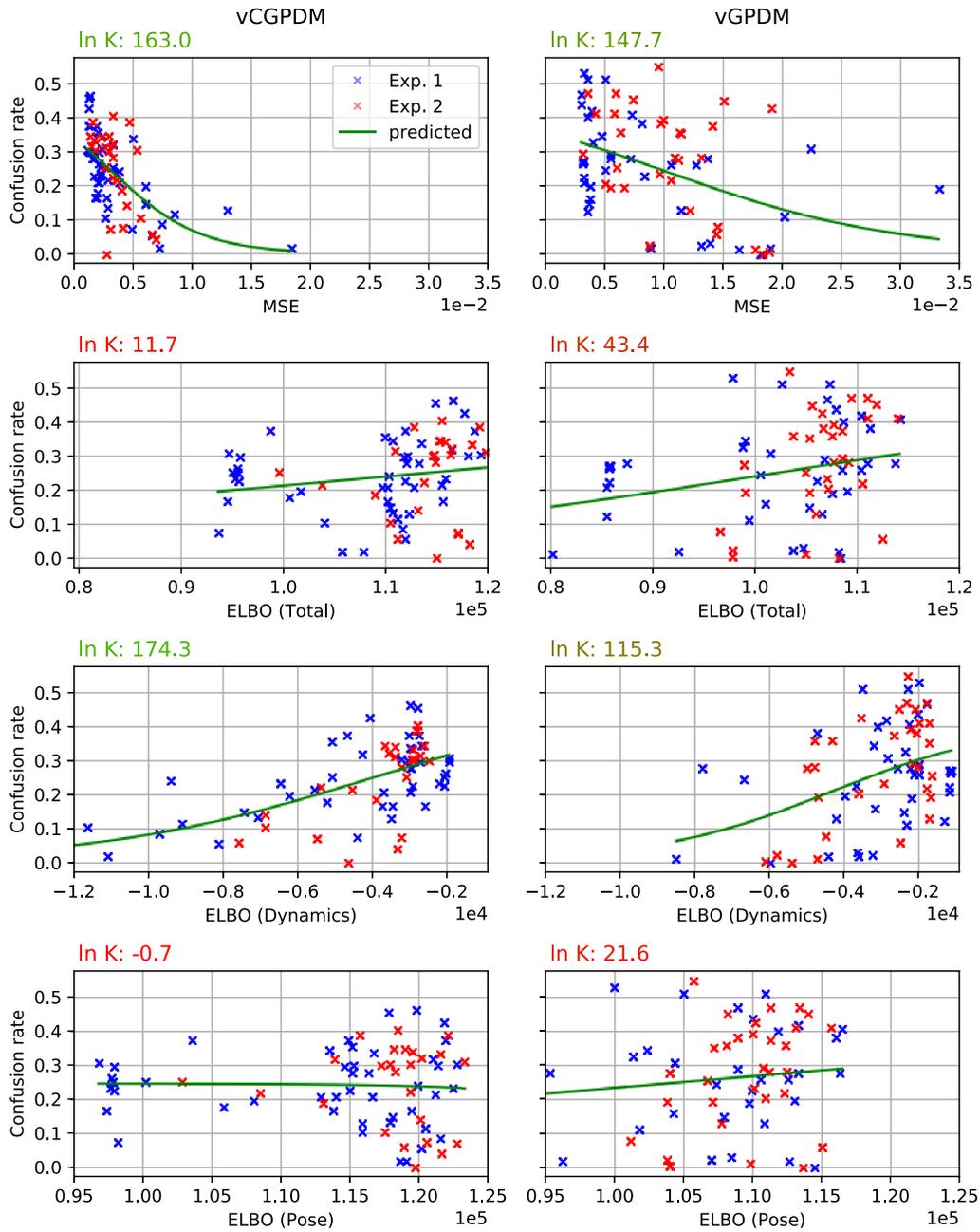


Fig. 6. Confusion rate of vCGPDM (left) and vGPDM (right) models as function of model scores: MSE, Total-, Dynamics-, Pose-ELBO (from top to bottom). Symbols have the same meaning as in Figure 5. See 3.3 for more detail.

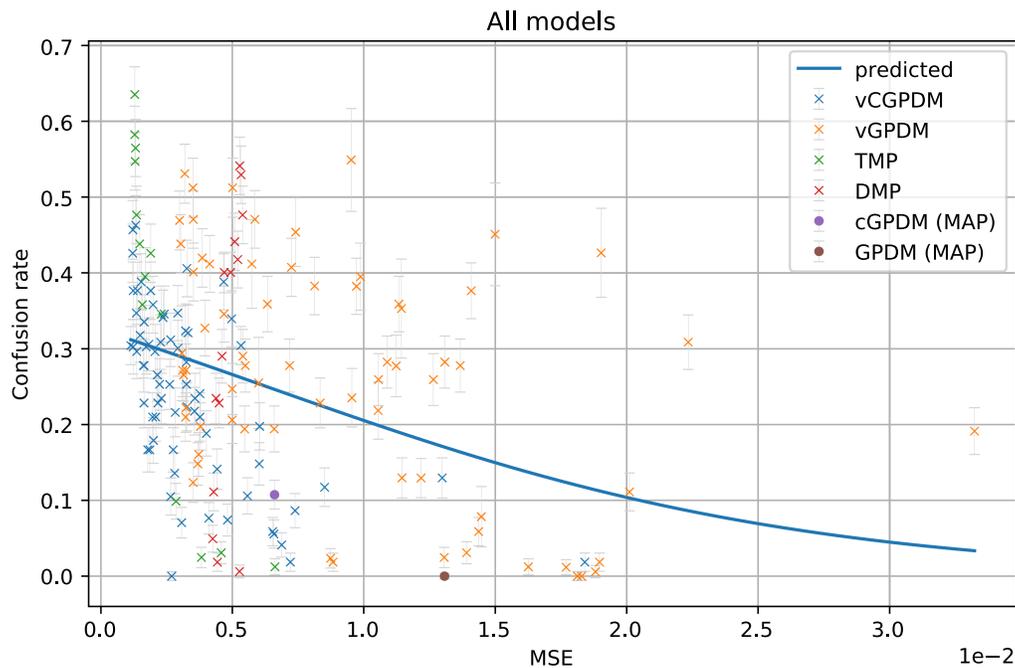


Fig. 7. Confusion rate of all models vs. test MSE and prediction learned over all models. Same data as in the first rows of Figures 5 and 6 plus cGPDM (MAP) and GPDM(MAP). Error bars denote beta standard deviation of the confusion rate.

should be near-Bayesian in general. Therefore, in particular, the perceived naturalness of a movement is expected to be predictable by approximate Bayesian model scores of the MPs. Our confirmation of this prediction adds evidence to the claim that human perception is nearly Bayes-optimal in many instances.

Comparison of total, dynamics, and pose ELBO as predictor for perceived naturalness of the v(C)GPDM models yields an interesting result: total ELBO is not a good predictor, because terms related to the latent-to-observed (pose) mapping apparently have no relevance for the perception of human animations. In contrast, dynamics ELBO scores indicate that a faithful dynamical mapping is more important than the pose mapping.

These computational level predictions might therefore also provide some insight into the perception of human animations on a algorithmic/mechanistic level: A feed-forward neural model (Giese and Poggio 2003) has been proposed arguing for the existence of separate motion and form pathways, where the motion pathway is performing a form of sequence recognition. Our results can thus be interpreted as additional evidence for importance of dynamics for perceiving human animations. Similar results have been derived from classical examinations of point light walkers (for a review, see Giese 2014): While local motion features form the simpler explanation for the perception of point light stimuli as biological motion than form features (Casile and Giese 2005), it has also been shown that biological motion perception can be induced in absence of local motion features (Beintema and Lappe 2002). For discrimination tasks, the information contained in the dynamics of the movement is more important than posture (Troje 2002).

Even though DMP models can generate highly realistic movement, a disadvantage is the unclear relation between MSE and perceptual validity. This finding demonstrates that the predictive MSE is not a sufficient indicator for perceptual performance: it is highly implausible that naturalness of a movement is evaluated by computing its point-wise deviation from an internal prototype for this movement.

The vGPDM performs comparable to the DMP, whereas the additional modular flexibility of the vCGPDM does not seem to be needed for our dataset: its best confusion rate is probably (86%) lower than that of the

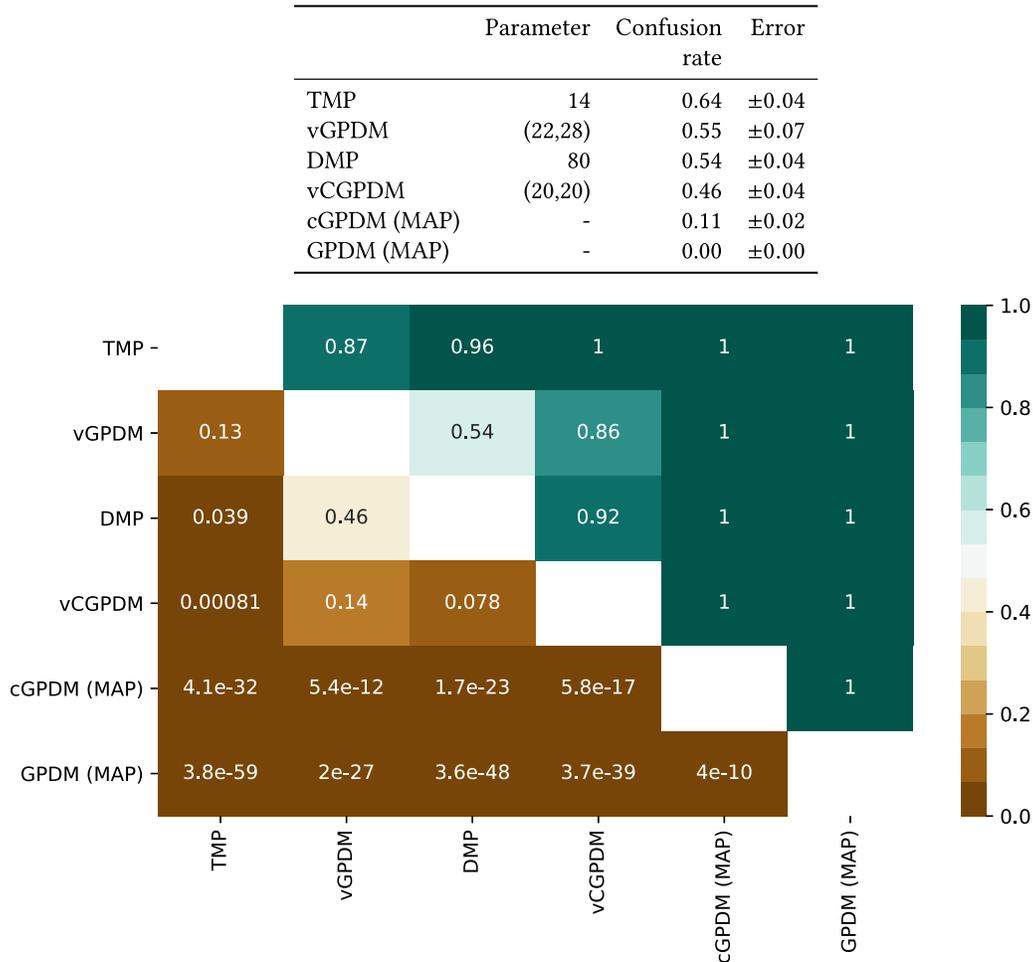


Fig. 8. Comparison of best models: (Top) Table of best models with corresponding parameter(-combinations), confusion rate and standard errors of beta posteriors. (Bottom) Bayesian ordering tests: probabilities that the best parameter combination of the models in the rows yields a higher confusion rate than the models in the columns. For example, the best TMP model (row) achieves a higher confusion rate than the best vGPDM model (column) with 87% certainty given our data.

vGPDM. This might also be due to the stochasticity in the training procedure: reachable optima depend on the random initial values of the optimization. Thus, the determined number of IPs where we suspected the perceptual optimum did not yield reliably high confusion rates or model scores for the second part of our experiment.

In our study, we only validated perceived naturalness of walking movements. We chose walking movements, because they are comparatively easy to model, yet highly important especially for animators. We are currently extending our investigation towards other, more complex movements, such as handling objects. Our hypothesis is that the main result—the Bayesian model score predicts naturalness perception—will generalize to these different movements as well, because at no point did we rely on features specific to walking.<sup>2</sup>

<sup>2</sup>The only exception is the specification of the DMP’s attractor model, which is not important for our main results.

In our experimental paradigm we chose simultaneous side-by-side presentation of generated and natural movement videos. Simultaneous presentation has two advantages: At any point in time there is a base-line for the participants. Presenting one after another would double the time of an already lengthy experiment. Still, the presentation time is short, thus the participants had to distribute their fixations across the two simultaneously presented videos. We will test and consider alternative paradigms, e.g., let participants rate naturalness on a scale. The gain of information per trial might be great enough to sacrifice the indistinguishability criterion. This might also enable inter-participant analysis, which is not possible in our paradigm, as described in 3.3 (Power Analysis).

## 6 CONCLUSIONS

Our study shows that MP models are capable of producing perceptually valid movements and we demonstrated that the prediction of naturalness is possible from model scores. These results add evidence for a shared MP-representation of action and perception and indicates the possibility of cheap, automated, and perceptually valid model selection for applications, e.g., in virtual reality. Finding a shared representation of MPs for perception and action could also provide a tool to study imitation learning in robots (Schaal 1999).

Congruent with previous studies, we found that parameters connected to dynamics are more relevant for perception than those connected with pose. This result could be useful to further improve generative models like the vCGPDM, and highlights the importance of prediction in the perception of human animations. While the Graphics Turing Test is a suitable tool for the estimation of perceived naturalness of movement, an analysis fixation data could shed some light on the features that drive this perception. Also, it would be interesting to determine what causes the hyper-realism of the TMP model.

Given that temporal and dynamical MP models have different advantages in movement planning and production, one of our current research directions is integrating such models into sensorimotor primitives, which are joint models of movement production and perception, with the aim of a computationally feasible instantiation of the common coding hypothesis. Sensory prediction during movement might not only be reflected in the movement itself, but also retrieved by an observer of biological movement, e.g., mime art. Applying such sensorimotor primitives to computer animation would enable a much more flexible interaction with avatars in virtual reality: Perceptually valid primitives could incorporate environmental constraints as well as the VR users movements, and be composed to form complex responsive behaviour of the avatar.

## ACKNOWLEDGMENTS

We thank Olaf Haag for help with rendering of the stimuli and collecting data.

## REFERENCES

- Jaap Beintema and Markus Lappe. 2002. Perception of biological motion without local image motion. *Proceedings of the National Academy of Sciences* 99, 8 (April 2002), 5661–5663. DOI: <https://doi.org/10.1073/pnas.082483699>
- Nikolai Bernstein. 1967. *The Co-ordination and Regulation of Movements*. Pergamon-Press. <https://books.google.de/books?id=kX5OQAAlAAJ>
- Bennett Bertenthal and Jeannine Pinto. 1994. Global processing of biological motions. *Psychological Science* 5, 4 (1994), 221–225. DOI: <https://doi.org/10.1111/j.1467-9280.1994.tb00504.x>
- Antonino Casile and Martin A. Giese. 2005. Critical features for the recognition of biological motion. *Journal of Vision* 5, 4 (April 2005), 6–6. DOI: <https://doi.org/10.1167/5.4.6>
- Enrico Chiovetto, Cristóbal Curio, Dominik Endres, and Martin A. Giese. 2018. Perceptual integration of kinematic components in the recognition of emotional facial expressions. *Journal of Vision* 18, 4 (April 2018), 13–13. DOI: <https://doi.org/10.1167/18.4.13>
- Debora Clever, Monika Harant, Henning Koch, Katja Mombaur, and Dominik Endres. 2016. A novel approach for the generation of complex humanoid walking sequences based on a combination of optimal control and learning of movement primitives. *Robotics and Autonomous Systems* 83 (Sept. 2016), 287–298. DOI: <https://doi.org/10.1016/j.robot.2016.06.001>
- Debora Clever, Monika Harant, Katja Mombaur, Maximilien Naveau, Olivier Stasse, and Dominik Endres. 2017. COCoMoPL: A novel approach for humanoid walking generation combining optimal control, movement primitives and learning and its transfer to the real robot HRP-2. *IEEE Robotics and Automation Letters* 2, 2 (2017), 977–984. DOI: <https://doi.org/10.1109/LRA.2017.2657000>

- Andrea d'Avella, Philippe Saltiel, and Emilio Bizzi. 2003. Combinations of muscle synergies in the construction of a natural motor behavior. *Nature Neuroscience* 6, 3 (March 2003), 300–308. DOI : <https://doi.org/10.1038/nn1010>
- Eran Dayan, Antonino Casile, Nava Levit-Binnun, Martin A. Giese, Talma Hendler, and Tamar Flash. 2007. Neural representations of kinematic laws of motion: Evidence for action-perception coupling. *Proceedings of the National Academy of Sciences* 104, 51 (Dec. 2007), 20582–20587. DOI : <https://doi.org/10.1073/pnas.0710033104>
- Dominik Endres, Enrico Chiovetto, and Martin A. Giese. 2013. Model selection for the extraction of movement primitives. *Frontiers in Computational Neuroscience* 7 (2013), 185. DOI : <https://doi.org/10.3389/fncom.2013.00185>
- Dominik Endres, Andrea Christensen, Lars Omlor, and Martin A. Giese. 2011. Emulating human observers with Bayesian binning: Segmentation of action streams. *ACM Transactions on Applied Perception (TAP)* 8, 3 (2011), 16:1–12.
- Karl Friston. 2010. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience* 11, 2 (February 2010), 127–138. DOI : <https://doi.org/10.1038/nrn2787>
- Martin A. Giese. 2014. Biological and body motion perception. *The Oxford Handbook of Perceptual Organization*. DOI : <https://doi.org/10.1093/oxfordhb/9780199686858.013.008>
- Martin A. Giese and Tomaso Poggio. 2000. Morphable models for the analysis and synthesis of complex motion patterns. *International Journal of Computer Vision* 38 (June 2000), 59–73. DOI : <https://doi.org/10.1023/A:1008118801668>
- Martin A. Giese and Tomaso Poggio. 2003. Neural mechanisms for the recognition of biological movements: Cognitive neuroscience. *Nature Reviews Neuroscience* 4, 3 (March 2003), 179–192. DOI : <https://doi.org/10.1038/nrn1057>
- Simon Giszter. 2015. Motor primitives—New data and future questions. *Current Opinion in Neurobiology* 33 (Aug. 2015), 156–165. DOI : <https://doi.org/10.1016/j.conb.2015.04.004>
- Simon Giszter, Emilio Bizzi, and Ferdinando A. Mussa-Ivaldi. 1992. Motor organization in the frog's spinal cord. In *Analysis and Modeling of Neural Systems*, Frank H. Eeckman (Ed.). Springer US, Boston, MA, 377–392. DOI : [https://doi.org/10.1007/978-1-4615-4010-6\\_38](https://doi.org/10.1007/978-1-4615-4010-6_38)
- Jessica K. Hodgins, James F. O'Brien, and Jack Tumblin. 1998. Perception of human motion with different geometric models. 4, 4 (1998), 307–316. DOI : <https://doi.org/10.1109/2945.765325>
- Bernhard Hommel, Jochen Müsseler, Gisa Aschersleben, and Wolfgang Prinz. 2001. The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences* 24 (2001), 849–937.
- Auke Jan Ijspeert, Jun Nakanishi, Heiko Hoffmann, Peter Pastor, and Stefan Schaal. 2013. Dynamical movement primitives: Learning attractor models for motor behaviors. *Neural Computation* 25, 2 (Feb. 2013), 328–373. DOI : [https://doi.org/10.1162/NECO\\_a\\_00393](https://doi.org/10.1162/NECO_a_00393)
- Yuri P. Ivanenko, Richard E. Poppele, and Francesco Lacquaniti. 2004. Five basic muscle activation patterns account for muscle activity during human locomotion: Basic muscle activation patterns. *The Journal of Physiology* 556, 1 (April 2004), 267–282. DOI : <https://doi.org/10.1113/jphysiol.2003.057174>
- Gunnar Johansson. 1994. Visual perception of biological motion and a model for its analysis. *Perceiving Events and Objects* 14 (1994), 185–207.
- Eric Jones, Travis Oliphant, and Pearu Peterson. 2001. SciPy: Open source scientific tools for Python. [Online; accessed 2015-10-09].
- David C. Knill and Alexandre Pouget. 2004. The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neuroscience* 27 (2004).
- Michael D. McGuigan. 2006. Graphics turing test. *CoRR abs/cs/0603132* (2006).
- Lars Omlor and Martin A. Giese. 2011. Anechoic blind source separation using Wigner marginals. *Journal of Machine Learning Research* 12 (2011), 1111–1148.
- Jonathan W. Peirce. 2009. Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics* 2 (2009). DOI : <https://doi.org/10.3389/neuro.11.010.2008>
- Felix Polyakov, Eran Stark, Rotem Drori, Moshe Abeles, and Tamar Flash. 2009. Parabolic movement primitives and cortical states: Merging optimality with geometric invariance. *Biological Cybernetics* 100, 2 (2009), 159.
- Wolfgang Prinz. 1997. Perception and action planning. *European Journal of Cognitive Psychology* 9, 2 (June 1997), 129–154. DOI : <https://doi.org/10.1080/713752551>
- Claire L. Roether, Lars Omlor, Andrea Christensen, and Martin A. Giese. 2009. Critical features for the perception of emotion from gait. *Journal of Vision* 9, 6 (June 2009), 15–15. DOI : <https://doi.org/10.1167/9.6.15>
- Stefan Schaal. 1999. Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences* 3, 6 (June 1999), 233–242. DOI : [https://doi.org/10.1016/S1364-6613\(99\)01327-3](https://doi.org/10.1016/S1364-6613(99)01327-3)
- Stefan Schaal. 2006. Dynamic movement primitives—a framework for motor control in humans and humanoid robotics. In *Adaptive Motion of Animals and Machines*, Hiroshi Kimura, Kazuo Tsuchiya, Akio Ishiguro, and Hartmut Witte (Eds.). Springer-Verlag, Tokyo, 261–280. DOI : [https://doi.org/10.1007/4-431-31381-8\\_23](https://doi.org/10.1007/4-431-31381-8_23)
- Krishna Shenoy, Maneesh Sahani, and Mark M. Churchland. 2013. Cortical control of arm movements: A dynamical systems perspective. 36, 1 (2013), 337–359. DOI : <https://doi.org/10.1146/annurev-neuro-062111-150509>
- Yun Kyoung Shin, Robert W. Proctor, and E. John Capaldi. 2010. A review of contemporary ideomotor theory. *Psychological Bulletin* 136, 6 (Nov. 2010), 943–974. DOI : <https://doi.org/10.1037/a0020541>
- David Sussillo, Mark M. Churchland, Matthew T. Kaufman, and Krishna V. Shenoy. 2015. A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience* 18, 7 (2015), 1025.

- Nick Taubert, Andrea Christensen, Dominik Endres, and Martin A. Giese. 2012. Online simulation of emotional interactive behaviors with hierarchical gaussian process dynamical models. *Proceedings of the ACM Symposium on Applied Perception (ACM-SAP 2012)* (2012), 25–32. DOI: <https://doi.org/10.1145/2338676.2338682>
- Emanuel Todorov and Michael I. Jordan. 2003. A minimal intervention principle for coordinated movement. In *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer (Eds.). MIT Press, 27–34. <http://papers.nips.cc/paper/2195-a-minimal-intervention-principle-for-coordinated-movement.pdf>.
- Matthew Tresch, Philippe Saltiel, and Emilio Bizzi. 1999. The construction of movement by the spinal cord. *Nature Neuroscience* 2, 2 (Feb. 1999), 162–167. DOI: <https://doi.org/10.1038/5721>
- Nikolaus Troje. 2013. What is biological motion? Definition, stimuli, and paradigms. *Social Perception: Detection and Interpretation of Animacy, Agency, and Intention*. 13–36. DOI: <https://doi.org/10.7551/mitpress/9780262019279.003.0002>
- Nikolaus F. Troje. 2002. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision* 2, 5 (Sept. 2002), 2–2. DOI: <https://doi.org/10.1167/2.5.2>
- Nikolaus F. Troje, Cord Westhoff, and Mikhail Lavrov. 2005. Person identification from biological motion: Effects of structural and kinematic cues. 67, 4 (2005), 667–675. DOI: <https://doi.org/10.3758/BF03193523>
- Dmytro Velychko and Dominik Endres. 2017. A method and algorithm for estimation of pose and skeleton in motion recording systems with active markers (pending patent).
- Dmytro Velychko, Dominik Endres, Nick Taubert, and Martin A. Giese. 2014. Coupling gaussian process dynamical models with product-of-experts kernels. In *Proceedings of the 24th International Conference on Artificial Neural Networks, Lecture Notes in Computer Science*, Vol. 8681. Springer, 603–610.
- Dmytro Velychko, Benjamin Knopp, and Dominik Endres. 2018. Making the coupled Gaussian process dynamical model modular and scalable with variational approximations. *Entropy* 20, 10 (Sept. 2018), 724. DOI: <https://doi.org/10.3390/e20100724>
- Jack Meng-Chieh Wang, David J. Fleet, and Aaron Hertzmann. 2008. Gaussian process dynamical models for human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 2 (Feb. 2008), 283–298. DOI: <https://doi.org/10.1109/TPAMI.2007.1167>
- Daniel M. Wolpert, Kenji Doya, and Mitsuo Kawato. 2003. A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences* 358, 1431 (March 2003), 593–602. DOI: <https://doi.org/10.1098/rstb.2002.1238>

Received July 2019; accepted August 2019

# E

Evaluating Perceptual Predictions based  
on Movement Primitive Models in VR-  
and Online-Experiments

# Evaluating Perceptual Predictions based on Movement Primitive Models in VR- and Online-Experiments

Benjamin Knopp  
Department of Psychology  
University of Marburg  
benjamin.knopp@uni-marburg.de

Dmytro Velychko  
Department of Psychology  
University of Marburg  
dmytro.velychko@uni-marburg.de

Johannes Dreibrodt  
Department of Psychology  
University of Marburg  
dreibrodt@students.uni-marburg.de

Alexander C. Schütz  
Department of Psychology  
University of Marburg  
alexander.schuetz@staff.uni-marburg.de

Dominik Endres  
Department of Psychology  
University of Marburg  
dominik.endres@uni-marburg.de

## ABSTRACT

We investigate the role of prediction in biological movement perception by comparing different representations of human movement in a virtual reality (VR) and online experiment. Predicting movement enables quick and appropriate action by both humans and artificial agents in many situations, e.g. when the interception of objects is important. We use different predictive movement primitive (MP) models to probe the visual system for the employed prediction mechanism. We hypothesize that MP-models, originally devised to address the degrees-of-freedom (DOF) problem in motor production, might be used for perception as well.

In our study we consider object passing movements. Our paradigm is a predictive task, where participants need to discriminate movement continuations generated by MP models from the ground truth of the natural continuation. This experiment was conducted first in VR, and later on continued as online experiment. We found that results transfer from the controlled and immersive VR setting with movements rendered as realistic avatars to a simple and COVID-19 safe online setting with movements rendered as stick figures. In the online setting we further investigate the effect of different occlusion timings. We found that contact events during the movement might provide segmentation points that render the lead-in movement independent of the continuation and thereby make perceptual predictions much harder for subjects. We compare different MP-models by their capability to produce perceptually believable movement continuations and their usefulness to predict this perceptual naturalness.

Our research might provide useful insight for application in computer animation, by showing how movements can be continued without violating the expectation of the user. Our results also contribute towards an efficient method of animating avatars by combining simple movements into complex movement sequences.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SAP '20, September 12–13, 2020, Virtual Event, USA

© 2020 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-7618-1/20/07.

<https://doi.org/10.1145/3385955.3407940>

## CCS CONCEPTS

• **Computing methodologies** → **Perception**; *Animation*; *Motion processing*; • **Theory of computation** → *Gaussian processes*.

## KEYWORDS

human animation, movement primitives, perception, dynamical systems, psychophysics, Gaussian process dynamical model, dynamical movement primitives

## ACM Reference Format:

Benjamin Knopp, Dmytro Velychko, Johannes Dreibrodt, Alexander C. Schütz, and Dominik Endres. 2020. Evaluating Perceptual Predictions based on Movement Primitive Models in VR- and Online-Experiments. In *ACM Symposium on Applied Perception 2020 (SAP '20)*, September 12–13, 2020, Virtual Event, USA. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3385955.3407940>

## 1 INTRODUCTION

Predictive coding is one of the hot topics in neuroscience [Friston 2010; Hohwy 2013]. In this framework, the brain is viewed as an engine generating predictions based on previous sensory input. These predictions are then compared to the current sensory input to refine a percept. The investigation of the prediction mechanism is directly relevant for areas of applied perception, such as computer animation: Generating realistic animation could be achieved in the most economical manner possible [Sattler et al. 2005].

Ways of economical movement production have also been proposed to facilitate the motor control problem: movement primitives (MPs) are hypothetical elements used by the central nervous system to build complex movements. Assuming a common code of action and perception [Friston 2010; Prinz 1997], MPs might be used in perception as well. If this would be the case, the MP representation used by the brain should yield the best animation results. Furthermore, we hypothesize that movement perception is Bayes-optimal [Knill and Pouget 2004], i.e. we assume that the complexity of the perceptual representation reflects Bayesian model comparison, which serves as our ideal observer model with MP-Type specific complexity parameters as input (see 3.1). The cross-validatory mean squared error (MSE) as approximate Bayesian model evidence can

then be used to predict the perceived naturalness of movements based on MPs (see 3.3.2).

We use a prediction task (adapted from Graf et al. [2007]) to compare MP representations with different predictive mechanisms. Participants rate movement continuations generated by MP models in a two-alternative forced choice task. One trial consists of two sequences, each with the same lead-in movement followed by a short occlusion, but with one sequence showing the generated movement continuation and the other one showing the actual recorded movement. We implemented this paradigm in VR, as well as a web browser based online experiment.

The movements we study either contained object contact or not. We furthermore manipulate the occlusion timing to control the visibility of the contact event. We can therefore investigate the role of segmentation by a contact event on the perceptual prediction [Zacks and Swallow 2007]: we hypothesize that a contact event breaks the continuity of movement necessary for prediction. A contact event during occlusion should thus widen the expectation of possible continuations.

## 2 RELATED WORK

This study is inspired by [Knopp et al. 2019], and shares the same assumptions about MPs as possible common representation for action and perception. While this previous work focused on the perception of naturalness of movements, the current study addresses the predictive mechanism inherent in MP models. Similar studies were also conducted for MP models of emotional handshakes [Taubert et al. 2012] and facial expressions [Chiovetto et al. 2018].

In our experiments, we investigate the perceptual extrapolation of a trajectory beyond the actual presented or implied movement of an object, which is termed representational momentum (RM), as a part of the visual prediction process [Bertamini 1993; Freyd and Finke 1984; Thornton and Hayes 2004]. Senior et al. [2000] reviewed functional magnetic resonance imaging (fMRI) results and used transcranial magnetic stimulation (TMS) to identify the middle temporal visual area (V5/MT) as involved in processing RM. Jarraya et al. [2005] found evidence of RM in memory tasks involving movements represented in point-dot figures. Brain areas that process motion, such as V5/MT, respond when motion is implied, for example in pictures, or occluded [Graf et al. 2007]. Kilner et al. [2004] found neural oscillations in the motor cortex without actual motor activity during expectation of a hand movement presentation prior to its onset, presumably due to visual prediction processes. These processes are also found in participants observing imitable actions [Buccino et al. 2004; Wilson and Knoblich 2005]. These studies suggest motor activity, or motor simulation [Stadler et al. 2012] to be involved in predicting future percepts of movements in real time, which further supports the functional framework of the mirror neuron system [MNS, Iacoboni and Dapretto 2006; Rizzolatti and Craighero 2004].

Besides the involvement of the MNS in RM, Graf et al. [2007] also show that visual movement prediction is a real-time process that includes effect estimations of motor commands before the motor action is performed. Visual Movement Prediction also requires prior information [Schröger et al. 2015], such as visual identifications of the percepts, therefore making tasks of visual prediction more

difficult compared to sheer tasks of identifying or distinguishing movements, such as in Knopp et al. [2019]. This is consistent with the predictive coding framework, which follows from a Bayesian view of the MNS and also explains how we can infer movement intentions from movement observations [Kilner et al. 2004]. Bayesian model scores would therefore not only serve to identify the model with the best prediction performance, but should also be diagnostic of visual movement prediction performance of humans.

## 3 MODELS AND METHODS

In this section we shortly review relevant features of the investigated MP model types to make this publication self-contained. We then describe the experimental paradigm and its implementation as VR- and web-browser based online experiment. Finally we describe our methods for data analysis.

### 3.1 Movement Primitives

MPs refer to building blocks of complex movements, but there is little consensus on an exact definition. Consequently, many different types of MPs have been proposed in literature [Endres et al. 2013]. We focus on dynamical and temporal MPs in this study, as we are interested in finding a higher level representation suitable for modeling perception,

We perceptually validate 3 generative MP Types: Temporal MPs, Dynamical MPs and the coupled Gaussian Process Dynamical Model. Each MP-Type has specific complexity parameters, which should ideally be selected to maximize the Bayesian model evidence. We use the cross-validatory MSE as approximation to the model evidence.

In this section we can only provide a rough overview, just enough to enable readers from different backgrounds to understand parameters of the stimuli for the psychophysical experiment. Please refer to the cited papers for detailed information. Velychko et al. [2018] also provide graphical model representations and summarize the features of the MP models presented in this chapter.

**3.1.1 Temporal Movement Primitives [TMP, Clever et al. 2016].** Temporal MPs describe the stereotyped temporal patterns of movement parameters, for example Electromyography (EMG) signals, but also joint trajectories as well as endpoint trajectories. We refer to all signals more generally as Degree-of-freedom (DOF). A possible biological implementation of temporal MPs might be central pattern generators (CPGs) [Ivanenko et al. 2004] combined with cortical top-down control. Temporal MPs incorporate a temporal predictive mechanism: the complete time-course of the movement is determined at its onset. This type of MPs allows for simple concatenation and temporal scaling.

The trajectory  $x_k(t)$  of a DOF, e.g. a joint angle, is a weighted sum of  $Q$  MPs  $y_q(t)$ , which are functions of time.  $\varepsilon_i(t) \sim \mathcal{N}(0, \sigma_i)$  is Gaussian observation noise:

$$x_k(t) = \sum_{q=1}^Q w_{k,q} y_q(t) + \varepsilon_i(t) \quad (1)$$

The posterior distribution of weights and MPs are learned by approximate Bayesian learning via free energy. We use the number of

MPs  $Q = 3 \dots 15$  as ideal observer parameter. In general, more MPs allow for more fine-grained temporal structure of the movement, but might lead to over-fitting.

**3.1.2 Dynamic Movement Primitives [DMP, Ijspeert et al. 2013].** While temporal MPs directly model the movement parameters, DMPs describe the stereotyped elements of movement as attractors of a dynamical system, thus enabling the prediction of the next state from the previous ones. Building on the hypothesis of separate brain areas for rhythmic and discrete movements, two kinds of dynamical systems are common: cyclic oscillators and point attractors [Schaal 2006].

More formally: DMP models represent a movement trajectory  $x_k(t)$  obeying a differential equation. They rely on a damped spring system which forces  $x_k(t)$  to contract to the specified goal  $g_k$ , if the dampening factor is high enough. Through the non-linear forcing function  $f_k$  (Eq. 2) the trajectories can be modified. This function is modeled as weighted sum of Gaussian basis functions  $\Psi_i(\tau)$  (Eq. 4). Time is replaced by  $\tau$ , which decays exponentially to zero (Eq. 3). DMPs are learned from training data by setting the weights  $w_i$  such that the training mean-squared error between predicted and actual movement (MSE) is minimal.

$$\tau \ddot{x}_k = \alpha_z (\beta_z (g_k - x_k) - \dot{x}_k) + f_k(\tau) \quad (2)$$

$$\dot{\tau} \propto -\tau \quad (3)$$

$$f_k(\tau) = \frac{\sum_{i=1}^N \Psi_i(\tau) w_{k,i}}{\sum_{i=1}^N \Psi_i(\tau)} \tau (g_k - x_k(0)) \quad (4)$$

We investigate  $N = 10, 20 \dots, 100$  basis functions as ideal observer parameters. Basis functions serve a similar role as the number of MPs in the TMP model: more basis functions allow for more complicated forcing functions, which enable richer temporal dynamics.

**3.1.3 Coupled Gaussian Process Dynamical Model [CGPDM, Velychko et al. 2014].** CGPDMs compose different dynamical models in a low dimensional latent space for  $M$  different body parts. This model is a generalization of the Gaussian Process Dynamical Model [GPDM, Wang et al. 2008]: By setting the whole body as one body part ( $M = 1$ ) the CGPDM becomes the GPDM. If there are more than one body parts, each dynamical system predicts not only the next time-step of their associated body part, but also the temporal evolution of other body parts via coupling functions. This way, flexible coupling between body-parts is possible. The CGPDM can be regarded as a middle ground between DMPs encoding single DOFs, and the monolithic GPDM. The latent dynamical systems can thus be thought of as flexibly coupled CPGs routing commands to the muscles.

In contrast to DMPs, CGPDMs learn a full dynamical model of latent variables  $Y$  in discrete time, which are mapped onto the observed DOFs  $X$ . Both the dynamics mapping  $f^{i,j}(\cdot)$  (Eq. 5) from the latent space of body part  $j$  to body part  $i$  ( $i, j = 1 \dots M$ ), as well as the mapping from latent to observed space  $g(\cdot)$  (Eq. 6) are drawn from Gaussian process priors.  $dt$  denotes the time discretization step-size:

$$Y^i(t) = f^{j,i}(Y^j(t-dt)) + \epsilon_{Y,t}^i \quad (5)$$

$$X^i(t) = g^i(Y^i(t)) + \epsilon_{X,t}^i \quad (6)$$

The model can be trained in two ways: by maximum-a-posteriori inference (MAP), or by free energy minimization using variational approximations (Variational (Coupled) Gaussian Process Dynamical Model [v(C)GPDM, Velychko et al. 2018]). In our study we use  $M = 1, 3$  body parts and use both training methods: GPDM ( $M = 1$ , MAP), CGPDM ( $M = 3$ , MAP), vGPDM ( $M = 1$ , variational), vCGPDM ( $M = 3$ , variational).

Without variational approximations, due to the non-parametric GPs prior, the movements *are* the movement representation, which is not compact. Therefore, MAP-trained (C)GPDMs, do not provide a complexity parameter.

The representation can be compressed by introducing sparse variational approximations. Now, each v(C)GPDM is parameterized by a small set of inducing points (IPs) and associated inducing values (IVs). The initial choice of IPs/IVs is the only remaining source of stochasticity in the training process. It may have measurable effects as we will show below.

We use IPs for both mappings, serving as ideal observer model parameters: “dynamics” IPs for the dynamical model mapping, and “pose” IPs for the latent-to-observed variable mapping. More dynamics IPs allow for richer dynamics (similar to the parameters of DMP and TMP), while more pose IPs will allow for more (spatial) variability of poses. An IP/IV pair might be thought of as a prototypical example for the mappings drawn from their associated Gaussian process. They thus provide some abstraction from the observed movement and might be implemented by small neuronal populations.

## 3.2 Experiments

This study includes two experiments: first, we conducted a highly controlled and ecologically valid VR-Experiment. Then, we decided to specifically study effects of contact events on perceptual predictions using the same paradigm with additional occlusion timing conditions. After we made this decision, the COVID-19 pandemic forced us to close our VR-Lab. This triggered us to port the VR-Experiment to an online setting. As benefit we could collect more data with less effort, but we as drawback we we could not control the viewing conditions under which participants performed the experiment. The VR experiment was implemented using Vizard 5 [WorldViz 2019] and the online experiment was implemented using the javascript library jsPsych [De Leeuw 2015] and WebGL. A test version of the online experiment can be tried online<sup>1</sup>.

In general, the methods of this work first comprise learning the recorded movements via extraction of MPs from mocap data, resulting in 3D joint locations and trajectories. The joint locations of both model-extracted and natural movement data are then connected (rigged) to a digital avatar (VR experiment) or a skeleton stick figure (online experiment). For the VR experiment, the rigged avatar, containing both natural and model-generated movements is then imported in a VR environment. For the online experiment, the movements are rendered in WebGL.

<sup>1</sup><http://vhrz1092.hrz.uni-marburg.de/javascriptbv/experiment.html?subject=xyz>.

We use this stimulus material for a psychophysical experiment in the form of a Graphics Turing Test [McGuigan 2006] on human movement prediction performance. In both experiments, the participants execute forced-choice trials, deciding which movement continuation fits best to a given beginning. The experiments' data comprises the relative frequency of a MP model successfully confusing participants to prefer its generated movement to a natural movement continuation. We call this frequency 'confusion rate'.

**3.2.1 Movements.** All presented movement consists of putting a bottle from one side of a table in front of the torso, where the bottle is passed to the other hand, to the other side of the table while sitting on a chair. Four kinds of movements are used: Passing the bottle from the left side to the right side (pass-bottle-movements), and vice versa (return-bottle- movements), and from the left to the right side without a pause (pass-bottle-hold-movements) but instead passing the bottle directly to the right hand, and vice versa (return-bottle-hold). Motor expertise/experience [Graf et al. 2007; Stadler et al. 2012] and visual familiarity of the movements to one's own movements [Loula et al. 2005] influences prediction performances. Simple movements of passing a bottle are actions with a low demand of motor expertise. Therefore, participants are not expected to strongly differ in their prediction performance due to expertise or familiarity.

**3.2.2 Stimulus Generation.** We recorded movements from one actor for the experiment with a PhaseSpace Impulse X2 System and 44 active LED markers. We inferred skeleton and joint angles from the recorded C3D-files, which contain marker positions in the recorded time frames using our own skeleton estimation software. These are used by computational implementations of the MP models to learn from five different bottle-passing movements for each movement type. The models then generate Biovision bvh-files containing joint locations and their trajectories from 5 different starting positions.

For the VR experiment, the bvh-files are then imported into the Autodesk MotionBuilder environment, where the bvh-joints are manually rigged onto a custom skeleton of a gray avatar polygon mesh. The rigging is adopted for all other bvh files with a custom script. The rigged avatar is then imported to the Autodesk 3dsMax environment, where the avatar and the movements were converted into a cfg-file, containing avatar mesh, skeleton and animation files, which was then importable for the Vizard 5 software, with which the experiment was designed.

For the online experiment, we used a simple stick figure to display the movement 2: the bvh-files produced by the MP-models are converted into pairs of 3D positions, where each pair is start and end-point of a segment specified by the skeleton. Each pair is then rendered using the GL\_LINES OpenGL-primitive. As we have no control over the setting and state of the subject when she is running the experiment, we added attention check trials. For this, we used movements generated by DMP models which obviously failed, such as avatars floating up from the chair. We excluded experimental runs where participants failed to correctly identify the floating movement in more than 40% of attention checks. In the VR experiment no such attention checks were needed, and we fixed the avatar's pelvis to the chair for these movements.

**3.2.3 Stimulus Presentation.** Elements of the trial structure were adopted from Sparenberg et al. [2012], who implemented experiments on internal simulation of movement and Graf et al. [2007], who tested various occlusion times in a psychophysics experiment of movement prediction performance, where participants were instructed to identify 1 of 2 action continuations as the most fitting to the beginning of the action before the occlusion. The structure of a trial can be viewed in Figure 1. Textures and objects for the experiment environment were provided by WorldViz and the website www.sketchfab.com. Presenting the two stimuli sequentially instead of simultaneously has the advantage of participants not having to distribute their fixations across the stimuli and instead could focus each stimulus separately.

**3.2.4 Catch Trials.** Instead of predicting the correct movement continuation, participants might instead use the unintended strategy of only distinguishing the first and second movement continuation as more or less natural-looking, ignoring the movement onset presented before the occlusion. Participants also might be less attentive to the experiment, resulting in higher confusion rates on average. To measure these variables, the experiment includes so-called "catch-trials", of which 24 were implemented for each participant in the VR experiment, and 2 for each experimental run in the online experiment. A catch-trial has the structure of a standard trial, but replaces the model-generated movement continuation with the same natural movement continuation as in the other movement sequence. This catch-continuation sequence will be time-incoherent to the movement-offset before the occlusion: catch-movements of the VR experiment start either 400 ms, 700 ms or 1000 ms (8 trials per participant) before the natural movements and therefore make the natural movement look as if they skipped movement frames during the occlusion. Catch-trials measure the rate of erroneously choosing the time-incoherent action continuation as a natural continuation. Time delays of the catch-trials were inspired from Graf et al. [2007]. We adapted the skip-timings for the online experiment slightly to 375 ms, 667 ms and 1000 ms to increase the range of investigated shifts.

**3.2.5 VR Experiment.**

**Participants:** N = 34 participants (23 female, 18-39 years old, mean age = 22,7 years, SE = 3,3 years) were recruited. As recruitment criteria, participants had to be 18 years or older and had to have no impaired vision. They also should not suffer from a disease of the musculoskeletal system, in order to handle HTC Vive controllers for the experiment. Participant recruitment was organized and promoted with the Sona Systems® participant management software. They received financial compensation (8€/h) or course credits for participation. An ethics application for the experiment had been approved by the local ethics commission (Ethikantrag 2015-19K). Participants received written information about the experiment in the participant management software and on the participant information sheet as well as in an instructional text in the VR environment. Participants gave their informed consent to participate.

**Experimental Procedure:** Participants were asked to sit on the experiment chair and were instructed to wear the head-mounted display (HMD) and HTC Vive controllers. As soon as participants

felt comfortable wearing the HMD, the experiment environment was loaded and the experiment started with an instructional text on the trial structure followed by nine familiarization trials where participants received feedback on their performance after each trial. After the familiarization trials, participants were additionally instructed to keep their gaze mainly focused on the avatar and their arms rested on their laps. Participants then started with the first of 269 trials. The trial number is derived from 24 catch-trials plus 5 repetitions of all 49 MP models. The trials were separated into four blocks of each 67 to 68 trials. After each block participants could take off the HMD and take a break of up to 5 minutes. Both catch- and normal trials were distributed randomly through all trials. Whether a trial presents a pass-hold-, pass- or return-movement was also randomized but is selected for both natural and model-generated movements presented in it. Each experiment run took about 70 to 85 minutes including all aforementioned procedures. Nine participants reported fatigue and one participant reported eye fatigue. One participant reported a headache after finishing a few trials of the first experiment block and aborted the experiment.

### 3.2.6 Online Experiment.

*Participants:* We collected data from 220 experiment runs of  $N = 98$  participants using the university’s participant management system (SONA System). The only metadata collected was a participant ID assigned by SONA. Participants were psychology students. They received course credits for participation.

*Experimental Procedure:* The experimental procedure is similar that of the VR experiment. We skipped the familiarization trials. Each experiment had 55 normal trials and two catch trials. This results in a length of approximately 15 minutes. Trials were sampled randomly from all possible trials for each experiment. Therefore each participant was allowed an arbitrary number of repetitions of the experiment. Each participant has a fixed anonymous ID assigned by the SONA participant management system. In the advertisement of the experiment we recommended 6 repetitions, but we did not control this number.

## 3.3 Data Analysis

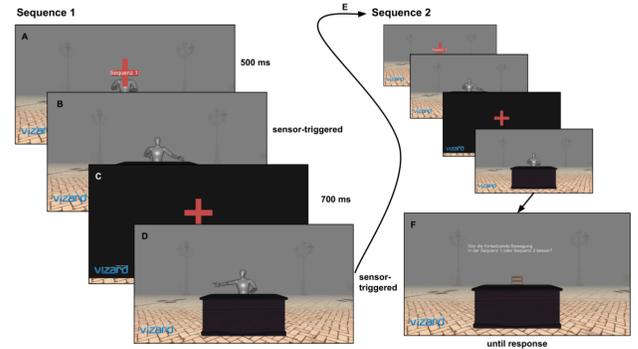
*3.3.1 Confusion Rate.* Participants were forced to choose one of the two sequences in each trial. Therefore, in trial  $i$  the participant’s response is  $r_i = 0$  if she guessed the wrong sequence and  $r_i = 1$  if the participant chose the correct sequence. We pooled across participants to achieve sufficient statistical power.

The confusion rate ( $p$ ) is defined as the number of times a participant erroneously chooses the sequence containing a model-generated movement continuation as more fitting to the movement onset divided by the total number of trials  $N$ .

$$p = \frac{N - \sum_i r_i}{N} \quad (7)$$

We assume that  $p$  approaches 0.5 if the model perfectly matches the observers perceptual predictions. The confusion rate measures the model success while  $1 - p_i$  measures human discrimination ability. We chose to report the confusion rate, as we are interested in comparing the models.

Each trial is specified by:



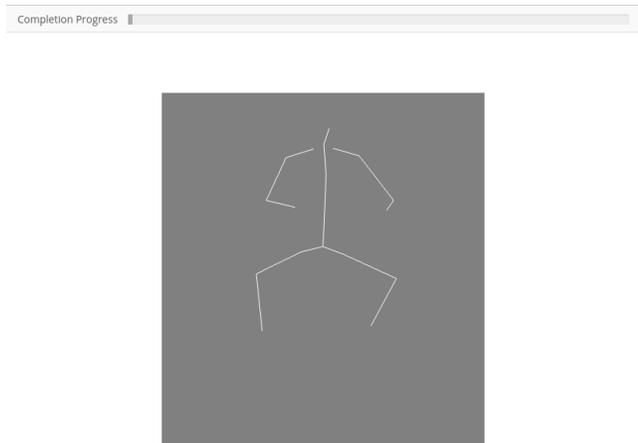
**Figure 1: Trial structure of the psychophysics experiment.** Each trial consists of two sequences, each beginning with (A) a red fixation cross appearing for 500ms in front of the desk for fixating the gaze towards the avatar, followed by (B) the onset of a natural movement randomly chosen from the set of 6 pass-hold-, 10 pass- and 9 return-movements, but based on the model-generated movement type. As soon as the hand returns to the front of the avatar, (C) an occlusion is triggered that lasts for 700ms. During the occlusion the movement is continued. After the occlusion, (D) the movement is continued by either the avatar performing the natural movement, with which the sequence has started, or an avatar performing the model-generated movement. The occurrence of the natural movement continuation in the first or second sequence is randomized. The end of the movement triggers either (E) starting sequence 2 or (F) making the visible avatar disappear and asking the participant for choosing the sequence with the correct movement continuation: “Which sequence did you perceive as more natural?”. The second sensor is activated 300 ms after the hand of the avatar enters it. This ends the movement sequence as soon as the hand is about to return to a position in front of the avatar. After choosing a sequence (by pressing the trigger-button on either the left HTC Vive controller for sequence 1, or the button on the right HTC Vive controller for sequence 2) the next trial starts.

- MP type with parameters:
  - TMP: Number of MPs  $Q$ .
  - DMP: Number of basis function  $N$ .
  - v(C)GPDM: Number of dynamical and pose IPs.
  - MAP-GPDM: No parameters.
- Movement: With or without table contact.
- Direction: From left to right, or vice versa.
- Training data set.
- Model scores after training.

In the online experiment there are furthermore three occlusion conditions:

- Occlusion timing: before, during, or after passing the center of the table

We assume that confusion rate  $p$  depends on a subset of these parameters. It might also be participant-specific.



**Figure 2: Screenshot of online experiment. The procedure was the same as described in Fig. 1, but participants respond by clicking buttons instead of using controllers.**

**3.3.2 Logistic Regression.** We assume that variable  $r_i$  is Bernoulli distributed and investigate the effect of cross-validators test set mean squared error (MSE), which is a proxy for the Bayesian model evidence. We obtain this MSE by training the MP models on 4 out of 5 movements, and then compute the mean squared residual between the reconstruction and actual observation of the 5th movement.

We use the *centered MSE* =  $\text{MSE} - \mathbb{E}[\text{MSE}]$  as predictor for the participants' responses using a Bayesian logistic regression model:

$$r_i \sim \text{Bernoulli}(p_i) \quad (8)$$

$$p_i = \frac{1}{1 + \exp(-(\alpha + \beta \cdot \text{MSE}_i))} \quad (9)$$

$$\alpha, \beta \sim \mathcal{N}(0, 10) \quad (10)$$

The participants' responses are Bernoulli distributed, with parameter  $p_i$  being the output of the sigmoid model with parameters  $\alpha$  and  $\beta$ . We set a wide Gaussian prior on these parameters and compute their posterior using Markov chain Monte Carlo<sup>2</sup>.

## 4 RESULTS

First, we compare the results of the VR- and the online experiments and contrast these with previous findings in a naturalness perception experiment [Knopp et al. 2019]. We demonstrate that our paradigm works as intended by presenting the catch trial results. We then show the predictions of logistic regression for different MP types and finally present results demonstrating the effect of contact events in our experiments.

### 4.1 Comparison of Experiments

Figure 3 compares the mean confusion rates over complexity parameters of MP-Types of a naturalness perception experiment [Knopp et al. 2019] with the two experiments described in this study. The

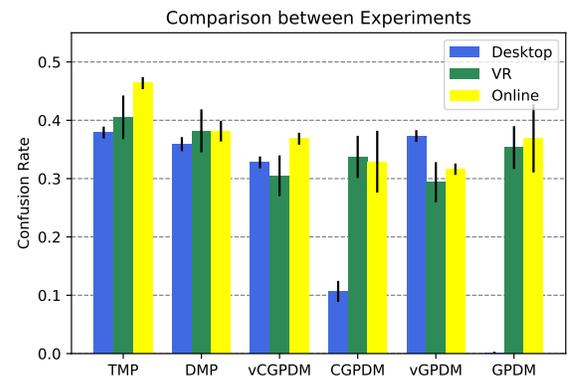
<sup>2</sup>We use the No-U-Turn Sampler implemented in Python library PyMC3 [Salvatier et al. 2016].

previous experiment measured confusion rate in a task where participants had to choose the more natural one of two walking movements. One of the movements in each trial of that experiment was MP generated, the other one was a replay of a natural movement recording.

Considering the differences regarding movement (walking vs. object-passing), experimental paradigm (prediction vs. identification), setting (desktop vs. VR vs. web-based), and representation (full avatar vs. stick-figure), the confusion rates are remarkably similar.

TMP models consistently perform best. DMP models perform well in all settings, too. The prediction and identification paradigms differ very much regarding the training mode of the (C)GPDMs: MAP training failed to fool subjects to mistake the generated walking movements in the identification task, but performed on par with the variationally trained models in the pass-object prediction task.

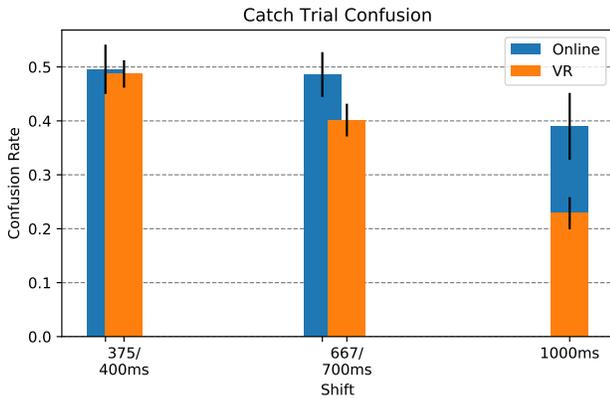
We used attention checks in the online experiment (highly unrealistic floating movements were shown), to filter experiment runs with inattentive subjects. Still, there is a slight tendency of slightly higher confusion rates in the online experiment compared to the VR setting.



**Figure 3: Confusion rate for MP models in three different experiments. 1. Previously published desktop experiment, 2. VR experiment and 3. online experiment from this study. Error bars depict beta-distributed standard error.**

### 4.2 Catch Trial Results

We recorded participants' performance of falsely identifying the discontinuous movement continuation as the one most fitting the movement onset before the occlusion in 809 catch trials in the VR experiment, and in 318 catch trials (up to now) in the online experiment. Figure 4 shows the resulting confusion rates. The smallest shift of 375/400 ms is not detected, as the confusion rate is close to 0.5. The rate decreases for the conditions with larger shifts. The decrease is more pronounced for the VR data compared to data collected by the online experiment.



**Figure 4: Confusion rate for catch trials with different time shifts for the two experimental conditions. The bars of the two experiments are shifted relative to each other, because we changed the shift timings slightly for the online experiment.**

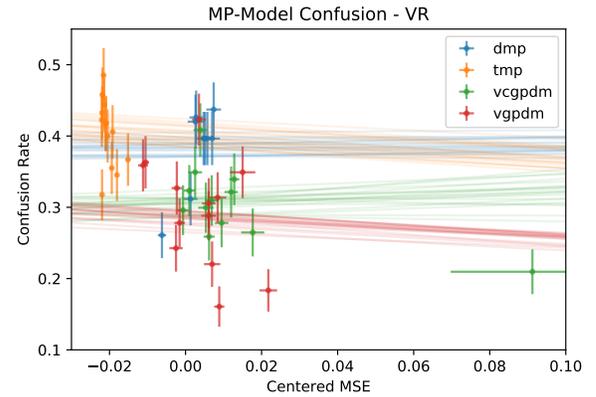
**Table 1: Mean and standard deviation of posterior samples of parameters  $\alpha$  and  $\beta$ .**

	VR	Online
$\alpha$		
DMP	$-0.48 \pm 0.05$	$-0.50 \pm 0.08$
TMP	$-0.41 \pm 0.04$	$-0.15 \pm 0.04$
vCGPDM	$-0.83 \pm 0.05$	$-0.55 \pm 0.05$
vGPDM	$-0.89 \pm 0.05$	$-0.78 \pm 0.05$
$\beta$		
DMP	$0.09 \pm 0.14$	$-0.65 \pm 0.18$
TMP	$-1.29 \pm 0.16$	$-0.09 \pm 0.21$
vCGPDM	$0.67 \pm 0.54$	$-3.54 \pm 0.95$
vGPDM	$-1.49 \pm 0.27$	$-3.09 \pm 0.72$

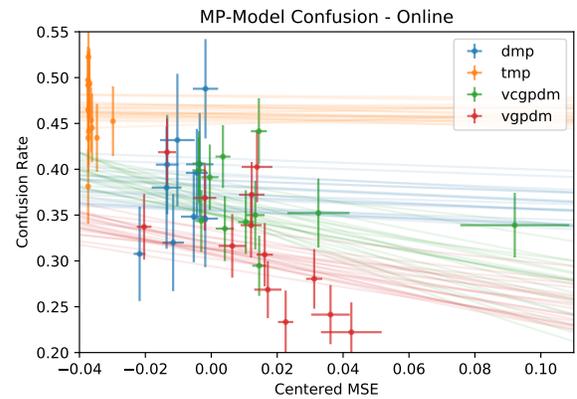
### 4.3 Predicting Perceptual Predictions from Centered MSE

We predict the confusion rate, which is our measure for the different MP types' ability to generate movements in line with human perceptual predictions, from centered MSE using logistic regression (3.3.2). Figure 5 shows confusion rates of MP models over mean MSE. In general, lower MSE corresponds to higher confusion rate (negative slope  $\beta$ ). We do not observe this relationship for DMP, vCGPDM models tested in the VR experiment. TMP models tested in the online experiment have a near-zero negative slope  $\beta$ . TMP models of the VR experiment on the other hand show the strongest dependence of the confusion rate on the MSE, together with vGPDM models. We summarize the posterior of parameters in Table 1.

**A**



**B**

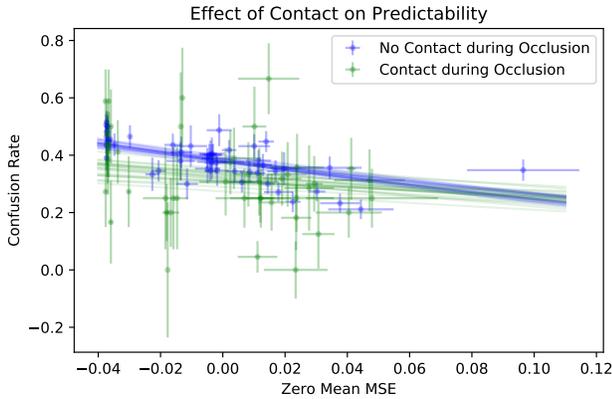


**Figure 5: Confusion rate and logistic regression for different MP model types for data collected in (A) VR- and (B) online experiment. Each point shows the mean confusion rate for a MP model with specific parameters against the centered MSE (which is the MSE with subtracted mean). Error bars show the beta distributed standard deviation and the standard error of the MSE. Lines are predictions of the logistic regression model with 20 samples of parameters  $\alpha$  and  $\beta$ .**

### 4.4 Effect of Contact Event on Perceptual Predictions

In our online experiment we collected data for movements where a bottle is passed from one hand to the other either with or without touching the table. We varied the occlusion timing to investigate the effect of the table contact on perceptual prediction performance of MSE: The movement was occluded before, during, or after bottle was passed. We compare the influence of MSE on the confusion rates of trials with occlusion during table contact with the rest of the trials. For this we use logistic regression [3.3]. Here, the slope  $\beta$  is a measure of influence of MSE on the confusion rate. Given the posteriors of  $\beta$  we can compute the probability of  $|\beta_{nc}|$  for trials without occluded contact being greater than  $|\beta_c|$  for trials with occluded contact:  $p(|\beta_{nc}| > |\beta_c|) = 0.998$ . We are thus fairly

certain that MSE loses predictive capability if object-table contact is occluded.



**Figure 6: Predictions by logistic regression model for trials either with or without contact during occlusion, and confusion rate of models plotted against the centered MSE (same as fig. 5). In case of contact during occlusion, the slope of the fit is smaller, making MSE less useful predictor.**

## 5 DISCUSSION

We compared 3 different types of MP-models using a predictive paradigm in two settings: VR and web-browser based. The representation of the movement was different as well: 3D avatars in the VR-, and stick figures in the online experiment. We also compared these results to published data in [Knopp et al. 2019], which used different movements, and a non-predictive paradigm. Our results indicate that measured confusion rates generalize across movements, paradigm, and rendering specifics. A notable exception is the dramatic performance increase of MAP-trained (C)GPDMs. We suspect that the initialization of this model in the previous experiment might have been unfavorable.

Participants of the online experiment have shown slightly worse prediction performance. We expect this is due to attentional and motivational shortcomings of a non-lab environment. Still, considering the substantially lower effort of running the experiment and highly increased reach for recruiting participants, this drawback is more than compensated. Reaching out to many participants is very important, as our experimental design, even though simple and elegant, is collecting very little information per trial (1 bit). Still, a problem remaining are potential inter-individual differences. Because participants are exhausted very quickly, we can only test a small subset of all models and conditions. Pooling across participants while still accounting for inter-individual differences might be useful and we will explore this in the next study.

Catch trials show decreasing confusion rate for increasing time-shift of natural movements, which indicates that participants actually predict the movement, instead of rating the naturalness of the movement continuation without regard to the lead-in movement. This decrease is less pronounced in the data of the online experiment.

In our experiments, we found that TMP-models produce the most realistic movement. This is in line with previous findings [Knopp et al. 2019]. Therefore, TMPs might be used by the visual system for perceptual predictions. Dynamical models might still be involved in movement production because of their ability to handle perturbations. The shared representation between perception and production may therefore be more abstract: one dynamics model paired with a corresponding TMP model that encodes typical (unperturbed) solutions of the dynamics model, for fast perceptual predictions [Giese and Poggio 2000].

We use the centered MSE to predict perceived naturalness by using a logistic regression model for the confusion rate. The prediction worked well for TMP and vGPDM models of the VR experiment, and for vCGPDM and vGPDM of the online experiment. The online experiment might be the decisive bit harder for subjects, such that many TMP models come close enough to indistinguishability, impeding prediction. The vCGPDM has increased number of IP sets (one set for each body part) compared to the monolithic vGPDM. This introduces more stochasticity during training, resulting in large variation of the MSE. This might explain the different prediction results of the vCGPDM for the different experiments. Compared to previous findings [Knopp et al. 2019], the predictions are less reliable. This is because our experimental design is more complex, adding different movements and switching to a predictive sequential task. As previously discussed, more data is required to disentangle effects of different MP types, movements, and occlusion conditions.

Contact events are a common heuristic for the task of segmenting movements. Yet, there is little psychophysical investigation measuring the effect of models on segmentation, but see [Endres et al. 2011]. In our online experiment, we manipulated the occlusion timing to investigate the existence of perceptual segmentations induced by contact events. We found a higher expected increase of confusion rate for increasing MSE in trials where table contact was not occluded, which we interpret as follows: participants, who expect a contact event based on the previous trajectory, but can not see it, will have a less precise expectation about the continuation of the movement, making them less susceptible to higher deviations from their expectations. This is not the case for participants who see the contact, and can use frames after contact to build a more precise expectation.

The current work is the new and unexplored implementation of a Graphics Turing Test of movement prediction performances. Even though the structure of the prediction task was mostly adapted from other works [Graf et al. 2007; Knopp et al. 2019], conducting this task in the context of a Graphics Turing Test in a VR and web-based environment is novel and has yet to be established more firmly in psychophysical research.

## 6 CONCLUSIONS

The present work created a psychophysical task for visual prediction performances in a VR and web-based environment and implemented it to gather psychophysical data on six different representations of motor actions based on MPs. MP models can be used to generate natural-appearing novel movements on virtual avatars, which is important for neuroscientists searching for a common code

of action and perception and might be applied to build realistic computer animation with less effort in the future. In future studies we want to validate the assumptions that the influence of different movement representations (stick-figure vs. 3D avatar) is small and compare different movements, to investigate the generalizability of our results.

## ACKNOWLEDGMENTS

This work was funded by DFG, IRTG1901 - The brain in action, and SFB-TRR 135 - Cardinal mechanisms of perception. We thank Olaf Haag for help with rendering of the stimuli and collecting data.

## REFERENCES

- Marco Bertamini. 1993. Memory for position and dynamic representations. *Memory & Cognition* 21, 4 (1993), 449–457.
- Giovanni Buccino, Ferdinand Binkofski, and Lucia Riggio. 2004. The mirror neuron system and action recognition. *Brain and language* 89, 2 (2004), 370–376.
- Enrico Chiovetto, Cristóbal Curio, Dominik Endres, and Martin A. Giese. 2018. Perceptual integration of kinematic components in the recognition of emotional facial expressions. *Journal of Vision* 18, 4 (April 2018), 13–13. <https://doi.org/10.1167/18.4.13>
- Debora Clever, Monika Harant, Henning Koch, Katja Mombaur, and Dominik Endres. 2016. A novel approach for the generation of complex humanoid walking sequences based on a combination of optimal control and learning of movement primitives. *Robotics and Autonomous Systems* 83 (Sept. 2016), 287–298. <https://doi.org/10.1016/j.robot.2016.06.001>
- Joshua R De Leeuw. 2015. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods* 47, 1 (2015), 1–12.
- Dominik Endres, Enrico Chiovetto, and Martin A. Giese. 2013. Model selection for the extraction of movement primitives. *Frontiers in Computational Neuroscience* 7 (2013), 185. <https://doi.org/10.3389/fncom.2013.00185>
- Dominik Endres, Andrea Christensen, Lars Omlor, and Martin A. Giese. 2011. Emulating human observers with Bayesian binning: segmentation of action streams. *ACM Transactions on Applied Perception (TAP)* 8, 3 (2011), 16:1–12.
- Jennifer J Freyd and Ronald A Finke. 1984. Representational momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 10, 1 (1984), 126.
- Karl Friston. 2010. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* 11, 2 (Feb. 2010), 127–138. <https://doi.org/10.1038/nrn2787>
- Martin A. Giese and Tomaso Poggio. 2000. Morphable Models for the Analysis and Synthesis of Complex Motion Patterns. *International Journal of Computer Vision* 38 (June 2000), 59–73. <https://doi.org/10.1023/A:1008118801668>
- Markus Graf, Bianca Reitzner, Caroline Corves, Antonino Casile, Martin Giese, and Wolfgang Prinz. 2007. Predicting point-light actions in real-time. 36 (2007), T22–T32. <https://doi.org/10.1016/j.neuroimage.2007.03.017>
- Jakob Hohwy. 2013. *The predictive mind*. Oxford University Press.
- Marco Iacoboni and Mirella Dapretto. 2006. The mirror neuron system and the consequences of its dysfunction. *Nature Reviews Neuroscience* 7, 12 (2006), 942.
- Auke Jan Ijspeert, Jun Nakanishi, Heiko Hoffmann, Peter Pastor, and Stefan Schaal. 2013. Dynamical Movement Primitives: Learning Attractor Models for Motor Behaviors. *Neural Computation* 25, 2 (Feb. 2013), 328–373. [https://doi.org/10.1162/NECO\\_a\\_00393](https://doi.org/10.1162/NECO_a_00393)
- Yuri P. Ivanenko, Richard E. Popple, and Francesco Lacquaniti. 2004. Five basic muscle activation patterns account for muscle activity during human locomotion: Basic muscle activation patterns. *The Journal of Physiology* 556, 1 (April 2004), 267–282. <https://doi.org/10.1113/jphysiol.2003.057174>
- Mohamed Jarraya, Michel-Ange Amorim, and Benoît G Bardy. 2005. Optical flow and viewpoint change modulate the perception and memorization of complex motion. *Perception & psychophysics* 67, 6 (2005), 951–961.
- James M Kilner, Claudia Vargas, Sylvie Duval, Sarah-Jayne Blakemore, and Angela Sirigu. 2004. Motor activation prior to observation of a predicted movement. *Nature neuroscience* 7, 12 (2004), 1299–1301.
- David C. Knill and Alexandre Pouget. 2004. The bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neuroscience* 27 (2004).
- Benjamin Knopp, Dmytro Velychko, Johannes Dreibrod, and Dominik Endres. 2019. Predicting Perceived Naturalness of Human Animations Based on Generative Movement Primitive Models. *ACM Trans. Appl. Percept.* 16, 3 (Sept. 2019), 15:1–15:18. <https://doi.org/10.1145/3355401>
- Fani Loula, Sapna Prasad, Kent Harber, and Maggie Shiffrar. 2005. Recognizing people from their movement. *Journal of Experimental Psychology: Human Perception and Performance* 31, 1 (2005), 210.
- Michael D. McGuigan. 2006. Graphics Turing Test. *CoRR abs/cs/0603132* (2006).
- Wolfgang Prinz. 1997. Perception and Action Planning. *European Journal of Cognitive Psychology* 9, 2 (June 1997), 129–154. <https://doi.org/10.1080/713752551>
- Giacomo Rizzolatti and Laila Craighero. 2004. The mirror-neuron system. *Annu. Rev. Neurosci.* 27 (2004), 169–192.
- J Salvatier, TV Wiecki, and C Fonnesbeck. 2016. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2, e55.
- Mirko Sattler, Ralf Sarlette, and Reinhard Klein. 2005. Simple and efficient compression of animation sequences. (2005), 209–217. <https://doi.org/10.1145/1073368.1073398>
- Stefan Schaal. 2006. Dynamic Movement Primitives -A Framework for Motor Control in Humans and Humanoid Robotics. In *Adaptive Motion of Animals and Machines*, Hiroshi Kimura, Kazuo Tsuchiya, Akio Ishiguro, and Hartmut Witte (Eds.). Springer-Verlag, Tokyo, 261–280. [https://doi.org/10.1007/4-431-31381-8\\_23](https://doi.org/10.1007/4-431-31381-8_23)
- Erich Schröger, Anna Marzecová, and Iria SanMiguel. 2015. Attention and prediction in human audition: a lesson from cognitive psychophysiology. *European Journal of Neuroscience* 41, 5 (2015), 641–664.
- Carl Senior, J Barnes, V Giampietroc, A Simmons, ET Bullmore, M Brammer, and AS David. 2000. The functional neuroanatomy of implicit-motion perception or ‘representational momentum’. *Current Biology* 10, 1 (2000), 16–22.
- Peggy Sparenberg, Anne Springer, and Wolfgang Prinz. 2012. Predicting others’ actions: Evidence for a constant time delay in action simulation. *Psychological Research* 76, 1 (2012), 41–49.
- Waltraud Stadler, Anne Springer, Jim Parkinson, and Wolfgang Prinz. 2012. Movement kinematics affect action prediction: comparing human to non-human point-light actions. *Psychological research* 76, 4 (2012), 395–406.
- Nick Taubert, Andrea Christensen, Dominik Endres, and Martin A. Giese. 2012. Online Simulation of Emotional Interactive Behaviors with Hierarchical Gaussian Process Dynamical Models. *Proceedings of the ACM Symposium on Applied Perception (ACM-SAP 2012)* (2012), 25–32. <https://doi.org/10.1145/2338676.2338682>
- Ian Thornton and Amy Hayes. 2004. Anticipating action in complex scenes. *Visual Cognition* 11, 2-3 (2004), 341–370.
- Dmytro Velychko, Dominik Endres, Nick Taubert, and Martin A. Giese. 2014. Coupling Gaussian Process Dynamical Models with Product-of-Experts Kernels. In *Proceedings of the 24th International Conference on Artificial Neural Networks, LNCS 8681*. Springer, 603–610.
- Dmytro Velychko, Benjamin Knopp, and Dominik Endres. 2018. Making the Coupled Gaussian Process Dynamical Model Modular and Scalable with Variational Approximations. *Entropy* 20, 10 (Sept. 2018), 724. <https://doi.org/10.3390/e20100724>
- Jack Meng-Chieh Wang, David J. Fleet, and Aaron Hertzmann. 2008. Gaussian Process Dynamical Models for Human Motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 2 (Feb. 2008), 283–298. <https://doi.org/10.1109/TPAMI.2007.1167>
- Margaret Wilson and Günther Knoblich. 2005. The case for motor involvement in perceiving conspecifics. *Psychological bulletin* 131, 3 (2005), 460.
- WorldViz. 2019. Vizard 6.
- Jeffrey M. Zacks and Khen M. Swallow. 2007. Event Segmentation. 16, 2 (2007), 80–84. <https://doi.org/10.1111/j.1467-8721.2007.00480.x>

# F

## Zusammenfassung in deutscher Sprache

Menschen können menschliche Bewegungen anhand von spärlichen visuellen Informationen ohne Mühe erkennen und verstehen. Wenige Punkte, welche die Position der Gelenke markieren, reichen aus um eine lebhafte und stabile Wahrnehmung der dahinter liegenden Bewegung zu erzeugen. Aufgrund dieser Fähigkeit erfordert die realistische Animation von 3D-Figuren großes Geschick. Das Studium der Bestandteile einer natürlich wirkenden Bewegung würde nicht nur diesen Künstlern helfen, sondern auch ein besseres Verständnis der zugrunde liegenden Informationsverarbeitung im Gehirn ermöglichen.

Analog zu den Herausforderungen bei der Animation spiegelt die Arbeit der Robotiker die die Komplexität der Bewegungserzeugung: Die Steuerung der vielen Freiheitsgrade eines Körpers erfordert zeitaufwändige Berechnungen. Modularität ist eine Strategie, um dieses Problem zu adressieren: Komplexe Bewegungen können in einfache Primitive zerlegt werden. Umgekehrt lassen sich aus wenigen Primitiven eine große Anzahl komplexer Bewegungen zusammensetzen. Viele Arten von Bewegungsprimitiven (MPs) sind auf verschiedenen Ebenen der Informationsverarbeitungshierarchie im dem Gehirn vorgeschlagen wurden.

MPs wurden meist im Kontext der Bewegungsproduktion vorgeschlagen und verwendet. Eine auf Primitiven basierende Modularität könnte jedoch in ähnlicher Weise eine robuste Bewegungswahrnehmung ermöglichen.

Für meine Dissertation habe ich Wahrnehmungsexperimente durchgeführt, die auf der Annahme einer gemeinsamen Repräsentation von Bewegungswahrnehmung und -Produktion basierend auf MPs. Die drei verschiedenen Typen von MPs, die ich untersucht habe, sind temporale MPs (TMP), dynamische MPs (DMP), und gekoppelte Gaussian Process Dynamical Models (cGPDM).

Die MP-Modelle wurden auf Basis von natürlichen Bewegungen trainiert um neue Bewegungen zu generieren. Diese künstlichen Bewegungen wurden dann perzeptuell validiert in psycho-physikalischen Experimenten. In allen Experimenten verwendete ich ein Forced-Choice-Paradigma mit zwei Alternativen, in

dem menschlichen Beobachtern eine Bewegung basierend auf Motion-Capturing-Daten, und eine durch ein MP-Modell generierte Bewegung präsentiert wurde. Danach mussten sie die Bewegung wählen, die sie als natürlicher empfanden.

Im ersten Experiment untersuchte ich die Wahrnehmung von Gehbewegungen: In Übereinstimmung mit früheren Ergebnissen ist die getreue Darstellung der Bewegungsdynamik wichtiger ist als eine gute Rekonstruktion der Pose. Im zweiten Experiment untersuchte ich die Rolle der Vorhersage in der Wahrnehmung anhand von Objekthandhabungsbewegungen. Es stellte sich heraus, dass die wahrgenommene Natürlichkeit der Bewegungsvorhersagen ähnlich ist zu der wahrgenommenen Natürlichkeit der Bewegungen selbst, die im ersten Experiment festgestellt wurde.

Ich habe herausgefunden, dass die MP-Modelle in der Lage sind, Bewegungen zu produzieren, die natürlich aussehen, wobei TMP-Modelle die höchsten Wahrnehmungswerte erzielen. Darüber hinaus ermöglichen sie die Vorhersagbarkeit der wahrgenommenen Natürlichkeit, was auf ihre Eignung für eine gemeinsame Darstellung von Bewegungs-Wahrnehmung und -Produktion hindeutet.



## Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Dissertation selbstständig, ohne unerlaubte Hilfe Dritter angefertigt und andere als die in der Dissertation angegebenen Hilfsmittel nicht benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen sind, habe ich als solche kenntlich gemacht. Dritte waren an der inhaltlich-materiellen Erstellung der Dissertation nicht beteiligt; insbesondere habe ich hierfür nicht die Hilfe eines Promotionsberaters in Anspruch genommen. Kein Teil dieser Arbeit ist in einem anderen Promotions- oder Habilitationsverfahren verwendet worden. Mit dem Einsatz von Software zur Erkennung von Plagiaten bin ich einverstanden.