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

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

15

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.

1544-3558/2019/09-ART15
https://doi.org/10.1145/3355401

1 INTRODUCTION

The perception of biological movement[1] 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 obseved 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).

[1]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 negligbile.

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:

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

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 τ, 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.

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

$$\dot{\tau} \propto -\tau \tag{3}$$

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

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:

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

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

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, seving 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

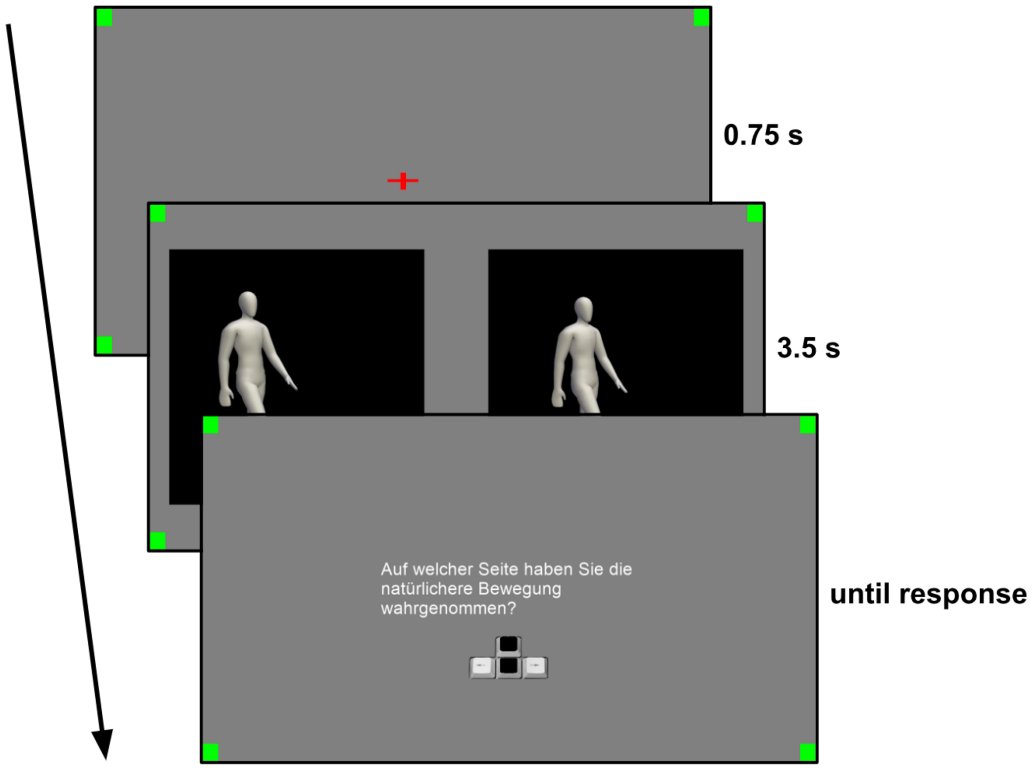

Fig. 1. Illustration of experimental procedure. Each trial begun with a fixation period of 0.75s. Then, participants watched simultaneous replays of natural and generated movements for 3.5s. After the presentation the participants were asked "On which side did you perceive the more natural movement?"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

a length of 3.5*s* 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 0.75*s*, (2) followed by simultaneous side-by-side

presentation of generated and natural stimuli for 3.5*s*, 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 30*s* 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):

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

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 H0: $p_i \in [0.45, 0.55]$ and H1: $p_i \notin [0.45, 0.55]$. We choose the number of trials such that the falsehood of H0 is discovered with power 0.8 when H1 is true, i.e., $1 - P(\mathrm{H}_0|\mathrm{H}_1) = 0.8$. 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:

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

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:

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

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

We now learn the weights $(w_0^\star, w_1^\star)$ 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. $p \neq 0.5$ ($p \sim beta(1, 1)$) is ≈ 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 ± 0.004 for the MAP-GPDM, and 0.11 ± 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.

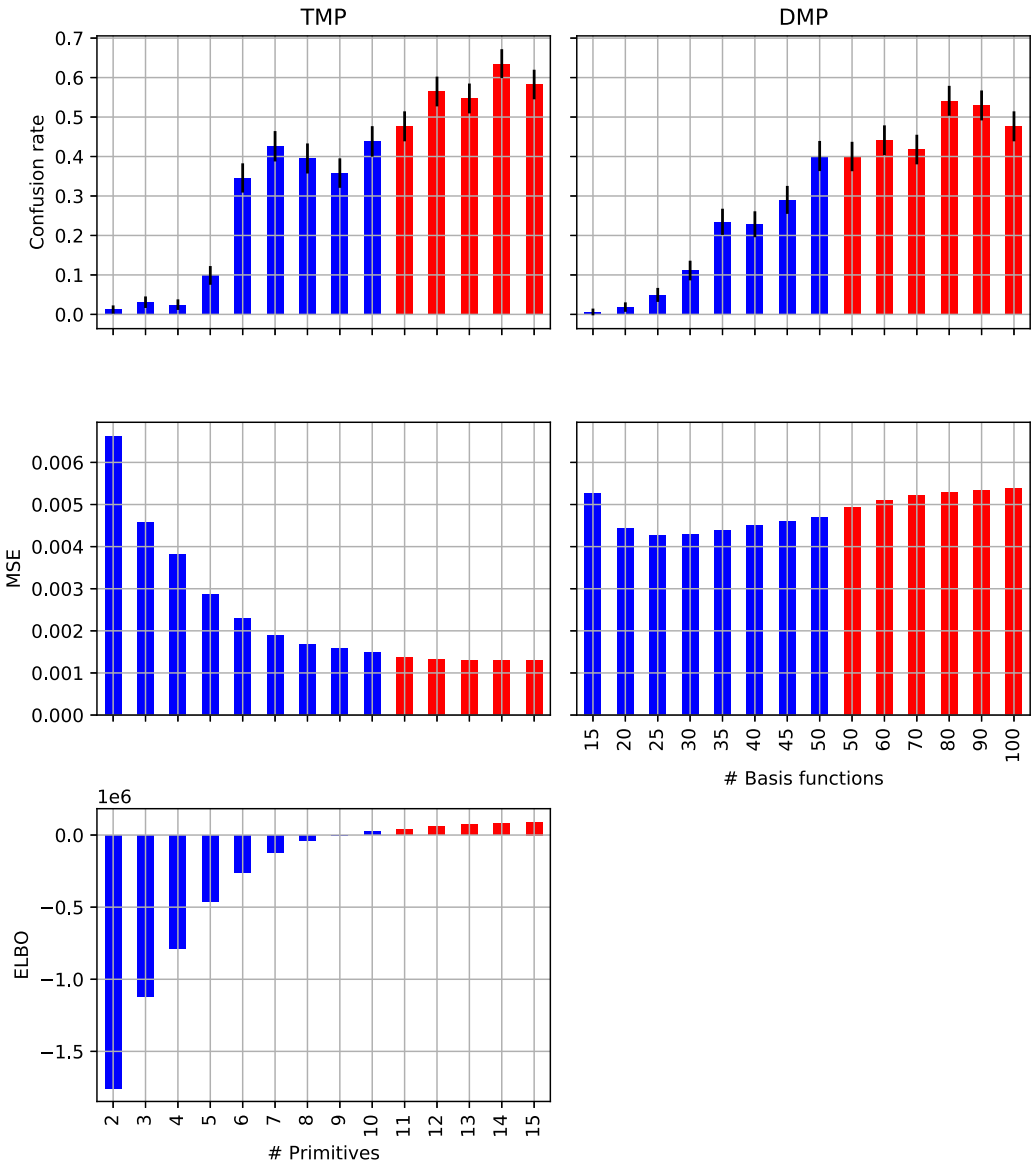

Fig. 2. Confusion rate, MSE, ELBO (from top to bottom) of TMP (left) and DMP (right) models for investigated model parameters. Data of first part of the experiment is colored blue, data of the second part is colored red.

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 beta(1,1) priors on the rate and a Bernoulli observation model, as before. The most probable ordering is TMP > vGPDM > DMP > vCGPDM > CGPDM(MAP) > GPDM(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 ≈ 0.99 sure of that.

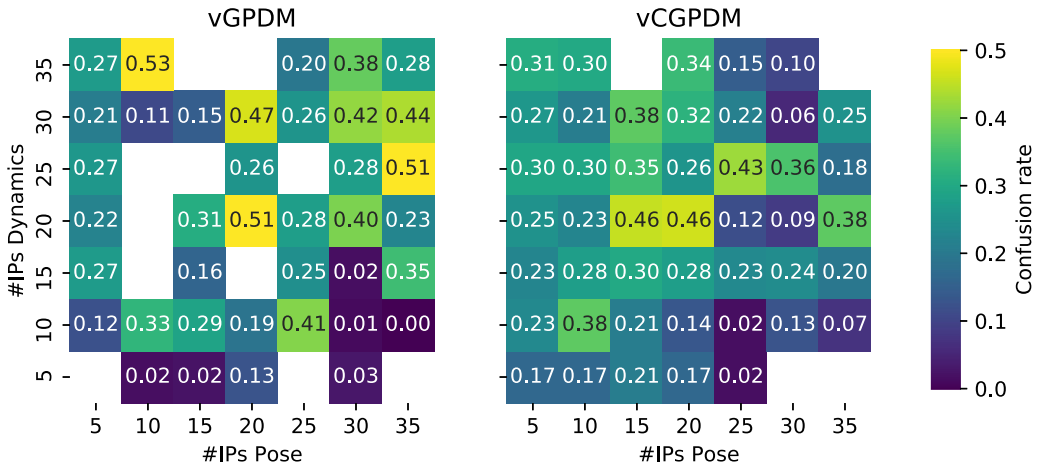

Fig. 3. Confusion rate of v(c)GPDM models in first part of experiment: Number of inducing points for the pose mapping on the x-axis, and for the dynamics mapping on the y-axis. The attention check parameter combinations are indicated by the white squares, where the model training procedure converged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

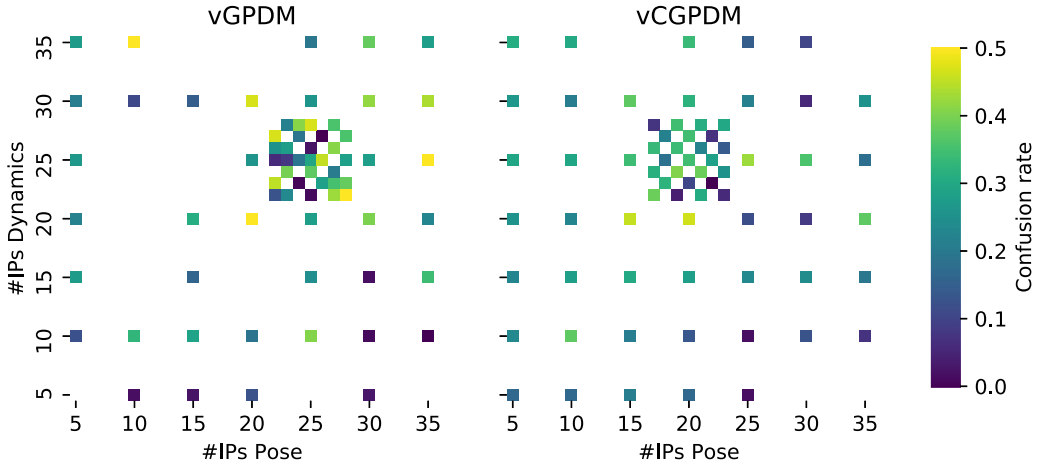

Fig. 4. Confusion rate of v(c)GPDM of first and second part of the experiment: Data of second part of the experiment are clustered around (25, 25) for vGPDM and (25, 20) for vCGPDM. Confusion rates are indicated by the same color-map as in Figure 3.

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

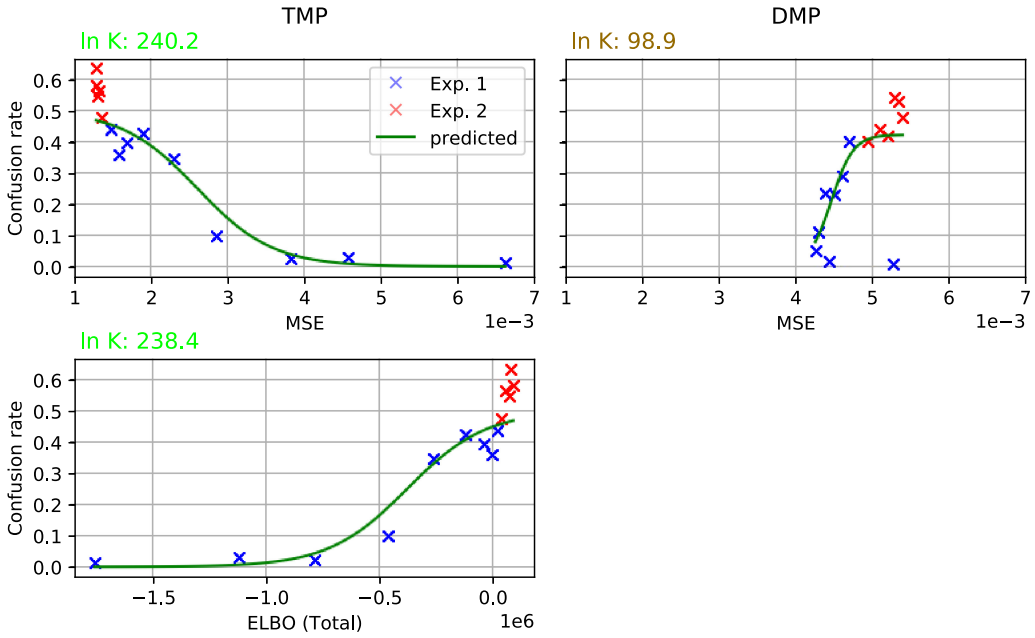

Fig. 5. Confusion rate of TMP (left) and DMP (right) models as function of model scores: MSE (top) and ELBO (bottom). Blue and red "X"s show confusion rates for model-parameters measured during experiment one and two. Green lines are predictions of the confusion rate (perceived naturalness) from the 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

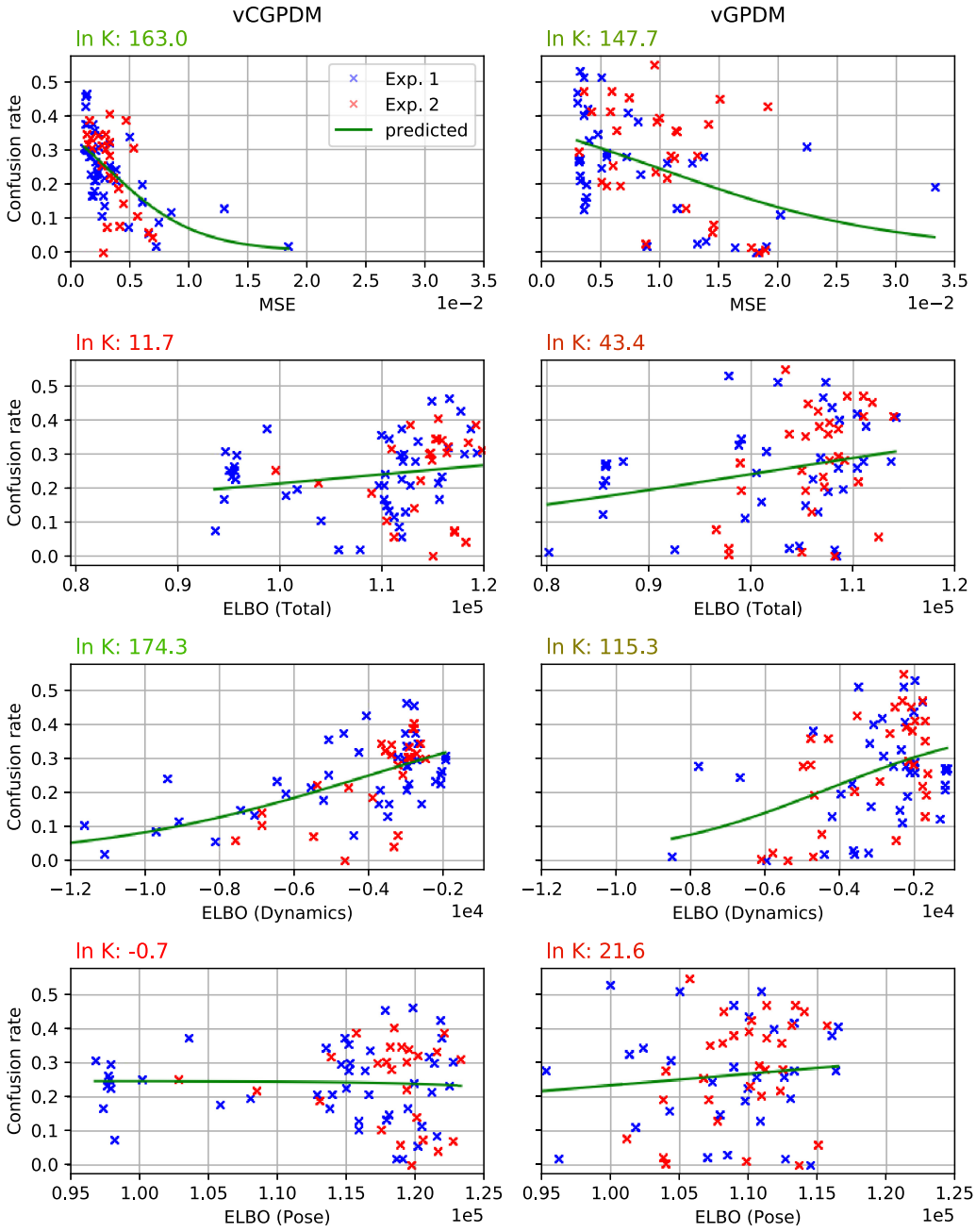

Fig. 6. Confusion rate of vCGPDM (left) and vGPDM (right) models as function of model scores: MSE, Total-, Dynamics-, Pose-ELBO (from top to bottom). Symbols have the same meaning as in Figure 5. See 3.3 for more detail.

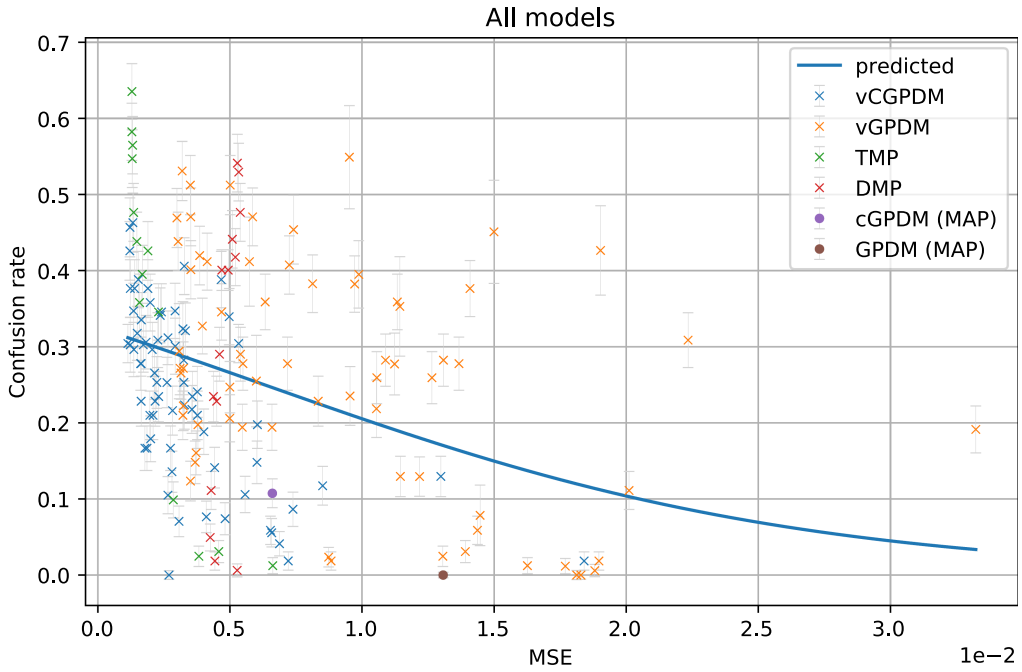

Fig. 7. Confusion rate of all models vs. test MSE and prediction learned over all models. Same data as in the first rows of Figures 5 and 6 plus cGPDM (MAP) and GPDM(MAP). Error bars denote beta standard deviation of the confusion rate.

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

	Parameter	Confusion rate	Error
TMP	14	0.64	±0.04
vGPDM	(22,28)	0.55	±0.07
DMP	80	0.54	±0.04
vCGPDM	(20,20)	0.46	±0.04
cGPDM (MAP)	-	0.11	±0.02
GPDM (MAP)	-	0.00	±0.00

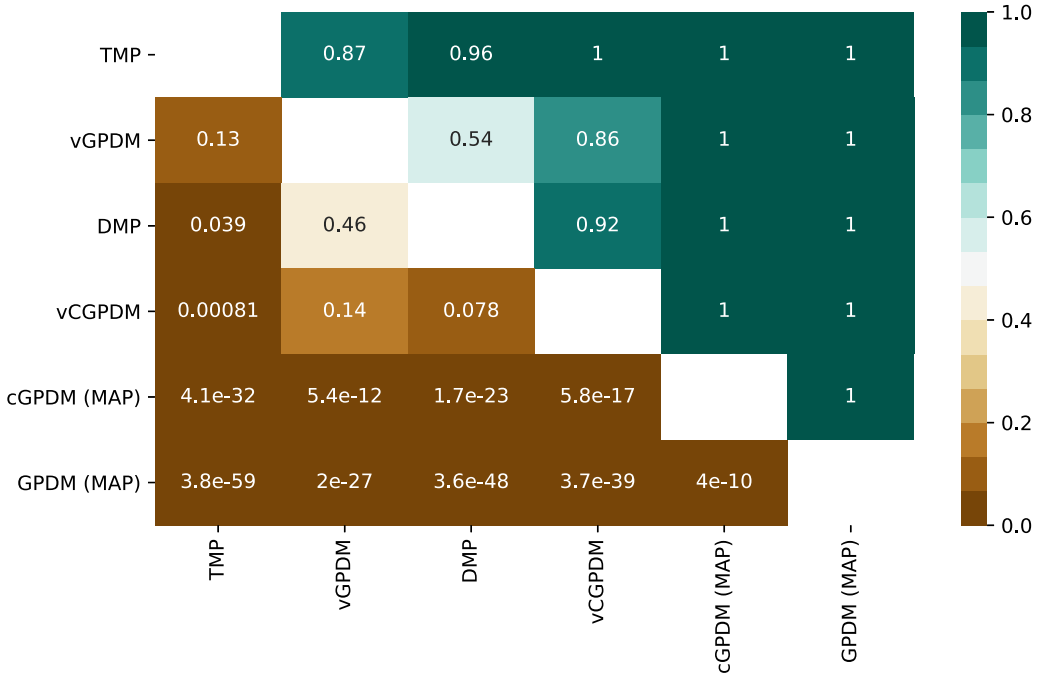

Fig. 8. Comparison of best models: (Top) Table of best models with corresponding parameter(-combinations), confusion rate and standard errors of beta posteriors. (Bottom) Bayesian ordering tests: probabilities that the best parameter combination of the models in the rows yields a higher confusion rate 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.[2]

[2]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. Simultanous 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=kX5OAQAAIAAJ

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

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

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

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

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