Online Simulation of Emotional Interactive Behaviors with Hierarchical Gaussian Process Dynamical Models

Nick Taubert[*], Andrea Christensen[†], Dominik Endres[‡], Martin A. Giese[§]
Section Computational Sensomotorics, Dept. of Cognitive Neurology,
Hertie Institute for Clinical Brain Sciences and Center for Integrative Neuroscience, University Clinic Tübingen,
Ottfried-Müller Str. 25, 72076 Tübingen, Germany

Figure 1: *Motion sequences of synthesized emotional handshakes. From left to right: neutral, angry, happy and sad. Different emotions are associated with different postures and ranges of joint movements.*

Abstract

The online synthesis of stylized interactive movements with high levels of realism is a difficult problem in computer graphics. We present a new approach for the learning of structured dynamical models for the synthesis of interactive body movements that is based on hierarchical Gaussian process latent variable models. The latent spaces of this model encode postural manifolds and the dependency between the postures of the interacting characters. In addition, our model includes dimensions representing emotional style variations (for neutral, happy, angry, sad) and individually-specific motion style. The dynamics of the state in the latent space is modeled by a Gaussian Process Dynamical Model, a probabilistic dynamical model that can learn to generate arbitrary smooth trajectories in real-time. The proposed framework offers a large degree of flexibility, in terms of the definition of the model structure as well as the complexity of the learned motion trajectories. In order to assess the suitability of the proposed framework for the generation of highly realistic motion, we performed a 'Turing test': a psychophysical study where human observers classified the emotions and rated the naturalness of the generated and natural emotional handshakes. Classification results for both stimulus groups were not significantly different, and for all emotional styles, except for neutral, participants rated the synthesized handshakes equally natural as animations with the original trajectories. This shows that the proposed method generates highly-realistic interactive movements that are almost indistinguishable from natural ones. As a further extension, we demonstrate the capability of the method to interpolate between different emotional styles.

CR Categories: I.2.9 [Artificial Intelligence]: Robotics—Kinematics and dynamics; I.2.6 [Artificial Intelligence]: Learning—Parameter Learning G.3 [Mathematics of Computing]: Probability and Statistics—Markov Processes G.3 [Mathematics of Computing]: Probability and Statistics—Probabilistic algorithms J.4 [Computer Applications]: Social and Behavioral Sciences—Psychology I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction techniques; I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Animation; I.5.1 [Pattern Recognition]: Models—Statistical;

Keywords: kinematics and dynamics, interaction techniques, animation, machine learning, probabilistic algorithms, gradient methods, nonlinear programming

1 Introduction

Accurate probabilistic models of interactive human motion are important for many technical applications, including computer animation, robotics, motion recognition, and for the analysis of data in neuroscience and motor control. A particular important problem in these fields is the online synthesis and simulation of human body motion, since often models have to react to inputs or behavior of the user, e.g. when real humans are immersed in an environment with virtual characters or humanoid robots. Alternatively, such movements might have to be fitted online to sensory inputs, for example when tracking people in computer vision. Such applications are difficult to address with methods for the offline synthesis of human motion, e.g. by motion capture and subsequent filtering or interpolation between stored trajectories [Bruderlin and Williams 1995; Wiley and Hahn 1997]. Ideal for many applications are representations in terms of dynamical systems, which can be directly embedded in control architectures or real-time processing loops for people tracking.

The development of such dynamical models is difficult because of the high dimensionality of human motion data (e.g. 159 dimensions for the motion capture data used in this paper). However, the

[*]e-mail:nick.taubert@klinikum.uni-tuebingen.de
[†]e-mail:andrea.christensen@uni-tuebingen.de
[‡]e-mail:dominik.endres@klinikum.uni-tuebingen.de, equal contrib. with M.A. Giese
[§]e-mail:martin.giese@uni-tuebingen.de

intrinsic dimensionality of such data is typically rather small. Thus, dimensionality reduction techniques offer a way to preserve the relevant variability in a low-dimensional latent space, and these methods are an important part of most human motion models. Gaussian process latent variable models (GP-LVM) have been proposed as a powerful approach for data modeling through dimensionality reduction [Lawrence 2004; Lawrence 2005]. The resulting low-dimensional latent representations are suitable for generalizations and extractions of characteristic features from few training examples. GP-LVMs are able to represent smooth nonlinear continuous mappings from the latent space to the data space and make it possible to model complex data manifolds in relatively low-dimensional latent spaces. Consequently, these methods have been used for the modeling of human movements in computer graphics and robotics. Since GP-LVMs are generative probabilistic models, they can serve as the building blocks of hierarchical architectures [Bishop and Tipping 1998; Lawrence and Moore 2007] to model conditional dependencies. In this paper we use them for the modeling of the kinematics of coordinated movements of multiple actors in an interactive setting.

2 Background

Statistical motion models have been used for the modeling and synthesis of human motion data (see e.g. [Grochow et al. 2004; Chai and Hodgins 2005]) and the editing of motions styles (e.g. [Li et al. 2002; Lau et al. 2009]) or style interpolation [Brand and Hertzmann 2000a; Ikemoto et al. 2009]. However, most of these techniques result in off-line models that are not suitable for the modeling of reactions of external inputs or to input from other characters in the scene.

The dominant approach for the construction of dynamic models is physics-based animation. Early approaches have been based on simple optimization principles, such as the minimization of energy ([Witkin and Kass 1988; Fang and Pollard 2003]). However, it turned out that the generation of realistic and stylized complex human body motion is difficult. Recently such physics-based animation has been combined with the use of motion capture data and learning techniques in order to improve the quality of physics-based animation (e.g. [Popovic and Witkin 1999; Safonova et al. 2004; Sulejmanpasic and Popovic 2005; Ye and Liu 2008]). A few approaches have used statistical approaches to learn dynamical systems that generate human motion [Brand and Hertzmann 2000a; Hsu et al. 2005; Wang et al. 2008], where the major problem is to develop methods that are accurate enough to capture the details of human motion, and at the same time generalize well for similar situations.

Gaussian Process latent variable models have been used in computer graphics for the modeling of kinematics and motion interpolation (e.g. [Grochow et al. 2004; Mukai and Kuriyama 2005; Ikemoto et al. 2009]) and for the learning of low-dimensional dynamical models [Ye and Liu 2010]. Here, we try specifically to devise a flexible method for learning hierarchical models of this type that can capture dependencies between multiple interacting characters.

The modeling of realistic motion with emotional style is a classical problem in computer graphics [Brand and Hertzmann 2000b; Rose et al. 1998; Unuma et al. 1995; Wang et al. 2006]. It is essential to capture the emotional movement with high accuracy, because the emotional perception depends more on the motion than on the body shape [Atkinson et al. 2004]. Even though neural activities differ when real and virtual stimuli are shown to human observers ([Perani et al. 2001; Hana et al. 2005]), the perception of emotional content is highly robust and for the most part independent of the character's body [McDonnell et al. 2008]. However, motion representations without body shapes are generally perceived as less emotionally intense than in the case where body shape is present [McDonnell et al. 2009].

Here, we develop a method that allows us to model emotional interactive motion with high accuracy in the context of real-time animation systems. For this purpose, we extend our previous hierarchical model [Taubert et al. 2011] in two directions:

- we introduce a dynamical nonlinear mapping similar to the GPDM to model (and learn) the dynamics of interactive movements, and
- we introduce style variables to modulate the emotional content and represent the personal style of the actors.

The nonlinear dynamical mapping replaces the Hidden Markov Model (HMM) which we used in [Taubert et al. 2011] for the generation of time series. The advantage of this mapping over the HMM is increased accuracy, which e.g. allows for the modeling of hand contact during handshakes in a top-down fashion. This was very hard to achieve with the HMM.

We exemplify our approach on emotional handshakes between two individuals. We validate the realism of the movements of the developed statistical model by psychophysical experiments and find that the generated patterns are hard to distinguish from natural interaction sequences.

3 Gaussian Process Latent Variable Model

Gaussian Process latent variable models are a special class of nonlinear latent variable model that map a low-dimensional latent space on the data [Bishop 1999]. Here, a high dimensional data set $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N]^T \in \Re^{N \times D}$ is represented by instances of low dimensional latent variables $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]^T \in \Re^{N \times q}$, where N is the length of the data set, D is the dimension of the data space and q is the dimension of the latent space. Data points are generated from points in the latent space $\mathbf{x}$ by applying a function $f(\mathbf{x})$ and adding isotropic Gaussian noise ε,

$$\mathbf{y} = f(\mathbf{x}) + \varepsilon, \qquad f(\mathbf{x}) \sim GP(m_Y(\mathbf{x}), k_Y(\mathbf{x}, \mathbf{x}')), \quad (1)$$

where $f(\mathbf{x})$ is drawn from a Gaussian process with mean function $m_Y(\mathbf{x})$ and kernel function $k_Y(\mathbf{x}, \mathbf{x}')$. We assume a zero mean function $m_Y(\mathbf{x}) = 0$ and use a non-linear radial basis function (RBF) kernel [Rasmussen and Williams 2008] for a high dimensionality reduction and smooth trajectories in latent space. Furthermore, the variance term for ε can be absorbed into this kernel via the noise precision γ_3:

$$k_Y(\mathbf{x}, \mathbf{x}') = \gamma_1 \exp\left(-\frac{\gamma_2}{2}|\mathbf{x} - \mathbf{x}'|^2\right) + \gamma_3^{-1}\delta_{\mathbf{x},\mathbf{x}'}, \quad (2)$$

where γ_1 is the output scale and γ_2 the inverse width of the RBF term. Let $\mathbf{K}_Y$ denote the $N \times N$ kernel covariance matrix, obtained by applying the kernel function (eqn. 2) to each pair of data points. Then the likelihood of the latent variables and the kernel parameters $\bar{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$ can be written as

$$p(\mathbf{Y}|\mathbf{X}, \bar{\gamma}) = \prod_{d=1}^{D} \mathcal{N}(\mathbf{y}_{:,d}|\mathbf{0}, \mathbf{K}_Y).$$

where $\mathbf{K}_Y$ depends on $\bar{\gamma}$ and $\mathbf{X}$. Learning the GP-LVM is accomplished by minimizing the negative log-posterior of the latent variables and the kernel parameters via scaled conjugate gradients

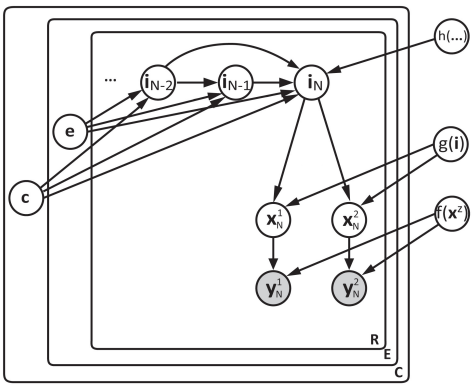

Figure 2: *Graphical model representation. A couple C consists of two actors $\in \{1;2\}$, which performed R trials of handshakes with N time steps per each trial and emotional style* $\mathbf{e}$*.* $\mathbf{y}_{\{1;2\}}$*: observed joint angles and their latent representations* $\mathbf{x}_{\{1;2\}}$ *for each actor. The* $\mathbf{x}_{\{1;2\}}$ *are mapped onto the* $\mathbf{y}_{\{1;2\}}$ *via a shared function* $f(\mathbf{x}^z)$ *which has a Gaussian process prior.* $\mathbf{i}$*: latent interaction representation, mapped onto the individual actors' latent variables by a function* $g(\mathbf{i})$ *which also has a Gaussian process prior. The dynamics of* $\mathbf{i}$ *are described by a second order dynamical model with a mapping function* $h(\mathbf{i}_{<n}, \mathbf{c}_{<n}, \mathbf{e}_{<n})$*, where* $<n$ *indicates time steps before* n*, here:* $n-1$ *and* $n-2$*. Prior mean and kernel functions have been omitted for clarity.*

(SCG) [Lawrence and Moore 2007],

$$\begin{aligned}\mathcal{L} &= -\ln(p(\mathbf{Y}|\mathbf{X},\bar{\gamma})p(\mathbf{X})p(\bar{\gamma})) \\ &= -\frac{DN}{2}\ln 2\pi - \frac{D}{2}\ln|\mathbf{K}_Y| - \frac{1}{2}\mathbf{Y}^T\mathbf{K}_Y^{-1}\mathbf{Y} \\ &\quad - \frac{1}{2}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T - \sum_j \ln\gamma_j.\end{aligned} \tag{3}$$

We choose an isotropic Gaussian prior over the latent variables and a scale-free prior over the kernel parameters, since we have no reason to believe otherwise.

4 Emotional interaction model

In order to learn the interactions between pairs of actors we devised a hierarchical model based on GP-LVMs with RBF kernels and a Gaussian Process Dynamical Model (GPDM) [Wang et al. 2007b] with linear+RBF kernel for the temporal evolution, see Figure 2. Furthermore, we introduce style variables [Wang et al. 2007a] to capture the emotional content ($\mathbf{e}$) and individual motion style of the actors ($\mathbf{c}$). Our model is comprised of three layers, which are learned jointly. This mode of learning includes both the bottom-up and top-down contributions at each hierarchy level.

4.1 Bottom Layer

In the bottom layer of our model each actor $\in \{1;2\}$ is modeled by a GP-LVM. Observed joint angles $\mathbf{Y}$ ($D = 159$) of *one individual* actor are represented by a 6-dimensional latent variable $\mathbf{X}$. The $\mathbf{Y}$ were treated as independent of actor identity, trials, emotional styles and time by sharing the mapping function in eq. (1) for the two GP-LVMs (see fig. 2)

$$\mathbf{y}^z = f(\mathbf{x}^z) + \varepsilon, \quad f(\mathbf{x}^z) \sim GP(\mathbf{0}, k_Y(\mathbf{x}^z, \mathbf{x}^{z\prime})), \quad z \in \{1,2\}. \tag{4}$$

This approach forces the GP-LVM to learn a latent representation which captures the variation w.r.t. these variables (in particular, variation across emotional style, actor style and time).

The negative joint log-probability for the two GP-LVMs in the bottom layer is

$$\mathcal{L} = -\ln(p(\mathbf{Y}^1|\mathbf{X}^1,\bar{\gamma})p(\mathbf{Y}^2|\mathbf{X}^2,\bar{\gamma})p(\mathbf{X}^1,\mathbf{X}^2)p(\bar{\gamma})). \tag{5}$$

Note that the shared mapping function (eq. (4)) is responsible for the dependency of both observations to the same kernel parameters $\bar{\gamma}$.

4.2 Interaction Layer

The latent representation $(\mathbf{X}^1, \mathbf{X}^2)$ of the joint angle data of the bottom layer model forms the 12-dimensional observation variable in the interaction layer. The mapping is represented by a GP-LVM and a 6-dimensional latent variable $\mathbf{I}$. The negative log-probability of the interaction layer model can be written as

$$\mathcal{L} = -\ln(p(\mathbf{X}^1, \mathbf{X}^2|\mathbf{I},\bar{\beta})p(\mathbf{I})p(\bar{\beta})). \tag{6}$$

We use a a non-linear RBF kernel in this layer, with parameters $\bar{\beta}$.

Furthermore, we want to model style modulation in this latent space, i.e. the characteristics of the couples of actors and their emotions. We make use of the multifactor model [Wang et al. 2007a], where a kernel matrix is constructed by an element-wise product of kernel matrices generated by different kernel functions. The element-wise multiplication of the matrices has a logical 'AND' effect in the resulting kernel matrix and is very useful for binding a specific style to pairs of latent points. For example, we enforce that points corresponding to different emotions do not correlate.

In this layer, we therefore introduce new latent variables in *1-of-K* encoding [Bishop 2007]: an emotional style (out of K possible styles) is represented by a unit vector along an axis of the latent space, e.g. $\mathbf{e} = (1,0,0,0)$ is 'neutral', $\mathbf{e} = (0,1,0,0)$ is 'angry' etc. The couple identity (and hence, individual motion style) $\mathbf{c}$ is encoded similarly. The total kernel function in the interaction model is thus:

$$\begin{aligned}&k_X([\mathbf{i},\mathbf{c},\mathbf{e}],[\mathbf{i}',\mathbf{c}',\mathbf{e}']) \\ &\quad = \quad (\mathbf{c}^T\mathbf{c}')(\mathbf{e}^T\mathbf{e}')\left\{\beta_1\exp\left(-\frac{\beta_2}{2}|\mathbf{i}-\mathbf{i}'|^2\right)\right\} + \beta_3^{-1}\delta_{\mathbf{i},\mathbf{i}'},\end{aligned}$$

where the factors $(\mathbf{c}^T\mathbf{c}')$ and $(\mathbf{e}^T\mathbf{e}')$ correlate training data points from the same couple and same emotional style, respectively. For these style variables we used a linear kernel because we would also like to interpolate between the styles (see the supplementary video for an interpolation example).

The resulting negative log-probability *with* style variables therefore has the form

$$\mathcal{L} = -\ln(p(\mathbf{X}^1, \mathbf{X}^2|\mathbf{I},\mathbf{C},\mathbf{E},\bar{\beta})p(\mathbf{I})p(\bar{\beta})), \tag{7}$$

and can be optimized w.r.t. the style variables, too.

4.3 Dynamic Layer

In the top layer we learn the temporal evolution of $\mathbf{I}$, i.e. the dynamics. To this end, we use a second-order GPDM [Wang et al. 2007b] which allows us to model e.g. dynamical dependencies on velocity and acceleration of $\mathbf{I}$. Furthermore, a GPDM can capture the non-linearities of the data without overfitting and it can learn complex dynamics from small training sets. The dynamic mapping

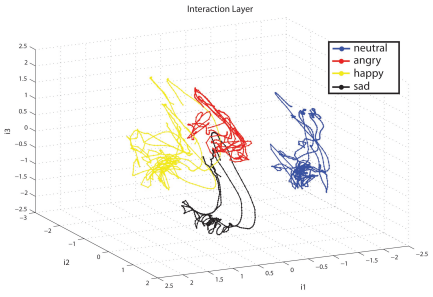

Figure 3: *Handshake trajectories in the latent spaces of the interaction layer. The separation between emotional styles is clearly visible.*

on the latent coordinates $\mathbf{I}$ is conceptually similar to the GP-LVM with the difference that the interaction variables from previous time steps $\mathbf{i}_{n-1}, \mathbf{i}_{n-2}$ are mapped onto the current $\mathbf{i}_n$,

$$\begin{aligned} \mathbf{i}_n &= h(\mathbf{i}_{n-1}, \mathbf{i}_{n-2}) + \xi, \\ h(\mathbf{i}_{n-1}, \mathbf{i}_{n-2}) &\sim GP(\mathbf{0}, k_I([\mathbf{i}_{n-1}, \mathbf{i}_{n-2}], [\mathbf{i}_{\nu-1}, \mathbf{i}_{\nu-2}])). \end{aligned} \tag{8}$$

where ξ is isotropic Gaussian noise. This kind of mapping results in a Markov chain and leads to a non-linear generalization of a hidden Markov model (HMM) [Li et al. 2000],

$$p(\mathbf{I}|\bar{\alpha}) = p(\mathbf{i}_1, \mathbf{i}_2) \prod_{n=3}^{N} p(\mathbf{i}_n|\mathbf{i}_{n-1}, \mathbf{i}_{n-2}, \bar{\alpha}) p(\bar{\alpha}), \tag{9}$$

where $\bar{\alpha}$ are the parameters of the kernel function and $p(\mathbf{i}_1, \mathbf{i}_2)$ is an isotropic Gaussian prior with zero mean and unit variance. Since this is a second-order GPDM, the kernel function depends on the last two latent positions. We use a linear+RBF kernel

$$\begin{aligned} &k_I([\mathbf{i}_{n-1}, \mathbf{i}_{n-2}], [\mathbf{i}_{\nu-1}, \mathbf{i}_{\nu-2}]) \\ &= \alpha_1 \exp\left(-\frac{\alpha_2}{2}|\mathbf{i}_{n-1} - \mathbf{i}_{\nu-1}|^2 - \frac{\alpha_3}{2}|\mathbf{i}_{n-2} - \mathbf{i}_{\nu-2}|^2\right) \\ &\quad + \alpha_4 \mathbf{i}_{n-1}^T \mathbf{i}_{\nu-1} + \alpha_5 \mathbf{i}_{n-2}^T \mathbf{i}_{\nu-2} + \alpha_6^{-1} \delta_{n,\nu}. \end{aligned}$$

Similar to the regular RBF kernel in eq. (2) is α_1 the output scale, α_2 and α_3 are the inverse width of the RBF terms. The parameters α_4, α_5 representing the output scales of the linear terms and α_6 is the precision of the distribution of the noise ξ. The RBF terms of the kernel allows us to model nonlinear dynamics in the GPDM, the linear terms enable interpolation.

Since individual and emotional style modulate the dynamics, we also extend this kernel function with factors depending on these styles:

$$\begin{aligned} &k_I([\mathbf{i}_{n-1}, \mathbf{i}_{n-2}, \mathbf{c}_{n-1}, \mathbf{c}_{n-2}, \mathbf{e}_{n-1}, \mathbf{e}_{n-2}], \\ &\qquad [\mathbf{i}_{\nu-1}, \mathbf{i}_{\nu-2}, \mathbf{c}_{\nu-1}, \mathbf{c}_{\nu-2}, \mathbf{e}_{\nu-1}, \mathbf{e}_{\nu-2}]) \\ &= (\mathbf{c}_{n-1}^T \mathbf{c}_{\nu-1} \cdot \mathbf{c}_{n-2}^T \mathbf{c}_{\nu-2})(\mathbf{e}_{n-1}^T \mathbf{e}_{\nu-1} \cdot \mathbf{e}_{n-2}^T \mathbf{e}_{\nu-2}) \\ &\times \Big\{ \alpha_1 \exp\left(-\frac{\alpha_2}{2}|\mathbf{i}_{n-1} - \mathbf{i}_{\nu-1}|^2 - \frac{\alpha_3}{2}|\mathbf{i}_{n-2} - \mathbf{i}_{\nu-2}|^2\right) \\ &\qquad + \alpha_4 \mathbf{i}_{n-1} \mathbf{i}_{\nu-1}^T + \alpha_5 \mathbf{i}_{n-2} \mathbf{i}_{\nu-2}^T \Big\} + \alpha_6^{-1} \delta_{n,\nu} \quad , \end{aligned} \tag{10}$$

to suppress correlation between points which differ in content and style in the kernel matrix of the GPDM. Let the training outputs $\mathbf{I}_{out} = [\mathbf{i}_3, ..., \mathbf{i}_N]^T$. From eq. (9) we obtain a negative log-probability of the dynamic layer

$$\begin{aligned} \mathcal{L} &= -\ln(p(\mathbf{I}|\mathbf{C}, \mathbf{E}, \bar{\alpha}) p(\bar{\alpha})) \\ &= -\frac{q(N-2)}{2} \ln 2\pi - \frac{q}{2} \ln |\mathbf{K}_I| - \frac{1}{2} \mathbf{I}_{out}^T \mathbf{K}_I^{-1} \mathbf{I}_{out} \\ &\quad - \frac{1}{2}(\mathbf{i}_1 \mathbf{i}_1^T + \mathbf{i}_2 \mathbf{i}_2^T) - \sum_j \ln \alpha_j, \end{aligned} \tag{11}$$

where $\mathbf{K}_I$ is the $(N-2) \times (N-2)$ kernel matrix constructed from the training inputs $\mathbf{I}_{in} = [[\mathbf{i}_2, ..., \mathbf{i}_{N-1}]^T, [\mathbf{i}_1, ..., \mathbf{i}_{N-2}]^T]$ and style variables $\mathbf{C}, \mathbf{E}$.

The whole model can be learned by adding the negative log-probabilities from the different layers (eqns. (5,7,11)), dropping the isotopic Gaussian priors of the bottom and interaction layer because of the top-down influence from the next higher layer:

$$\begin{aligned} \mathcal{L} = -\ln \{ &p(\mathbf{Y}_1|\mathbf{X}_1, \bar{\gamma}) p(\mathbf{Y}_2|\mathbf{X}_2, \bar{\gamma}) \\ &\times p(\mathbf{X}_1, \mathbf{X}_2|\mathbf{I}, \mathbf{C}, \mathbf{E}, \bar{\beta}) \\ &\times p(\mathbf{I}|\mathbf{C}, \mathbf{E}, \bar{\alpha}) p(\bar{\alpha}) p(\bar{\beta}) p(\bar{\gamma}) \} , \end{aligned} \tag{12}$$

and minimizing it w.r.t. the latent variables, style variables and the kernel parameters.

So far we have described the learning for one motion sequence. Learning multiple sequences proceeds along the same lines, but at the beginning of each sequence, the first two points have no predecessors and hence have the isotropic Gaussian prior from eqn. 9. In other words, the Markov chain is restarted at the beginning of each sequence.

4.4 Motion Generation

Our model is fully generative. We generate new interaction sequences in the latent space of the interaction layer by running mean prediction on the GPDM.

Given the training pairs, $\mathbf{I}_{in}, \mathbf{I}_{out}$, the training style variables, $\mathbf{C}, \mathbf{E}$, the target style variables, $\tilde{\mathbf{C}}, \tilde{\mathbf{E}}$ and the kernel function in eq. (10), the predictive distribution for generating new target outputs, $\tilde{\mathbf{I}}_{out}$, can be derived. With the (second order) Markov property (eq. 9) of the GPDM, a new target output point can be generated given the previous steps and style values,

$$\tilde{\mathbf{i}}_n \sim \mathcal{N}(\mu_I([\tilde{\mathbf{i}}_{n-1}, \tilde{\mathbf{i}}_{n-2}, \tilde{\mathbf{c}}, \tilde{\mathbf{e}}]), \sigma_I^2([\tilde{\mathbf{i}}_{n-1}, \tilde{\mathbf{i}}_{n-2}, \tilde{\mathbf{c}}, \tilde{\mathbf{e}}])\mathbf{1}).$$

where

$$\mu_I([\tilde{\mathbf{i}}_{n-1}, \tilde{\mathbf{i}}_{n-2}, \tilde{\mathbf{c}}, \tilde{\mathbf{e}}]) = k_I([\tilde{\mathbf{i}}_{n-1}, \tilde{\mathbf{i}}_{n-2}, \tilde{\mathbf{c}}, \tilde{\mathbf{e}}]) \mathbf{K}_I^{-1} \mathbf{I}_{out}, \tag{13}$$

$$\begin{aligned} &\sigma_I^2([\tilde{\mathbf{i}}_{n-1}, \tilde{\mathbf{i}}_{n-2}, \tilde{\mathbf{c}}, \tilde{\mathbf{e}}]) \\ &= k_I([\tilde{\mathbf{i}}_{n-1}, \tilde{\mathbf{i}}_{n-2}, \tilde{\mathbf{c}}, \tilde{\mathbf{e}}], [\tilde{\mathbf{i}}_{n-1}, \tilde{\mathbf{i}}_{n-2}, \tilde{\mathbf{c}}, \tilde{\mathbf{e}}]) \\ &\quad - k_I([\tilde{\mathbf{i}}_{n-1}, \tilde{\mathbf{i}}_{n-2}, \tilde{\mathbf{c}}, \tilde{\mathbf{e}}])^T \mathbf{K}_I^{-1} k_I([\tilde{\mathbf{i}}_{n-1}, \tilde{\mathbf{i}}_{n-2}, \tilde{\mathbf{c}}, \tilde{\mathbf{e}}]), \end{aligned} \tag{14}$$

and $\mathbf{1}$ is the identity matrix. The function $k_I([\tilde{\mathbf{i}}_{n-1}, \tilde{\mathbf{i}}_{n-2}, \tilde{\mathbf{c}}, \tilde{\mathbf{e}}])$ results in a vector, constructed from the test style values, the previous test inputs, all the training inputs and their associated style values,

$$\begin{aligned} &k_I([\tilde{\mathbf{i}}_{n-1}, \tilde{\mathbf{i}}_{n-2}, \tilde{\mathbf{c}}, \tilde{\mathbf{e}}]) \\ &= k_I([\tilde{\mathbf{i}}_{n-1}, \tilde{\mathbf{i}}_{n-2}, \tilde{\mathbf{c}}, \tilde{\mathbf{e}}], [\mathbf{I}_{in}, \mathbf{C}, \mathbf{E}]). \end{aligned} \tag{15}$$

For trajectory generation, we set the latent position at each time step to be the *expected* point given the previous steps,

$$\tilde{\mathbf{i}}_n = \mu_I([\tilde{\mathbf{i}}_{n-1}, \tilde{\mathbf{i}}_{n-2}, \tilde{\mathbf{c}}, \tilde{\mathbf{e}}]). \quad (16)$$

Similarly, new poses can be generated by going top-down through the hierarchy of the model to produce new target outputs in each layer, given the generated outputs of the next upper layer,

$$\tilde{\mathbf{X}}^{\{1;2\}} = \mu_{X^{\{1;2\}}}(\tilde{\mathbf{I}}),$$

$$\tilde{\mathbf{Y}}^{\{1;2\}} = \mu_{Y^{\{1;2\}}}(\tilde{\mathbf{X}}^{\{1;2\}}).$$

We applied this method to generate online four types of emotional handshakes. We processed 12 motion capture trajectories with three repetitions for each emotion. On an Intel Quad-Core 2.7GHz computer this took about 20 hours. The generated movements look quite natural (to see in the supplementary video), as was further corroborated by our psychophysical analysis below.

5 Results

5.1 Psychophysical validation

To test whether the accuracy of the developed probabilistic model is sufficient for the generation of emotionally expressive computer animations that are perceived as natural movements we conducted a 'Turing test', where human observers had to judge naturalness and emotional content of both natural and generated movements in a psychophysical study.

Nine participants (3 female, mean age: 32 years, 4 months) took part in this experiment. All were naïve with respect to the purpose of the study. Each participant was tested individually and the experimenter left the testing room when the testing started to avoid that participants would feel observed while entering their responses.

Stimuli were rendered video clips showing two gray avatars shaking hands on a light gray background. Rendering was done in a pipeline working process using commercial software from Autodesk. With Motion Builder we applied the natural and generated motion data to the avatars and rendered them in a next step with 3d studio MAX where the length of the stimuli varied between 2 and 4 seconds.

See Figure 1 and the supplementary video for an illustration of the appearance of the avatars. The appearance of the avatars was kept as simple as possible (e.g. omitting clothing, hair, and facial expressions) to keep the participants focussed on the bodily movements. The complete stimulus set consisted of the animated movements of three original handshakes and one generated handshake per emotion (neutral, happy, angry, and sad), it thus consisted of 16 different video clips, each repeated three times in randomized order. The used emotions were chosen from the set of basic emotions described by Ekman [Ekman and Friesen 1971] that have been shown previously to be conveyed by full body movements [Roether et al. 2009; Taubert et al. 2011].

Presentation of the stimuli and recording of the subjects' responses was done using Matlab and the Psychophysics toolbox [Brainard 1997].

Participants performed two tasks: an emotion classification task and a naturalness rating task. In the emotion classification task we tested whether the generated handshake movements still conveyed the intended emotion. In each trial participants observed the video clip twice before they answered the question: "'Which emotion did the avatars express?"' by pressing one out of four keys on standard keyboard marked with letters corresponding to the tested emotions. Stimuli were repeated three times in random order. The discrimination performance of the participants was determined using a contingency-table analysis ($\chi^2 - test$).

synthesized handshakes

perceived emotion	intended emotion			
	neutral	happy	angry	sad
neutral	**66.33**	7.33	11	18.33
happy	7.33	**73.78**	22	0
angry	18.44	18.33	**66.22**	3.67
sad	7.33	0	0	**77.56**
class. rate	**70.97 ± 17.6**			

natural handshakes

perceived emotion	intended emotion			
	neutral	happy	angry	sad
neutral	**78.78**	3.67	23.3	13.44
happy	18.41	**82.52**	25.74	0
angry	2.44	13.44	**50.41**	0
sad	0	0	0	**86.26**
class. rate	**74.52 ± 15.4**			

Table 1: *Classification Results. Emotion classification of synthesized (top) and natural (bottom) handshakes. Intended affect is shown in columns, percentages of subjects' (N=9) responses in rows. Bold entries on the diagonal mark rates of correct classification.* ***class. rate****: overall mean correct classification rate ± standard deviation across subjects for generated and natural movements. This standard deviation, a measure for the agreement between subjects, is similar for synthesized and natural handshakes.*

In the second part of the experiment we investigated whether the computer animations of synthesized movements were perceived as less natural than animations created from original motion capture data. For this purpose we conducted a rating task in which participants rated the naturalness of the displayed video clips. Participants were instructed that the stimuli they would see throughout the experiment could either show recorded movements of real humans interacting naturally or artificial computer generated movements. They rated the stimuli on a Likert scale ranging from 1 (very artificial) to 6 (very natural). The same stimulus set as in the emotion classification task was used for the naturalness rating task. Again, each stimulus was repeated three times and the presentation order was randomized. Since the number of natural ($4\ emotions \times 3\ prototypes \times 3\ repetitions$) and synthesized ($4\ emotions \times 1\ generated\ movement \times 3\ repetitions$) stimuli were not counterbalanced we used non-parametric statistical testing to investigate paired differences between these stimulus classes. Again, participants observed each stimulus twice before they responded by keypress.

5.2 Results

The classification results are shown in Table 1. Participants classified the expressed emotions of the handshake movements with high accuracy. Overall, participants classified the original natural handshakes in 74.52% of the trials correctly. The synthesized movements were on average classified correctly in 70.97% of the trials. These results were confirmed by a contingency-table analysis testing the null hypothesis that the variables 'intended emotion' and 'perceived emotion' were independent. We found highly significant violations of this null hypothesis (generated movements: $\chi^2 = 501.22, d.f. = 9, p < 0.001$; original/natural movements: $\chi^2 = 591.01, d.f. = 9, p < 0.001$). Comparisons of the percentages of correct classification of synthesized and natural animations

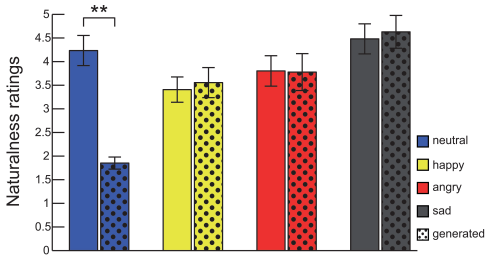

Figure 4: *Mean ratings of naturalness for original (colored) and synthesized (colored, dotted areas) handshake movements. On average the naturalness ratings are comparable for both, synthesized and original movements. Only the neutral synthesized handshake movement was rated as less natural than its original counterpart. For all other emotions participants were unable to distinguish original from synthesized movements. Error bars show standard errors, asterisks indicate significant pairwise difference (** $p< 0.01$, Wilcoxon rank-sum test).*

for the four tested emotions revealed no differences between animation classes for any of the emotions (paired Wilcoxon rank-sum test, all $p > 0.1$). Thus, the animations created from generated handshake movements still conveyed correctly the intended emotions. Furthermore, the standard deviation across subjects, a measure for between-subjects agreement, is similar for synthesized (17.6%) and natural (15.4%) handshakes.

The mean results of the naturalness ratings are shown in Figure 4. For happy, angry, and sad handshakes participants were unable to distinguish between natural and synthesized movements. The mean naturalness ratings for those emotional movements did not differ from each other as has been tested with a non-parametric paired difference test (Wilcoxon rank-sum tests, all $p < 0.16$). Only the neutral synthesized handshake movement was rated as less natural than its original counterpart ($p < 0.01$).

This surprising result for the neutral stimuli can be likely explained by the nature of the motion capture prototypes for the neutral handshakes that were used to generate the synthesized pattern. In one of the three original movements one actor made a communicative gesture after the handshake: he raised his right arm instead of taking it down to the starting position (see the supplementary video for an illustration). This gesture had effects on the naturalness ratings in two ways. On the one hand, it made this movement unique which lead to the participants' perception that this movement is very natural (median rating: 5) because they would not assume a computer to generate such individual gestures. On the other hand, this gesture also altered the generated movement that interpolates between the training examples. As a result, in the synthesized movement the right arm of the character shows a slight raising movement, which might have looked unnatural. Future experiments will use training movements with improved segmentation in order to avoid such problems.

6 Conclusion

We have devised a new online-capable method for the accurate simulation of interactive human movements. The method is based on a hierarchical probabilistic model that combines layers containing Gaussian Process Latent Variable Models, and a dynamics in the latent space that is implemented with a Gaussian Process Dynamical Model. In addition, the proposed method is able to model the emotional style by appropriate additional dimensions in the latent space.

The results of our psychophysical experiments confirm the suitability of the derived probabilistic model for the generation of emotionally expressive movements. The animations of those synthesized movements conveyed the correct information about style changes. Furthermore, when the motion capture examples used to train the system were appropriately segmented, the resulting online generated patterns were perceived as almost as natural as the original motion capture data. This highlights the importance of a robust automatic segmentation process, for which we will employ our own previously developed Bayesian segmentation approach [Endres et al. 2011].

Future work will extend the modular architecture of our model to build algorithms for continuous interpolation of emotional styles (an early example of this is contained in the supplementary video). In addition, we will embed the learned model in an online animation pipeline in order to study interactions between real and virtual humans.

Acknowledgments

We thank E.M.J. Huis in 't Veld for help with the movement recordings and all our participants for taking part in the psychophysical study. This work was supported by EU projects FP7-ICT-215866 SEARISE, FP7-249858-TP3 TANGO, FP7-ICT-248311 AMARSi, the DFG (GI 305/4-1, and GZ: KA 1258/15-1), and the German Federal Ministry of Education and Research (BMBF; FKZ: 01GQ1002).

References

ATKINSON, A. P., DITTRICH, W. H., GEMMELL, A. J., AND YOUNG, A. W. 2004. Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception. 33*, 6 (June), 717–746.

BISHOP, C. M., AND TIPPING, M. E. 1998. A hierarchical latent variable model for data visualization. *IEEE Transactions on Pattern Analysis and Machine Intelligence 20(3)*, 281–293.

BISHOP, C. M. 1999. Latent variable models. In *Jordan, M. I. (Ed.), Learning in Graphical Models*. MIT Press, 371–403.

BISHOP, C. M. 2007. *Pattern Recognition and Machine Learning*. Springer.

BRAINARD, D. 1997. The psychophysics toolbox. *Spatial Vision*, 10, 433:436.

BRAND, M., AND HERTZMANN, A. 2000. Style machines. In *SIGGRAPH*, 183–192.

BRAND, M., AND HERTZMANN, A. 2000. Style machines. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, SIGGRAPH '00, 183–192.

BRUDERLIN, A., AND WILLIAMS, L. 1995. Motion signal processing. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, ACM, New York, NY, USA, SIGGRAPH '95, 97–104.

CHAI, J., AND HODGINS, J. K. 2005. Performance animation from low-dimensional control signals. *ACM Transactions on Graphics 24*, 3, 686–696.

EKMAN, P., AND FRIESEN, W. V. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology 17*, 124–129.

ENDRES, D., CHRISTENSEN, A., OMLOR, L., AND GIESE, M. A. 2011. Emulating human observers with Bayesian binning: segmentation of action streams. *ACM Transactions on Applied Perception (TAP) 8*, 3, 16:1–12. DOI: 10.1145/2010325.2010326.

FANG, A. C., AND POLLARD, N. S. 2003. Efficient synthesis of physically valid human motion. *ACM Transactions on Graphics 22*, 3, 417–426.

GROCHOW, K., MARTIN, S. L., HERTZMANN, A., AND POPOVIC, Z. 2004. Style-based inverse kinematics. *ACM Transactions on Graphics 23*, 3, 522–531.

HANA, S., JIANGA, Y., HUMPHREYSC, G. W., ZHOUD, T., AND CAID, P. 2005. Distinct neural substrates for the perception of real and virtual visual worlds. *NeuroImage 24*, 928 935.

HSU, E., PULLI, K., AND POPOVIC, J. 2005. Style translation for human motion. *ACM Transactions on Graphics 24*, 3, 1082–1089.

IKEMOTO, L., ARIKAN, O., AND FORSYTH, D. A. 2009. Generalizing motion edits with gaussian processes. *ACM Transactions on Graphics 28*, 1.

LAU, M., BAR-JOSEPH, Z., AND KUFFNER, J. 2009. Modeling spatial and temporal variation in motion data. *ACM Transactions on Graphics 28*, 5.

LAWRENCE, N. D., AND MOORE, A. J. 2007. Hierarchical gaussian process latent variable models. In *Proceedings of the International Conference in Machine Learning*, Omnipress, 481–488.

LAWRENCE, N. D. 2004. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in Neural Information Processing Systems*, MIT Press, 329–336.

LAWRENCE, N. D. 2005. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research 6*, 1783–1816.

LI, X., PARIZEAU, M., AND PLAMONDON, R. 2000. Training hidden markov models with multiple observations-a combinatorial method. *IEEE Trans. Pattern Anal. Mach. Intell. 22*, 4, 371–377.

LI, Y., WANG, T., AND SHUM, H.-Y. 2002. Motion texture: a two-level statistical model for character motion synthesis. In *SIGGRAPH*, 465–472.

MCDONNELL, R., JÖRG, S., MCHUGH, J., NEWELL, F., AND O'SULLIVAN, C. 2008. Evaluating the emotional content of human motions on real and virtual characters. In *Proceedings of the 5th symposium on Applied perception in graphics and visualization*, ACM, New York, NY, USA, APGV '08, 67–74.

MCDONNELL, R., JÖRG, S., MCHUGH, J., NEWELL, F. N., AND O'SULLIVAN, C. 2009. Investigating the role of body shape on the perception of emotion. *ACM Trans. Appl. Percept. 6*, 3 (Sept.), 14:1–14:11.

MUKAI, T., AND KURIYAMA, S. 2005. Geostatistical motion interpolation. *ACM Transactions on Graphics 24*, 3, 1062–1070.

PERANI, D., FAZIO, F., BORGHESE, N., TETTAMANTI, M., FERRARI, S., DECETY, J., AND GILARDI, M. 2001. Different brain correlates for watching real and virtual hand actions. *NeuroImage 14*, 749 758.

POPOVIC, Z., AND WITKIN, A. P. 1999. Physically based motion transformation. In *SIGGRAPH*, 11–20.

RASMUSSEN, C. E., AND WILLIAMS, C. K. I. 2008. Gaussian processes for machine learning. *Journal of the American Statistical Association 103*, 429–429.

ROETHER, C., OMLOR, L., CHRISTENSEN, A., AND GIESE, M. A. 2009. Critical features for the perception of emotion from gait. *Journal of Vision 6*, 10.1167/9.6.15.

ROSE, C., BODENHEIMER, B., AND COHEN, M. F. 1998. Verbs and adverbs: Multidimensional motion interpolation using radial basis functions. *IEEE Computer Graphics and Applications 18*, 32–40.

SAFONOVA, A., HODGINS, J. K., AND POLLARD, N. S. 2004. Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces. *ACM Transactions on Graphics 23*, 3, 514–521.

SULEJMANPASIC, A., AND POPOVIC, J. 2005. Adaptation of performed ballistic motion. *ACM Transactions on Graphics 24*, 1, 165–179.

TAUBERT, N., ENDRES, D., CHRISTENSEN, A., AND GIESE, M. A. 2011. Shaking hands in latent space: modeling emotional interactions with gaussian process latent variable models. In *KI 2011: Advances in Artificial Intelligence, LNAI*, S. Edelkamp and J. Bach, Eds. Springer, 330–334.

UNUMA, M., ANJYO, K., AND TAKEUCHI, R. 1995. Fourier principles for emotion-based human figure animation. In *In Proceedings of Computer Graphics*, SIGGRAPH'95.

WANG, Y., LIU, Z.-Q., AND ZHOU, L.-Z. 2006. Learning style-directed dynamics of human motion for automatic motion synthesis. In *Proc. IEEE International Conference on Systems, Man, and Cybernetics*.

WANG, J. M., FLEET, D. J., AND HERTZMANN, A. 2007. Multifactor gaussian process models for style-content separation. In *ICML*.

WANG, J. M., FLEET, D. J., MEMBER, S., AND HERTZMANN, A. 2007. Gaussian process dynamical models for human motion. *IEEE Trans. Pattern Anal. Machine Intell.*

WANG, J. M., FLEET, D. J., AND HERTZMANN, A. 2008. Gaussian process dynamical models for human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*, 2, 283–298.

WILEY, D., AND HAHN, J. 1997. Interpolation synthesis for articulated figure motion. In *Virtual Reality Annual International Symposium*, IEEE 1997, 156–160.

WITKIN, A. P., AND KASS, M. 1988. Spacetime constraints. In *SIGGRAPH*, 159–168.

YE, Y., AND LIU, C. K. 2008. Animating responsive characters with dynamic constraints in near-unactuated coordinates. *ACM Transactions on Graphics 27*, 5, 112.

YE, Y., AND LIU, C. K. 2010. Synthesis of responsive motion using a dynamic model. *Computer Graphics Forum 29*, 2, 555–562.