Shaking Hands in Latent Space – Modeling Emotional Interactions with Gaussian Process Latent Variable Models.

Conference Paper · January 2011

Source: DBLP

CITATIONS
6

READS
96

4 authors:

Nick Taubert
University of Tübingen
42 PUBLICATIONS **172** CITATIONS
SEE PROFILE

Dominik Endres
Philipps University of Marburg
96 PUBLICATIONS **1,913** CITATIONS
SEE PROFILE

Andrea Christensen
University of Tübingen
32 PUBLICATIONS **822** CITATIONS
SEE PROFILE

Martin A. Giese
University of Tübingen
463 PUBLICATIONS **9,213** CITATIONS
SEE PROFILE

Shaking Hands in Latent Space:
Modeling Emotional Interactions with Gaussian Process Latent Variable Models

Nick Taubert, Dominik Endres, Andrea Christensen, Martin A. Giese

{nick.taubert,dominik.endres}@klinikum.uni-tuebingen.de
{andrea.christensen,martin.giese}@uni-tuebingen.de
Theoretical Sensomotorics, Cognitive Neurology, University Clinic Tübingen,
CIN, HIH and University of Tübingen, Frondsbergstr. 23, 72070 Tübingen, Germany

Abstract. We present an approach for the generative modeling of human interactions with emotional style variations. We employ a hierarchical Gaussian process latent variable model (GP-LVM) to map motion capture data of handshakes into a space of low dimensionality. The dynamics of the handshakes in this low dimensional space are then learned by a standard hidden Markov model, which also encodes the emotional style variation. To assess the quality of generated and rendered handshakes, we asked human observers to rate them for realism and emotional content. We found that generated and natural handshakes are virtually indistinguishable, proving the accuracy of the learned generative model.

Cite as:
Taubert N., Endres D., Christensen C. and Giese M.A. (2011). Shaking Hands in Latent Space: Modeling Emotional Interactions with Gaussian Process Latent Variable Models. In *KI2011: Advances in Artificial Intelligence, LNAI 7006, Springer*, 330-334.
The original publication can be found at www.springerlink.com,
DOI: 10.1007/978-3-642-24455-1.

1 Introduction

Accurate probabilistic models of interactive human motion are important for many applications, including computer animation, motion recognition and emotional feature analysis. Gaussian processes provide a powerful framework for the modeling of human motion since they permit to approximate complex trajectories with high accuracy, at the same time guaranteeing successful generalization from few training examples [11]. Gaussian process latent variable models (GP-LVM) have been proposed for the modeling of the motion of individual humans [3]. The resulting low-dimensional representations are suitable for feature extraction and the modeling of style. The GP-LVM can also be extended towards

hierarchical architectures [4], making it possible to model the conditional dependencies induced by the coordinated movements of multiple actors / agents in an interactive setting.

The modeling of emotional styles is a classical problem in computer graphics, see e.g. [2,9,10]. The modeling of interactions between multiple characters has often been based on physical interaction models [6]. In this paper we take an approach from machine learning and try to learn the joint statistics of the interactive movements (represented as joint angles from motion capture) using a hierarchical Bayesian approach. This approach was applied to emotional handshakes between two individuals. We validated the realism of the movements of the developed statistical model by psychophysical experiments and find that the generated patterns are virtually indistinguishable from natural interactions.

2 Model

In order to learn the interactions between pairs of actors we devised a hierarchical model based on GP-LVMs with radial basis function (RBF) kernels [7] and hidden Markov models (HMM) [5]. Our model is comprised of three layers (see fig. 1, left), which were learned in a layer-wise bottom-up fashion:

GP-LVM-single: the bottom layer. Observed joint angles $\mathbf{y}$ of *one individual* actor were mapped onto a 3-dimensional latent variable $\mathbf{x}$. The $\mathbf{y}$ were treated as i.i.d. across actors, trials, emotional styles and time. This approach forced the GP-LVM to learn a latent representation which captures the variation w.r.t. these variables (in particular, variation across emotional style and time).

GP-LVM-interaction: the interaction layer. For pairs of joint angles of interacting actors (say, actors 1 and 2), we computed the corresponding latent representation $(\mathbf{x}_1, \mathbf{x}_2)$ with the learned bottom layer model. This latent representation $(\mathbf{x}_1, \mathbf{x}_2)$ forms the 6-dimensional observation variable in the interaction layer, which maps $(\mathbf{x}_1, \mathbf{x}_2)$ onto a 3-dimensional latent variable $\mathbf{i}$. The mapping is represented by a GP-LVM. Similar to the bottom layer, the $(\mathbf{x}_1, \mathbf{x}_2)$ were treated as i.i.d. across *pairs* of actors, trials and time, sorted by emotional styles. Consequently, $\mathbf{i}$ is a latent representation of the *interaction* which captures the variability w.r.t. emotional style and time.

Latent variables and kernel parameters were optimized with scaled conjugate gradients (SCG) [4].

HMM-dynamic: the top layer. Left-to-right HMMs (7 states, Gaussian observation models, initial mean 0 and diagonal covariance 0.3) learned the temporal evolution of $\mathbf{i}$, i.e. the dynamic. We trained one HMM per emotional style, across all pairs of actors and their trials.

Our model is fully generative. Since we learned one HMM per emotional style in layer **HMM-dynamic**, we can switch between styles simply by choosing the appropriate HMM and GP-LVM-interaction. We generate new interaction sequences in the latent space of GP-LVM-interaction by running the HMM forward, to compute a state-probability weighted mean from the means of the emission models. This weighted mean generates smooth input sequences in the

latent space of the interaction layer. Using the learned probabilistic generative model we then back-project [7] to the joint angles at the lowest level of the hierarchy.

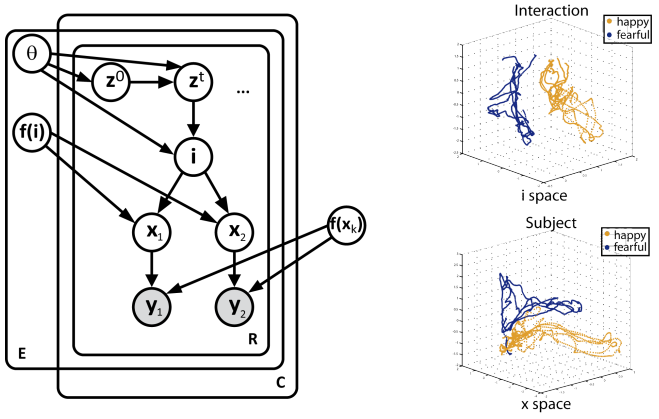

Fig. 1. *Left*: Graphical model representation. A couple C consists of two actors $\in \{1; 2\}$, which performed R trials of handshakes with emotional style E. $y_{1,2}$: observed joint angles and their latent representations $x_{1,2}$ for each actor. The $x_{1,2}$ are mapped onto the $y_{1,2}$ via a function $f(x)$ which has a Gaussian process prior. i: latent interaction representation, mapped onto the individual actors' latent variables by a function $f(i)$ which also has a Gaussian process prior. The dynamics of i are described by a HMM with hidden states z_t and parameters Θ. *Right*: Handshake trajectories in the latent spaces of layer **GP-LVM-single** (bottom panel) and layer **GP-LVM-interaction** (upper panel). The separation between emotional styles (happy and fearful) is clearly visible. For details, see section 2.

3 Results and Conclusion

We learned emotional handshakes represented as joint angles (in radians) derived from motion capture data. Movements were executed three times with five different emotional styles for each couple: *neutral*, *fearful*, *happy*, *angry* and *sad*. We fitted a commercial character model with 38 joint angles to the data from each subject, thus obtaining the training data for the model.

To illustrate that we succeeded in learning latent representations which encode emotional style variations, see Fig. 1, right. The upper and lower panels show the trajectories of one actor for a fearful and a happy handshake in the latent representations of layers **GP-LVM-single** and **GP-LVM-interaction**, respectively. The two trajectories are clearly separated.

We designed a psychophysical study to test whether the accuracy of the developed probabilistic model is good enough for computer animation. Nine

participants (4 female, mean age: 31 years, 7 months) took part in the first experiment. All were naïve with respect to the purpose of the study.

Rendered video clips showed two uniform, androgynous gray avatars without facial expressions to keep the focus on the bodily movements. In each trial two videos representing the same emotion were displayed side by side on a computer screen. The two videos in each trial could either both display natural handshakes, or a natural and a generated movement. Viewing time was not restricted, allowing participants to search for subtle differences between the animations. Participants classified the *emotion* of the stimulus and the *naturalism* of both displayed movie clips.

Participants failed to reliably assess naturalism, see table 1, top. They had a strong bias to classify every movement as being natural, which is indicated by a high hit rate (correctly identified natural movements) of 67.5%, a high false alarm rate (generated movements classified as natural) of 42.5% and a low sensitivity measure (d' = 0.64).

	animation	
judgment	generated movement	natural movement
not natural	**57.5**	32.5
natural	42.5	**67.5**

	intended emotion									
judgment	neutral	sad	happy	fearful	angry	neutral	sad	happy	fearful	angry
neutral	**70.83**	4.17	12.5	0	4.17	**91.67**	0	0	0	0
sad	8.33	**95.83**	0	4.17	0	4.17	**100**	0	12.5	0
happy	4.17	0	**87.5**	0	0	4.17	0	**91.67**	0	37.5
fearful	16.67	0	0	**91.67**	12.5	0	0	0	**87.5**	0
angry	0	0	0	4.17	**83.33**	0	0	8.33	0	**62.5**
class. rate	**85.83**					**86.67**				

Table 1. Classification Results. *Upper part*: discrimination performance for natural versus synthesized handshake movements. Columns represent the original movement on which the animation based, rows show judgments of the participants (N=9) in percent. *Lower part*: emotion classification of natural and synthesized handshakes separately. Intended affect is shown in columns, percentages of subjects' (N=12) responses in rows. Bold entries on the diagonal mark rates of correct classification. *class. rate* overall mean correct classification rates for generated and natural movements.

In contrast, participants classified the expressed emotions of the handshake movements with very high accuracy. Confusions occurred mainly for emotions that are comparable in their motion energy; i.e. 'angry' and 'happy' or 'sad' and 'fearful'. These results confirm that emotions can be reliably detected from bodily movements and are in line with findings for emotional gait [8,1].

Displaying animations of natural and generated handshakes side by side allows participants to directly match stimuli features. Specifically, if there are

differences in the naturalism between those video clips, it would be more likely that the observer detects them compared to a sequential display of the movies. Since participants were not even able to discriminate natural movements from completely generated ones in a task were they had the opportunity of a direct comparison, it may be concluded that the generated movements look indeed very natural. On the other hand, this form of presenting the videos next to each other has the disadvantage that the affect classification might be confounded. The perception of expressed emotion could be mainly driven by one of the two videos that are simultaneously presented. To validate that the generated movements are perceived as emotional as the natural ones we conducted a second control experiment. In this experiment participants observed the identical animations as described above but now sequentially.

The classification rates of twelve participants (6 female, mean age 29 years, 9 month) are depicted in table 1, bottom. Both kinds of animations conveyed enough information about the emotional context of the handshake to recognize the intended affects more than 85 % of all trials. The recognizability was highly significant for both video types, as revealed in a contingency-table analysis testing the null hypothesis that the variables 'intended emotion' and 'perceived emotion' are independent (generated movements: $\chi^2 = 1398, d.f. = 16, p < 0.001$; natural movements: $\chi^2 = 1488, d.f. = 16, p < 0.001$). Further, the percentages of correct classification did not differ between synthesized and natural animations (paired t-test, $t_{59} = -0.163, p = 0.87$).

To conclude, the results of these psychophysical experiments demonstrate clearly that the generated movements are not distinguishable by human observers from animations derived from original motion capture data. Furthermore, the fact that emotions were well identified shows that the animations were sufficient to convey very subtle information about style changes.

For the future, we plan to exploit the modular architecture of our model for feature analysis and to build algorithms that automatically generate emotional interactive action sequences.

Acknowledgements: this work was supported by EU projects FP7-ICT-215866 SEARISE, FP7-249858-TP3 TANGO, FP7-ICT-248311 AMARSi and the DFG. We thank E. Huis in 't Veld for help with the movement recordings.

References

1. Atkinson, A.P., Dittrich, W., Gemmell, A., Young, A.: Emotion perception from dynamic and static body expressions in point-light and full-light displays. Perception (33), 717–746 (2004)
2. Brand, M., Hertzmann, A.: Style machines. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques. pp. 183–192. SIGGRAPH '00, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA (2000), http://dx.doi.org/10.1145/344779.344865
3. Lawrence, N.D.: Probabilistic non-linear principal component analysis with gaussian process latent variable models. Journal of Machine Learning Research 6, 1783–1816 (2005)

4. Lawrence, N.D., Moore, A.J.: Hierarchical gaussian process latent variable models. In: Proceedings of the International Conference in Machine Learning. pp. 481–488. Omnipress (2007)
5. Li, X., Parizeau, M., Plamondon, R.: Training hidden markov models with multiple observations-a combinatorial method. IEEE Trans. Pattern Anal. Mach. Intell. 22(4), 371–377 (2000)
6. Nguyeny, N., Wheatland, N., Brown, D., Parise, B., Liu, C.K., Zordan, V.B.: Performance capture with physical interaction. In: Symposium on Computer Animation. pp. 189–195 (2010)
7. Rasmussen, C.E., Williams, C.K.I.: Gaussian processes for machine learning. Journal of the American Statistical Association 103, 429–429 (2008)
8. Roether, C., Omlor, L., Christensen, A., Giese, M.A.: Critical features for the perception of emotion from gait. Journal of Vision 6, 10.1167/9.6.15 (2009)
9. Rose, C., Bodenheimer, B., Cohen, M.F.: Verbs and adverbs: Multidimensional motion interpolation using radial basis functions. IEEE Computer Graphics and Applications 18, 32–40 (1998)
10. Unuma, M., Anjyo, K., Takeuchi, R.: Fourier principles for emotion-based human figure animation. In: In Proceedings of Computer Graphics. SIGGRAPH'95 (1995)
11. Wang, J.M., Fleet, D.J., Member, S., Hertzmann, A.: Gaussian process dynamical models for human motion. IEEE Trans. Pattern Anal. Machine Intell (2007)