A Virtual Reality Setup for Controllable, Stylized Real-Time Interactions between Humans and Avatars with Sparse Gaussian Process Dynamical Models

Nick Taubert[*] Martin Löffler[†] Nicolas Ludolph[‡] Andrea Christensen[§] Dominik Endres[¶] Martin A. Giese[‖]
Section Computational Sensomotorics, Department of Cognitive Neurology, University Clinic Tübingen,
CIN, HIH and University of Tübingen, Otfried-Müller-Str. 25, 72076 Tübingen, Germany

Abstract

Building on our previous work [Taubert et al. 2012], we present an approach for real-time interaction between a real human and an avatar. We generate reactive motions by a dynamical extension of a hierarchical Gaussian process latent variable model, including latent dimensions for emotional style variation and target positions. This allows the avatar to produce accurate reactive motions to the human. To validate our approach, we developed a real-time application where an avatar and a human actor engage in emotional 'high fives'. Furthermore, we show preliminary results indicating that humans do perceive emotions more accurately when engaging in interaction as opposed to passive observation.

CR Categories: 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: real-time system, reactive avatar, style control, Gaussian Process Dynamical models

1 Introduction

The probabilistic modeling of stylized interactive human motion is useful in technical fields like computer animation, but also for conducting experiments in psychology and neuroscience. For example, it has been shown that personal involvement established by direct eye contact with an emotional second agent can alter neurophysiological responses [Wicker et al. 2003]. However, participants in the experiment of Wicker and colleagues remained quite passive. We implemented a novel virtual reality environment in which participants are interacting in an emotional scene with a virtual avatar. Besides addressing basic research questions about perception in an interactive setting, such environments might also be interesting for studying changes of emotional interaction in specific patient groups such as autism, [Philip et al. 2010], schizophrenia [Chan et al. 2010] or various affective diseases [Schaefer et al. 2010], going beyond simple passive recognition studies.

Many existing algorithms for the real-time synthesis of emotional movements, e.g. based on motion morphing or simple generative models, lack the capability to react in real-time to movements of an interaction partner. This is difficult to accomplish with offline synthesis methods, e.g. by motion capture and subsequent filtering or interpolation between stored trajectories [Bruderlin and Williams 1995; Wiley and Hahn 1997]. Instead we propose a new algorithm which offers the possibility to represent fully stylized motion as part of a dynamical control system.

The intrinsic dimensionality of human motion data is rather small compared to the typical dimensionality of the raw motion capture data. Therefore, motion modeling often begins with a dimensionality reduction step (e.g. principal component analysis) to enable generalization to novel motion from small amounts of data. To capture nonlinearities in the data, we use Gaussian process latent variable models (GP-LVM) which are a state-of-the-art approach for dimensionality reduction in continuous spaces, and can serve as the building blocks of hierarchical architectures [Lawrence and Moore 2007]. In this paper we use them for the modeling of the kinematics of coordinated movements of two actors in an interactive setting.

2 Background

Physics-based animation is the most dominant approach for the construction of dynamical models in computer graphics. While early approaches simply minimized the energy [Witkin and Kass 1988] and could not synthesize stylized motion well, later approached combined motion capture data and learning techniques to address this shortcoming (e.g. [Safonova et al. 2004; Ye and Liu 2008]). However, many of these techniques are not suitable for modeling reactive stylized movements.

Statistical motion models have been used for the modeling and synthesis of human motion data, see e.g. [Chai and Hodgins 2005], and the editing of motions styles (e.g. [Lau et al. 2009]) or style interpolation [Brand and Hertzmann 2000a; Ikemoto et al. 2009]. Some of these models learn dynamical systems ([Brand and Hertzmann 2000a; Wang et al. 2008; Ye and Liu 2010]). However, most of these approaches result either in non-reactive off-line models, or are not accurate enough for stylized motion.

The modeling of emotional styles is a classical problem in computer graphics [Brand and Hertzmann 2000b; Rose et al. 1998]. GP-LVMs have been used for the modeling of kinematics and motion interpolation (e.g.[Grochow et al. 2004; Mukai and Kuriyama 2005; Ikemoto et al. 2009]), inverse kinematics [Levine et al. 2012] and style interpolation [Taubert et al. 2012]. Below, we present a method that allows to model emotional style for real-time animation. To this end, we extend our previous hierarchical model [Taubert et al. 2012] in three directions:

1. for real-time capability, we implement fast approximate inference of the latent variables via back-projection with a Gaussian process,
2. we introduce a dynamical mapping similar to the Gaussian process dynamical model (GPDM) [Wang et al. 2008] to

[*] nick.taubert@klinikum.uni-tuebingen.de
[†] martin.loeffler@uni-tuebingen.de
[‡] nicolas.ludolph@student.uni-tuebingen.de
[§] andrea.christensen@klinikum.uni-tuebingen,de
[¶] dominik.endres@klinikum.uni-tuebigen.de, equal. contrib. with M.A.Giese
[‖] martin.giese@uni-tuebingen.de

SAP 2013, August 22 – 23, 2013, Dublin, Ireland.

ACM 978-1-4503-2262-1/13/08 $15.00

model and learn the dynamics of interactive movements, and

3. we introduce style variables and back constraints to factorize the emotional content from the motion.

We illustrate our approach on emotional 'high five' expressions between a human actor and an avatar, and report preliminary tests of perceptual differences between interactive and non-interactive experiences of whole-body expressions. The preliminary psychophysical results indicate that the perception of emotional movements is indeed different when the observer is actively interacting with an emotional agent, rather than if such an agent is passively observed. This supports the importance of the development of models for stylized interactive motion.

3 Interactive system setup

We developed a pipeline from a Vicon motion capture system to the game engine Ogre 3D (see fig. 1). The head and arm positions of the human actor were recorded at a sampling rate of 120 Hz using a VICON MX motion capture system with 10 cameras and 5 reflecting markers on the head and 11 reflecting markers on the right arm. The markers are identified in real-time using Vicon Nexus software, which sends marker positions to the marker tracker module. This module uses a Bayesian tracking model, similar to a Kalman Filter, to infer the positions of occluded markers. The joint angles of a human actor's arm are computed from the complete marker set, and sent to the synchronization server. The purpose of this server is to synchronize the joint angles computed by GPDM model for the avatar (see section 4 for details) with the data-stream of the human arm. Both are then sent on to the Ogre engine for rendering, which runs with 68 fps on average in our setup. The rendered scene is displayed on a head mounted display worn by the human participant.

4 Emotional interaction model

We learn interactions between a couple of actors and then replace one actor by the motion capture data of a human user during real time interaction. To this end, we devised a hierarchical model based on GP-LVMs with radial basis function (RBF) and linear kernels associated with a nonlinear auto-regression model [Wang et al. 2008] for the temporal evolution. This leads to a hierarchical Gaussian process dynamical model (GPDM), see Fig. 2. To control the strength of an emotional style and the characteristics of the reacting avatar during the generation of an interactive sequence we introduce style variables [Wang et al. 2007] to capture and to separate the emotional content $(\mathbf{e})$ from the basic motion. Our model is comprised of three layers, which were learned jointly. This mode of learning includes both the bottom-up and top-down contributions at each hierarchy level.

To learn this model from large data sets we used sparse approximations techniques, employing the 'deterministic training conditional' [Lawrence 2007]. It can be shown (see [Lawrence 2007] for details) that the computational cost per learning step is then effectively linear in the number of data-points, as opposed to cubic for the exact solution.

Bottom Layer: each actor $\in \{1; 2\}$ is represented by a GP-LVM mapping the latent kinematic variables $\mathbf{x}^i$ of each actor onto the observed marker positions or joint angles $\mathbf{y}^i$. Actor 1 is the computer-generated avatar, actor 2 is the human. We denote the dimensionality of the observed spaces by D, that of the latent spaces by q.

Model of the reacting avatar. The latent variable $\mathbf{X}^1$ $(q = 2)$ of the avatar maps onto observed joint angles $\mathbf{Y}^1$ $(D = 159)$. We represent the $\mathbf{Y}^1$ across trials, emotional styles and time as points in the same latent space. Additional dimensions of this latent space are reserved for the emotional style, i.e. we separate style from motion content. For this purpose we engineer a prior that promotes factorization of the latent variables into motion dimensions and style dimensions during learning via back constraints [Lawrence et al. 2006] expressing the approximately periodic nature of the movements (i.e. the avatar begins and ends a trial in approximately the same pose). Furthermore we make use of the multifactor model [Wang et al. 2007], where a total kernel function for the GP-LVM is constructed by a product of different kernel functions for motion and style $\mathbf{e}$. This multiplication has a logical 'AND' effect in the resulting kernel matrix and is very useful for binding a specific style to pairs of latent points. Style is represented in *1-of-K* encoding [Bishop 2007]: each emotion is a unit vector along an axis of the style dimensions of the latent space, e.g. $\mathbf{e} = (1, 0, 0, 0)$ is 'neutral', $\mathbf{e} = (0, 1, 0, 0)$ is 'angry' etc. Interpolating between emotions is accomplished by computing weighted sums of these vectors.

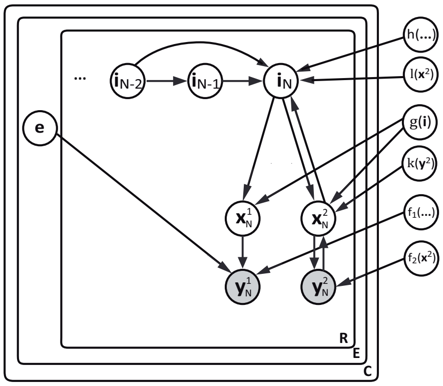

Figure 2: *Graphical model representation of our emotional interaction model. A couple* $\mathbf{c}$ *consists of two actors* $\in \{1; 2\}$*, which performed* R *trials of 'high-fives' with* N *time steps per each trial and emotional style* $\mathbf{e}$*.* $\mathbf{y}^1$*: observed joint angles of the avatar and their latent representations* $\mathbf{x}^1$ *and* $\mathbf{e}$ *(emotional style);* $\mathbf{y}^2$*: observed marker positions of the human actor and their latent representation* $\mathbf{x}^2$*. The* $\mathbf{x}^{\{1;2\}}$ *are mapped onto the* $\mathbf{y}^{\{1;2\}}$ *via function* $f_1(\mathbf{x}^1, \mathbf{e})$ *and* $f_2(\mathbf{x}^2)$ *which have a Gaussian process prior.* $\mathbf{i}$*: latent interaction representation, mapped onto 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* $h(\mathbf{i}_{n-1}, \mathbf{i}_{n-2})$*. For actor 2 (human) we also learn back projections from observation space to latent space with mapping functions* $r(\mathbf{y}^2)$ *and* $s(\mathbf{x}^2)$*, also with Gaussian process priors, to facilitate fast joint angle reconstruction of actor 1. Prior mean and kernel functions have been omitted for clarity.*

Modeling the human hand position. For the human actor, we only learn a model of the hand position, which is used to drive the interactive motion via inference of the corresponding point in the latent space. The latent variable $\mathbf{X}^2$ $(q = 2)$ of actor 2 maps onto observed marker positions of the hand $\mathbf{Y}^2$ $(D = 6)$. After learning the corresponding GP-LVM, we also learn a back projection: a Gaussian process that maps the $\mathbf{Y}^2$ onto the corresponding $\mathbf{X}^2$ for fast inference, to keep the model real-time capable.

Interaction Layer: the interaction of both actors is represented by a GP-LVM and a 3-dimensional latent variable $\mathbf{I}$ which maps onto the latent representation $(\mathbf{X}^1, \mathbf{X}^2)$ of the bottom layer model. Here, too, we learn a back projection from $\mathbf{X}^2$ to $\mathbf{I}$ for fast inference, as indicated by the upward arrow in 2.

Dynamic Layer: in the top layer the temporal evolution of $\mathbf{I}$ is represented with a second-order GPDM [Wang et al. 2008], a nonlinear auto-regressive model which allows us to model the dynam-

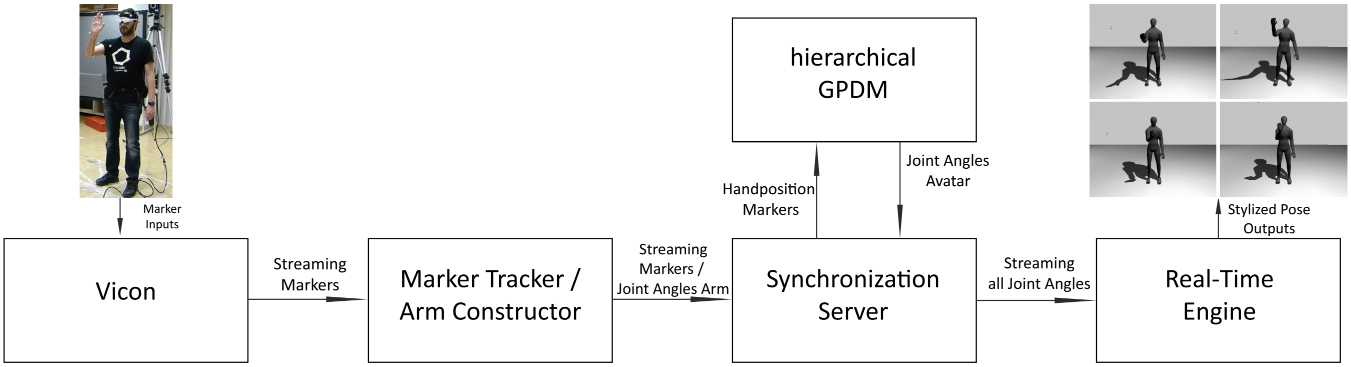

Figure 1: *The interactive system layout. From left to right: human marker positions are recorded with a* ***Vicon*** *system. A* ***marker tracker*** *infers the positions of occluded markers, the human* ***arm*** *is constructed from the full marker set. A* ***synchronization server*** *streams the arm representation and the interactive response motion generated by the* ***hierarchical GPDM*** *to the Ogre 3D* ***Real-Time Engine*** *for rendering. The resulting movie is displayed on a head-mounted display worn by the acting human (far left).*

ical dependencies on velocity and acceleration of $\mathbf{I}$. The current time point $\mathbf{i}_n$ is represented by the previous time steps $\mathbf{i}_{n-1}, \mathbf{i}_{n-2}$. This mapping generalizes a hidden Markov model (HMM) [Li et al. 2000] in a non-linear way. Here we used a RBF+linear kernel depending on the last two latent positions, to model non-linear dependencies near the data and enable extrapolation beyond the learned data.

Interaction and Motion Generation: we initiate the generation of a new interactive motion at time-step N with the GPDM by providing hand position data for the human actor $\mathbf{y}^2_N$. We then use the back projection and infer the latent variables $\mathbf{x}^2_N$. The current interaction layer representation $\mathbf{i}_N$ is computed as the expected posterior combining the prediction obtained from the two previous time-steps $\mathbf{i}_{N-2}, \mathbf{i}_{N-1}$ via the dynamical model, and the back-projections of the human latent variables $\mathbf{x}^2_N$ onto the interaction variable $\mathbf{i}_N$. This interaction variable is projected down the hierarchy to generate joint angles $\mathbf{x}^1_N$ for rendering, setting the desired emotional style $\mathbf{e}$ in the bottom layer.

subject	**emo. class. rates ± std.error**		**P(i>non-i)**
	interactive	non-interactive	
S1	0.617 ± 0.054	0.519 ± 0.055	0.911
S2	0.741 ± 0.048	0.716 ± 0.050	0.653
S3	0.765 ± 0.047	0.802 ± 0.044	0.252
S4	0.768 ± 0.050	0.696 ± 0.055	0.866
S5	0.716 ± 0.050	0.728 ± 0.049	0.422
Overall P(i>non-i)			**0.970**

Table 1: *Emotion classification rate comparison between the 'interactive' and the 'non-interactive' condition in the tipping experiment, and probabilities* ***P(i>non-i)*** *that classification rates in the interactive condition is larger than in the non-interactive one. The trend* ***i>non-i*** *seems to hold in additional subjects which we have recorded meanwhile (not shown). For details, see text.*

5 Tests

Tipping In the first test of our interactive system we tried to determine if the active involvement of a human in the elicitation of a *natural* (i.e. without style morphing) emotional response from the avatar would improve emotion perception. The human had to tip the avatar on its shoulder from behind, which triggered the avatar's turning around. We also placed a dummy in the scene for haptic feedback. The avatar turned in one of three emotional styles: neutral, fearful or angry. We chose these emotions because they are on one emotional 'axis'. Participants classified the perceived emotion. The experiment consisted of 3 blocks of 12 trials with stimuli presented in pseudo-random order. Furthermore, we also asked the participants to repeat the classification non-interactively, by viewing recordings of their *own* previous trials, to minimize the risk that any differences between the conditions could be caused by factors other than interactivity. To reduce memorization effects, which might reduce the classification/rating difference between conditions, interactive and corresponding non-interactive (i.e. recorded) blocks were presented with one different interactive block in between. The experiment took 60-90 minutes in total.

While we have not yet collected enough data for definitive conclusions, the preliminary results are promising, see table 1. Shown in this table are emotion classification rates (ECRs) and standard errors, assuming a Beta distribution [Bishop 2007] as the posterior distribution over ECR. This posterior is the canonical choice for a binomial observation model, here: correct vs. incorrect answers. Standard errors are computed from this posterior, as is the probability **P(i>non-i)** that the interactive ECR is larger than the non-interactive ECR. The overall P(i>non-i) is not the average of the individual probabilities, but rather the probability that *all* subjects have a higher ECR in the interactive condition, versus the hypothesis that all subjects have a lower ECR in the interactive condition. This preliminary result indicates that an interactive setting improves emotion perception, which further motivates the importance of the development of systems that can synthesize reactive emotional movements in real-time.

'High-five' interactions. We applied our approach to generate online four types of emotional 'high fives' (neural, happy, angry, sad). We learned 105 motion capture trajectories which were performed on an imaginary 3×3 grid for hand contacts. 'High fives' were repeated ≈ 3 times in each emotional style at every location. The whole data-set includes 5605 data points, where we represented the joint angle trajectories as quaternions to enable good style interpolation. We trained the model with 200 inducing inputs for each latent space, on an AMD Phenom X6 3.3 GHz computer this took about 5 hours.

The synthesized motions looking quite natural, see supplementary video for an examples movie of the avatar and the human in the motion-capture setup. Also, the hand contacts match the desired position for reconstruction without emotional style. With differentiated emotional style, the hand position of the avatar is slightly off, an effect that is most noticeable for a 'happy' avatar. We plan to improve this hand contact by applying effectors at the end of the kinematic chain and optimizing the pose w.r.t the position of these effectors.

6 Summary

We have implemented a new real-time-capable system for the accurate simulation of interactive human movements. The method is based on a hierarchical probabilistic model that combines layers of GP-LVMs with a GPDM on top. The proposed method is able to model the emotional style by appropriate additional dimensions in the latent space. We are now able to study the influence of interaction on human perception; first result indicate that involvement in a scene improves emotion classification. Future work will exploit morphing between emotions to reduce memory effects between conditions further.

Acknowledgments: This work was supported by EU projects FP7-249858-TP3 TANGO, FP7-ICT-248311 AMARSi, FP7-PEOPLE-2011-ITN(Marie Curie): ABC PITN-GA-011-290011, the Human Brain Project; 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

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

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.

CHAN, R., LI, H., CHEUNG, E., AND GONG, Q.-Y. 2010. Impaired facial emotion perception in schizophrenia: A meta-analysis. *Psychiatry Research 178*, 381–390.

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

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., COURT, R., AND SCIENCE, C. 2006. Local Distance Preservation in the GP-LVM through Back Constraints. In *ICML*, 513 – 520.

LAWRENCE, N. 2007. Learning for larger datasets with the Gaussian process latent variable model. *Journal of Machine Learning Research - Proceedings Track 2*, 243–250.

LEVINE, S., WANG, J. M., HARAUX, A., POPOVIĆ, Z., AND KOLTUN, V. 2012. Continuous character control with low-dimensional embeddings. *ACM Transactions on Graphics 31*, 4 (July), 1–10.

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.

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

PHILIP, R., WHALLEY, H., STANFIELD, A., SPRENGELMEYER, R., SANTOS, I., YOUNG, A., ATKINSON, A., CALDER, A., JOHNSTONE, E., LAWRIE, S., AND HALL, J. 2010. Deficits in facial, body movement and vocal emotional processing in autism spectrum disorders. *Psychological Medicine 40*, 1919–1929.

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.

SCHAEFER, K., BAUMANN, J., RICH, B., LUCKENBAUGH, D., AND ZARATE JR, C. 2010. Perception of facial emotion in adults with bipolar or unipolar depression and controls. *Journal of Psychiatric Research 44*, 1229–1235.

TAUBERT, N., CHRISTENSEN, A., ENDRES, D., AND GIESE, M. A. 2012. Online simulation of emotional interactive behaviors with hierarchical Gaussian process dynamical models. In *Proceedings of the ACM Symposium on Applied Perception - SAP '12*, ACM Press, New York, New York, USA, vol. 1, 25–32.

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

WICKER, B., PERRETT, D. I., BARON-COHEN, S., AND DECETY, J. 2003. Being the target of anothers emotion: a pet study. *Neuropsychologia 41*, 2, 139–146.

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.