Stylistic Locomotion Modeling with Conditional Variational Autoencoder

H. Du[1,2], E. Herrmann[1,2], J. Sprenger[1,2], N. Cheema[1,2,3], S. Hosseini[1,2], K. Fischer[1,2] and P. Slusallek[1,2]

[1]German Research Center for Artificial Intelligence (DFKI) Saarbrücken, [2]Saarland University, [3]Max-Planck Institute for Informatics; Germany

Abstract

We propose a novel approach to create generative models for distinctive stylistic locomotion synthesis. The approach is inspired by the observation that human styles can be easily distinguished from a few examples. However, learning a generative model for natural human motions which display huge amounts of variations and randomness would require a lot of training data. Furthermore, it would require considerable efforts to create such a large motion database for each style. We propose a generative model to combine the large variation in a neutral motion database and style information from a limited number of examples. We formulate the stylistic motion modeling task as a conditional distribution learning problem. Style transfer is implicitly applied during the model learning process. A conditional variational autoencoder (CVAE) is applied to learn the distribution and stylistic examples are used as constraints. We demonstrate that our approach can generate any number of natural-looking human motions with a similar style to the target given a few style examples and a neutral motion database.

CCS Concepts

• ***Computing methodologies*** → ***Animation;*** *Motion processing;*

1 Introduction

The synthesis of human motion with large variation and different styles has a growing demand for simulation applications such as games, psychological experiments and ergonomic analysis. Therefore, data-driven motion synthesis approaches based on machine learning have captured considerable research interest. However, constructing a representative motion model usually requires a large amount of example data. Therefore, it is of interest to efficiently reuse recorded motion data for different scenarios [MC12]. Motion style transfer provides the possibility of creating a synthetic stylized motion database from existing data without additional motion capturing efforts.

In this work, we present a novel approach to combine statistical motion modeling and style transfer. We follow the assumption from [MC12], although human motion has infinite variations, the high level structures (motion primitive) are finite. A motion primitive contains structurally and semantically similar motion clips and can be modeled by a statistical distribution. We formulate the stylistic motion modeling task as a conditional distribution learning problem. A large amount of neutral motions which contain rich variation in content is used as training data, and a limited number of style examples are taken as constraints. The content and style motions are encoded by a pre-trained convolutional autoencoder and used to train a conditional variational autoencoder (CVAE) to model the conditional distribution. In summary, our contribution is to propose a novel method to create a generative model for the style motion by combining the variation from a large neutral motion database and a few style examples.

2 Background

Our work is constructed upon previous successful work on the generative statistical models for human motion synthesis and human motion style transfer. In this section, we first briefly discuss the development of generative models for human motion data. Then we review the various methods for motion style transfer.

2.1 Generative Statistical Motion Modeling

Bowden [Bow00] uses Hidden Markov Models (HMMs) to model human motion distribution. Min and Chai [MC12] project motion data into low-dimensional space and use Gaussian Mixture Model (GMM) to learn the distribution of motions in low-dimensional space. Wang et al. [WFH08] address the dynamics of human motion by introducing the Gaussian Process Dynamical Model (GPDM) to model the temporal sequence of human motion. A smooth low-dimensional space is found for motion data by taking dynamics as the constraint. haha Recently generative models based on deep neural networks have demonstrated outstanding performance on motion modeling. Holden et al. [HSKJ15] employ convolutional autoencoders to learn a continuous manifold space for motion representation. Motegi et al. [MHM] apply variational autoencoder to model the distribution of a large motion database. Variants of recurrent variational autoencoders are used to model

DOI: 10.2312/egs.20191002

the dynamics of motion data. Habibie et al. [HHS*17] use a recurrent variational autoencoder to model the sequence of motion data. Fragkiadaki et al. [FLFM15] propose the Encoder-Recurrent-Decoder model for human body pose prediction.

Crowd simulation [NGCL09] models the large-scale behavior of human as well. However, they are more interested in high-level planning, for instance, crowd steering, collision avoidance and so on. In this work, we focus on enriching the variations and styles of the motion itself.

2.2 Style Transfer for Motion Data

A significant amount of effort has been spent on the problem of style transfer for human motion data. One set of approaches is trying to separate style components from the content of the motion. Brand and Hertzmann [BH00] apply HMMs to learn a set of style-specific models and a generic model to encode style from motion content and generate new stylistic motions. Min et al. [MLC10] construct a multi-linear model to learn the parameters for content and style from a motion database of the same action with different styles. Xia et al. [XWCH15] use local mixtures of autoregressive models to achieve realtime style transfer for unlabeled heterogeneous human motion. Yumer and Mitra [YM16] observe that the magnitude of spectrum is more relevant to the style of action and the phase is more relevant to the content. They formulate motion style transfer as an optimization problem and achieve style transfer between different actions in the spectral domain. Deep learning has also been applied to learn style transfer. Holden et al. [HHKK17] automate this requirement by using the Gram matrix [GEB15] to extract the style of motions from features learned by 1D convolution. In this work, we do not focus on explicit style transfer between motions, instead we apply the style as a constraint in conditional distribution modeling.

3 System Overview

The goal of the work is to create a generative model for stylistic motion modeling and synthesis, based on a neutral motion database and a few of style examples. Our motion synthesis pipeline is based on previous work [MC12, DHM*16]. Each action is represented by a directed graph. The nodes are high-level structures of the action named motion primitives. The edges represent the possible transitions between the nodes. For instance, six motion primitives can well represent arbitrary normal walking: "leftStance", "rightStance", "beginLeftStance", "endLeftStance", "beginRightStance" and "endLeftStance". The main difference between our work and previous work [DHM*16] is that instead of a parametric model like GMM, we use a variational autoencoder (VAE) to model the distribution of motion primitives and encode the style of motion as a conditional distribution. New motion can be generated by first taking a graph walk in the motion graph, then sampling each motion primitive to find the target motion clip. The advantage of this framework is that variations of motions are encoded in the model while the high-level structures of the action are also maintained.

4 Motion Data Acquisition

Our motion database is recorded by an OptiTrack system with three male actors. The whole database contains 40 minutes of locomotion, including walking, running and jogging. We use our captured motion as a neutral motion database since the style in the captured motions are relatively similar. For the style data, we use stylistic walking data from [XWCH15]. A retargeting approach [MBBT00] is implemented to retarget all motion data to a MakeHuman† game engine skeleton.

All motions in our motion dataset are parameterized as 3D joint positions. The reason we use joint position is because joint positions are more coherent with the visual observation of motion [DHM*16], which is suitable for training. In our work, we use a skeleton with 21 joints. In addition to joint positions, the global speed on the 2D ground and the rotational velocity about the vertical axis are computed and added to each frame. So in our dataset, each frame is represented as a vector of length 66.

All frames are normalized to have the same global position on 2D ground and face the same direction. Similar to [HSK16], we decompose the long recordings into small motion clips using an overlapping window size of 60 frames with an overlap of 30 frames.

5 Stylistic Motion Modeling

Figure 1 shows the modeling pipeline for each motion primitive. Our goal is to learn a conditional distribution $P(\mathbf{X}|\mathbf{x}_s)$ based on a large number of content motion clips $\mathbf{X}$ and a style constraint $\mathbf{x}_s$. If there is no style constraints given, the model simply learns the distribution of content motion clips $P(\mathbf{X})$ for each motion primitive. If there are style examples, the distribution can be deformed to a new distribution based on different style input $\mathbf{x}_s$. Similar to [HSK16], we use a single layer convolutional autoencoder to encode motions into feature space $\mathbf{Y} = \Phi(\mathbf{X})$. Style is encoded in feature space using a Gram matrix. The distribution of motion clips in feature space is learned by VAE. We formulate the Gram matrix as a style constraint term in the loss function of VAE. Therefore, a conditional distribution can be learned by training the model.

5.1 Motion Feature Extraction

Using a Gram matrix to extract styles from features produced by convolutional neural networks has achieved great success in image style transformation [GEB15]. Similar to [HSK16, HHKK17], we construct a single layer convolutional autoencoder to perform 1d convolution on motion data to extract features.

5.2 Variational Autoencoder for Human Motion Model

In this section, we will briefly review variational autoencoder (VAE) proposed by [KW13]. VAE assumes that data $\mathbf{Y}$ can be encoded in a low-dimensional latent space $\mathbf{z}$ and the distribution $P(\mathbf{Y})$ can be computed by the integral of the marginal likelihood as $P(\mathbf{Y}) = \int P_\theta(\mathbf{Y}|\mathbf{z})P_\theta(\mathbf{z})d\mathbf{z}$. The detailed explanation of VAE can be found in [KW13].

In our work, the encoder and decoder are both modeled by a four-layer feed-forward network. The hidden units for the encoder are 512, 256, 128 and 32, and for the decoder are 32, 128, 256 and

† MakeHuman: http://www.makehumancommunity.org/

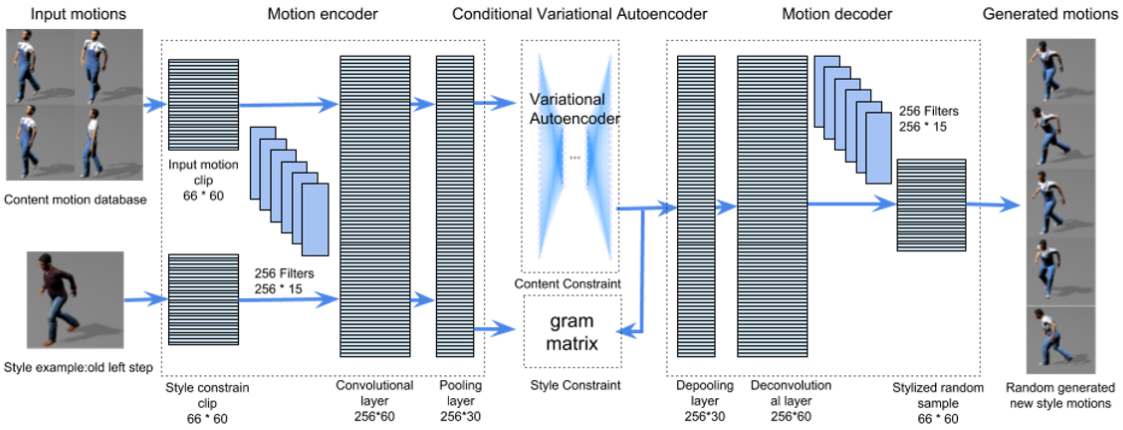

Figure 1: *The overview of stylistic motion primitive modeling. The input is a large set of content motions with rich variations and a single style example. The input motions are encoded into feature space by motion decoder, which is trained on all motion clips. The conditional variation autoencoder (CVAE) is used to model the conditional distribution. New motion can be generated from CVAE and back to motion space via motion decoder.*

512. Tanh is used as activation function for all layers except for output layer in both encoder and decoder. The loss function $\mathbf{L}$ is defined as follows:

$$\begin{aligned}\mathbf{L} &= \mathbf{L}_{content} + \mathbf{L}_{KL} \\ &= -E_z[log(P_\theta(\mathbf{Y}|\mathbf{z}))] + D[Q_\phi(\mathbf{z}|\mathbf{Y})||P(\mathbf{z})]\end{aligned} \tag{1}$$

For the first reconstruction term, $E_\mathbf{z}$ can be approximated by sampling $\mathbf{z}$:

$$-E_z[log(P_\theta(\mathbf{Y}|\mathbf{z}))] = \frac{1}{L}\sum_{l=1}^{L}(\frac{1}{N}\sum_{i=1}^{N} log(e^{-\frac{1}{2}||decoder(\mathbf{z}_{i,l})-\mathbf{y}_i||^2})) \tag{2}$$

D is the Kullback-Leibler divergence, which serves as regularization term. The analytical solution can be computed since P_θ and Q_ϕ are both multivariate Gaussian.

5.3 Conditional Variational Autoencoder

Creating a generative model for a certain style can be formulated as a conditional distribution modeling $P(\mathbf{Y}|\mathbf{y_s})$. The stylistic example $\mathbf{x_s}$ is first encoded into feature space $\mathbf{y_s}$ using convolutional encoder Φ. The style is extracted using a Gram matrix, which is defined as the sum of inner product of features over the temporal axis.

$$Gram(\mathbf{y}_s) = \sum_i \mathbf{y_{s,i}}\mathbf{y_{s,i}^T} \tag{3}$$

As the Gram matrix sums over frames, it does not require frame alignment between content motion and style motion. So for style constraints, single or multiple style examples can be used. The distribution should deform based on different style inputs. This is achieved by adding style constraint into the loss function of VAE.

$$\mathbf{L}_{style} = \alpha||Gram(decoder(\mathbf{z})) - Gram(\mathbf{y}_s)|| \tag{4}$$

where α is the style weight to control the magnitude of the effect of the style. We empirically set α to 200. Our final loss function for conditional VAE model is:

$$\mathbf{L} = \mathbf{L}_{content} + \mathbf{L}_{KL} + \mathbf{L}_{style} \tag{5}$$

6 Experiments

We evaluate our method on a fairly large content database and six distinctive styles, which are: depressed, proud, old, childlike, sexy and angry.

The convolutional autoencoder is trained on a large dataset with 122624 clips, which not only contains locomotion, but other actions as well. We train the model with 300 epochs and training rate 0.00001 on a NVIDIA GeForce GTX 760. The training takes roughly 10 hours. For training each motion primitive, we use the pre-trained convolutional autoencoder to encode motion clips into feature space. We set epochs to 300 with a training rate 0.0001. The training time depends on the number of samples in each motion primitive. For instance, for walk leftStance, there are 749 clips which takes about 50 minutes to train. All networks are implemented using Tensorflow.

6.1 Motion Primitive Evaluation

Motion primitives serve as the core of our motion synthesis framework. The quality and variation of generated motions will decide the quality of the completed motion. Figure 2 shows some random samples generated from two motion primitives: walk leftStance and run rightStance. They are modeled using VAE on neutral database without style constraints. All characters start in a line with equal spacing. The last frame of each clip is displayed in Figure 2.

6.2 Stylistic Motion Primitive Evaluation

Figure 3 shows random samples from six stylistic variants of walk left and right stance. We use one stylistic clip as style constraint to train conditional variational autoencoder. All the style examples are selected to walk in a straight line to minimize the variations between style examples. From the sampling results, we can see that the stylistic models have good variations in both poses and trajectories.

Figure 2: *Random samples from motion primitives trained on neutral dataset. From top to down: walk leftStance, run rightStance.*

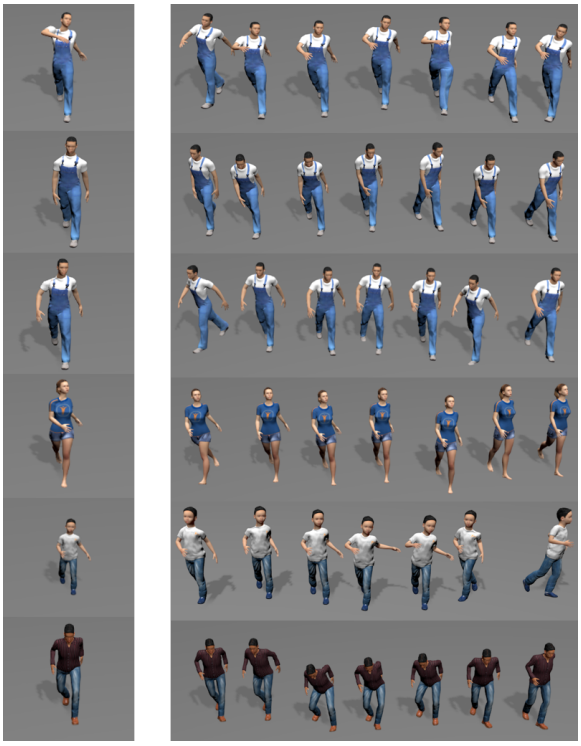

Figure 3: *Random samples from six stylistic motion primitives deformed from neutral walk leftStance. The left side is the style constraints. For top to bottom, the styles are: proud, depressed, angry, sexy childlike and old. The right side is the samples generated from each stylistic motion primitive.*

7 Conclusion

In this work, a new approach to create generative models for stylistic locomotion has been presented. Our approach does not require a large amount of stylistic motions for training. Our model makes use of motion style transfer to implicitly convert neutral motion to a target style during training. In order to demonstrate our method, we test six different styles on walking. For each style, a walking motion graph with six stylistic motion primitives are constructed. The work we presented in this paper focuses on stylistic locomotion modeling and synthesis. However, our approach is not limited to locomotion. We believe the approach to combine the variations in a large neutral motion database and the style from a few stylistic examples to learn generative models could be applied to all kinds of actions.

Acknowledgements

This work is funded by the German Federal Ministry of Education and Research (BMBF) through the projects React (grant agreement No.: 01IW17003) and Hybr-iT (grant agreement No.: 01IS16026A) and the ITEA3 project MOSIM (grant number: 01IS18060A-H).

References

[BH00] BRAND M., HERTZMANN A.: Style machines. In *Proceedings of ACM SIGGRAPH* (2000). 2

[Bow00] BOWDEN R.: Learning statistical models of human motion. In *IEEE Workshop on Human Modeling, Analysis and Synthesis, CVPR* (2000), vol. 2000. 1

[DHM*16] DU H., HOSSEINI S., MANNS M., HERRMANN E., FISCHER K.: Scaled functional principal component analysis for human motion synthesis. In *Proceedings of the 9th International Conference on Motion in Games* (2016), ACM, pp. 139–144. 2

[FLFM15] FRAGKIADAKI K., LEVINE S., FELSEN P., MALIK J.: Recurrent network models for human dynamics. In *Proceedings of the IEEE ICCV* (2015), pp. 4346–4354. 2

[GEB15] GATYS L. A., ECKER A. S., BETHGE M.: A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015). 2

[HHKK17] HOLDEN D., HABIBIE I., KUSAJIMA I., KOMURA T.: Fast neural style transfer for motion data. *IEEE computer graphics and applications 37*, 4 (2017), 42–49. 2

[HHS*17] HABIBIE I., HOLDEN D., SCHWARZ J., YEARSLEY J., KOMURA T.: A recurrent variational autoencoder for human motion synthesis. *BMVC17* (2017). 2

[HSK16] HOLDEN D., SAITO J., KOMURA T.: A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG) 35*, 4 (2016), 138. 2

[HSKJ15] HOLDEN D., SAITO J., KOMURA T., JOYCE T.: Learning motion manifolds with convolutional autoencoders. In *SIGGRAPH Asia 2015 Technical Briefs* (2015), ACM, p. 18. 1

[KW13] KINGMA D. P., WELLING M.: Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013). 2

[MBBT00] MONZANI J.-S., BAERLOCHER P., BOULIC R., THALMANN D.: Using an intermediate skeleton and inverse kinematics for motion retargeting. In *Computer Graphics Forum* (2000), vol. 19. 2

[MC12] MIN J., CHAI J.: Motion graphs++: A compact generative model for semantic motion analysis and synthesis. *In ACM Transactions on Graphics 31*, 6 (2012), 153:1–153:12. 1, 2

[MHM] MOTEGI Y., HIJIOKA Y., MURAKAMI M.: Human motion generative model using variational autoencoder. 1

[MLC10] MIN J., LIU H., CHAI J.: Synthesis and editing of personalized stylistic human motion. In *Proceedings of the 2010 ACM SIGGRAPH I3D* (2010), ACM, pp. 39–46. 2

[NGCL09] NARAIN R., GOLAS A., CURTIS S., LIN M. C.: Aggregate dynamics for dense crowd simulation. In *ACM transactions on graphics (TOG)* (2009), vol. 28, ACM, p. 122. 2

[WFH08] WANG J. M., FLEET D. J., HERTZMANN A.: Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence 30*, 2 (2008), 283–298. 1

[XWCH15] XIA S., WANG C., CHAI J., HODGINS J.: Realtime style transfer for unlabeled heterogeneous human motion. *ACM Transactions on Graphics (TOG) 34*, 4 (2015), 119. 2

[YM16] YUMER M. E., MITRA N. J.: Spectral style transfer for human motion between independent actions. *ACM Transactions on Graphics (TOG) 35*, 4 (2016), 137. 2