Full-Body Gait Reconstruction Using Covariance-Based Mapping Within a Realtime HMM-Based Framework

Joëlle Tilmanne[1] and **Nicolas d'Alessandro**[1] and **Thierry Ravet**[1] and **Maria Astrinaki**[1] and **Alexis Moinet**[1]

Abstract.

In this paper we propose a new HMM-based framework for the exploration of realtime gesture-to-gesture mapping strategies. This framework enables the realtime HMM-based recognition of a given gesture sequence from a subset of its dimensions, the covariance-based mapping of the gesture stylistics from this subset onto the remaining dimensions and the realtime synthesis of the remaining dimensions from their corresponding HMMs. This idea has been embedded into a proof-of-concept prototype that "reconstructs" the lower-body dimensions of a walking sequence from the upper-body gestures in realtime. In order to achieve this reconstruction, we adapt various machine learning tools from the speech processing research. Notably we have adapted the HTK toolkit to motion capture data and modified MAGE, a HTS-based library for reactive speech synthesis, to accommodate our use case. We have also adapted a covariance-based mapping strategy used in the articulatory inversion process of silent speech interfaces to the case of transferring stylistic information from the upper- to the lower-body statistical models. The main achievement of this work is to show that this reconstruction process applies the inherent stylistics of the input gestures onto the synthesised motion thanks to the mapping function applied at the state level.

1 INTRODUCTION

It is pretty straightforward to assume that we live in a technological context where the capture, understanding and synthesis of human gestures lead to unprecedented opportunities. Indeed motion capture (mocap) technologies are finally moving out of the experimental era and a growing amount of research groups and companies are now putting their hands on very accurate and easy-to-use motion capture systems. This intense development of body tracking solutions has definitely brought Natural User Interaction (NUI) to the mainstream.

With this massive increase of NUI opportunities comes a growing demand for the further understanding of human gestures and the leveraging of such advanced knowledge in new realtime applications. Encountered issues are related to both the recognition of ongoing gestures and the generation of humanlike motion sequences. In many scenarios machine learning has become a very privileged way of addressing these issues and various classes of algorithms based on statistical models have emerged over the last decades [9]. Approaches based on Hidden Markov Models (HMMs) are probably among the most popular [4, 12] but human motion and action modelling is now considered under an increasing amount of viewpoints.

In our work we are very interested in gestures that require a certain level of motoric skills and for which the use of machine learning can lead to better tracking, representation or query of those skills: dance, musical practice, craftsmanship, gait, etc. Beyond the functional classification of gestures, we put the focus on motion *stylistics*. By styles, we mean the possible variations encountered in the realisation of a gesture for an identical functional pattern: a step in gait, grabbing an object, playing a musical phrase, etc. There are many causes for the variability of a gesture: intra- or inter-personal differences, emotional state, school of practice (e.g. French vs. Russian piano technique), etc. We think that considering these styles in both the recognition of gestures and the synthesis of motion sequences can greatly improve the expressivity of resulting applications.

In this project we have developed a realtime application that aims at addressing several interesting questions about gesture stylistics. As a proof of concept, our applications is "reconstructing" full-body human gait motion in realtime. The reconstruction process relies on three components using various machine learning concepts:

1. online gesture recognition: HMMs are used to recognise and follow the ongoing step sequence from an input stream of mocap data corresponding to the upper body joints (head, torso and arms);
2. covariance-based mapping: the full covariance matrices of HMM emitting distributions are used to "project" the stylistics of the upper-body HMMs onto the lower-body HMMs (hips and legs);
3. realtime motion synthesis: motion trajectories of the lower body joints are synthesised in realtime from the lower-body HMMs.

In this paper we first give an overview of the literature in gesture recognition and motion synthesis in Section 2. In Section 3 we describe our mocap data (a database of stylistic human gait) and our motion model. Then we describe our gait reconstruction process in Section 4. Finally we conclude and prospect about future works.

2 RELATED WORK

In Section 1 we have shown that our approach lies between two very specific problems involving machine learning: the online recognition of input gestures and the realtime synthesis of humanlike motion. Here we give an overview of related work in these fields.

2.1 Gesture recognition

There is an inherent variability in the realisation of human gestures. Literature mentions several classes of approaches to deal with such variability in gesture recognition, depending on the considered application. Particularly, online detection of human activity appears to

[1] numediart Institute for Creative Technologies, University of Mons (UMONS), Belgium, email: firstname.lastname@umons.ac.be

be a difficult problem. One popular approach is the Dynamic Time Warping (DTW) process and its numerous variants [2]. DTW consists in the realignment and distance evaluation between the ongoing gesture and a series of reference gestures. Another very successful class of algorithms use HMMs to "summarise" the time series corresponding to the reference gestures, taking into account their inherent variability. Then the likelihood of the ongoing gesture is computed from these models to recognise the reference gesture. Running HMM-based recognition in realtime is not trivial. The Gesture Follower [3] lies somewhere between DTW- and HMM-based gesture decoding. Indeed a single occurrence of the reference gesture is captured and used to build a giant HMM. Then subsequent occurrences are simultaneously realigned and evaluated with the likelihood measure, as the gesture is happening. We can also find gesture recognition approaches based on Hidden Conditional Random Forests (HCRF), where motion is modelled as a fixed amount of poses represented by hidden states [17]. Other approaches include, for instance, particle filtering, condensation algorithms, Deep Neural Networks (DNN), etc. Exhaustive reviews of the literature can be found in [9, 5].

If we think beyond motion, the area where time series recognition is the most advanced is speech processing. Indeed speech recognition has brought very sophisticated and matured toolkits, especially using HMMs (see for instance HTK, the Hidden Markov Model Toolkit [16]). In this research we have taken the approach of considering that speech and motion exhibit similar trends. Therefore we use HMM-based speech recognition tools for gesture recognition.

2.2 Motion synthesis

Machine learning has brought several representations and generative models for trajectory generation. Some techniques exhibit interesting properties for handling styles, such as the automatic separation of the stylistic component in the model: PCA [15], Conditional Restricted Boltzmann Machine (CRBM) [10] or Dynamic Bayesian Network [8] can be found in the literature. HMMs have also widely been used for motion synthesis [4]. In previous work about gait synthesis, we have also proved that HMM-based generative approaches proposed in the speech processing research – like HTS, the HMM-Based Speech Synthesis System [14] – could advantageously be adapted to motion synthesis and handle stylistic information in a flexible way [12].

3 MOTION MODELLING WITH HMM

Our approach towards HMM-based gesture recognition and motion synthesis is to adapt the advanced speech recognition and synthesis toolkits so that the phonetic sequence is replaced by motion segments and the algorithms take mocap data as their feature vectors. In HMM-based speech processing, the synthesis task has "emerged" from the recognition research, with the complementarity between HTS and HTK. So we do with motion data, i.e. we train HMMs on a corpus of motion-captured data and we use these trained models both for the recognition and the synthesis steps. As described in Section 4, only working within this unified recognition/synthesis framework allows us to achieve the reconstruction process, i.e. generate new motion trajectories from realtime-recognised gestures.

Our proof-of-concept application benefits from previous work in gait synthesis and the availability of the Mockey database [13]. In the Mockey database, one actor performs eleven different styles arbitrarily chosen for their recognisable expressivity, such as proud, macho, afraid, drunk, etc. The data was recorded with an inertial motion capture suit: the Animazoo IGS 190 [1] and acquired at 30 frames per second. The skeleton is represented by 18 3D joints, hence giving 54 dimensions to the motion data which consists in 3D angles parameterised as exponential maps. The human walk is labelled as an alternate sequence of left and right steps. We use one HMM for each step. The topology is five-state left-to-right with no skip and observations are modelled by one multidimensional full-covariance Gaussian distribution. This setup is similar to the ones used in speech processing. Figure 1 illustrates this topology for a simplified case of three emitting states instead of five, for readability reasons. In this use case, each HMM encounters all the stylistic variations, as performed by the actor in the inertial suit. Therefore we end up with two HMMs, corresponding to a clustering of the gait space in two steps.

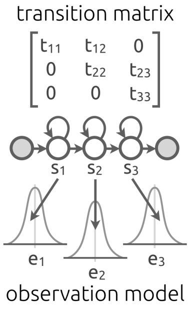
transition matrix
$\begin{bmatrix} t_{11} & t_{12} & 0 \\ 0 & t_{22} & t_{23} \\ 0 & 0 & t_{33} \end{bmatrix}$
s_1
s_2
s_3
e_1
e_2
e_3
observation model

Figure 1. Topology of the HMMs that we have used to model walking sequences: left-to-right with no skip with Gaussian distributions in emitting states. In this example, we use 3 states instead of 5 for readability reasons.

4 REALTIME GAIT RECONSTRUCTION

Our work with motion stylistics draws attention towards two different approaches. In previous work we have explored the idea that gesture styles can be labelled explicitly [11]. This differentiation leads to a large amount of small clusters in the motion space (Figure 2 left). This choice impacts on application design in various ways. Notably it suggests that any new gesture, to recognise or generate, should be seen as a combination of these small stylistic leaves in which it is difficult to navigate. In this work we adopt an implicit approach towards style control since no explicit tagging of the styles is achieved. Each gesture is a large cluster in the motion space and navigating inside the cluster is made possible with an appropriate mapping function (Figure 2 right).

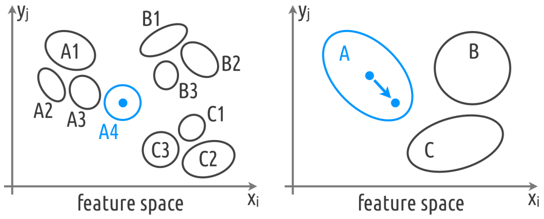

Figure 2. Comparaison between an explicit labelling of motion styles (many clusters and no mapping possible) and an implicit approach towards the same question (few clusters and an in-cluster mapping required).

Our realtime gait reconstruction application was built to validate the concept of the exploration of a stylistic motion space using a subset of the motion dimensions to recognise the motion and its style in realtime and drive the stylistic synthesis of the remaining dimensions in realtime. We have built a prototype that will synthesise the stylistic gait (motion + style) of the legs of a virtual character using, as an input, the gestures captured from the upper part of the body during the gait sequence. The process is hence split into three realtime-performed tasks: online HMM-based gesture recognition, covariance-based mapping and reactive HMM-based synthesis. The whole process is illustrated in Figure 3. An illustration of the realtime stylistic walk reconstruction can be found at `http://youtu.be/gB2Bz5Nx8oU`.

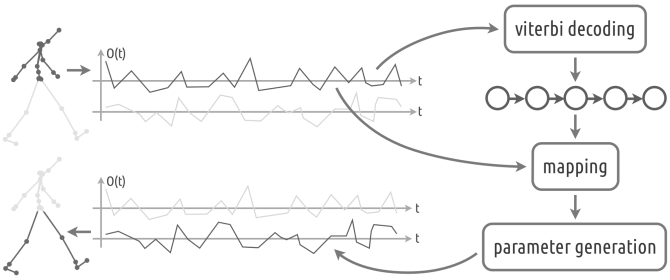

Figure 3. Illustration of the overall process used in the gait reconstruction example: continuous inputs are decoded with a realtime Viterbi algorithm. This decoding generates an ongoing state sequence that is used for synthesis of the outputs. Before emitting distributions are used for synthesis, means are modified by a mapping function based on covariance.

4.1 Online Gesture Recognition

The first task in the gait reconstruction process is the decoding of the input gestures. For testing our application, we simulate the input motion by sending a realtime data stream with the upper-body joints to the recognition module. In the current stage, this module performs a simple realtime Viterbi algorithm. Indeed it only performs the forward step of the standard Viterbi procedure, as illustrated in Figure 4. In this forward-only algorithm, the probability of being in each state at each time t is computed in the same way as for the standard Viterbi algorithm. However at each time t, a decision is taken and the most likely state is considered as the decoded state. The path is hence defined at each increment in the input sequence.

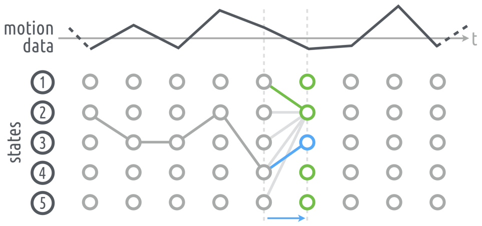

Figure 4. Realtime Viterbi decoding: illustration of the forward-only approach on a simplified five-state model. The "best" state sequence is determined for each increment in the Viterbi lattice.

Since the decoding is performed on a subset of the dimensions of the original data, we implemented a mask vector that inhibits the dimensions corresponding to the outputs of the reconstruction process. The realtime Viterbi decoding provides the most likely HMM (and hence label) that corresponds to the streamed data and the most likely current state of the model. Once the most likely current state has been decoded, it can be used to build the state sequence for the synthesis stage. Before the stack of observation density functions can be accumulated for the synthesis step, the means are modified in order to take into account the style of the incoming streamed data, which is the next step of the reconstruction procedure.

4.2 Covariance-Based Mapping

Once the underlying model and state is determined for each observation of the input gesture (in our case: upper-body motion), our system owns two important pieces of information. On the one hand, we can associate one lower-body emitting distribution (multivariate Gaussian of the lower-body joints) for each sample of the recognised upper-body gesture. It is important to highlight that these queried means and covariances correspond to the large clusters described in Figure 2 and therefore to a gait model summarising the eleven styles. On the other hand, we can compare the queried upper-body Gaussians with the ongoing upper-body motion values. This "distance" between the average motion statistics and the ongoing input motion conveys information about the motion stylistics, i.e. it informs about where the ongoing input motion is located in its cluster.

Then we need a mapping function in order to convert this estimated distance in the upper-body space into a model transformation in the lower-body space, transformation that we aim at being stylistically consistent. In their work on articulatory inversion for the creation of silent speech interfaces, Hueber *et al.* have suggested the basics of such a mapping function [7]. The algorithm proposes to shift the means and variances of the target models (in our case: lower-body motion) according to the evaluated distance between input models (in our case: upper-body motion) and ongoing input data. The mapping function uses the covariance between input and target data as a way of projecting the distance evaluated in the input space onto the target space. Once the distance vector has been projected, we can shift the means of the target models accordingly. In the silent speech interface, it enables the speech sound to drive the articulatory models of the jaw and the tongue. In the preliminary results of our gait reconstruction prototype, we show that it allows the style of the upper-body to get transferred on lower-body motion.

4.3 Reactive HMM-Based Synthesis

Every time a state model is generated by the two previous steps of the reconstruction, it is then pushed into a queue that we name the *state queue*. We already use the state queue for reactive HMM-based synthesis with MAGE and it is introduced in [6]. Each element of the state queue corresponds to an analysis frame of the upper part of the body that has been recognized and for which we are going to use the state models to synthesize a set of features for the lower part of the body. The state queue is implemented as a ring buffer, as illustrated in Figure 5 and between its tail and its head, it contains the last N states pushed. These states are divided into three groups:

- past states (Q_P): the first P state models in the queue (i.e. the oldest models) correspond to frames of lower features that have already been computed in previous iteration(s)
- current states (Q_C): the next C state models corresponds to frames of lower features that will be computed with the current iteration
- future states (Q_F): the last F state models corresponds to frames of lower features that will be computed in upcoming iteration(s)

Q_P and Q_F give some contextual information around Q_C necessary to compute the C current frames in a smooth continuity with

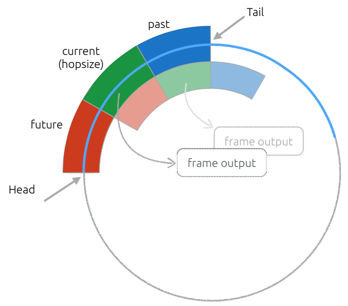

Figure 5. State queue implemented as a ring buffer. Showing two iterations of the process, past iteration shown in dim colours and current iteration shown in bright.

the past and future frames and, as such, large values of P and F ensure a better reconstruction. However, increasing P and F also increases the cost of computing the current set of features. Besides, each "future" state in the queue actually corresponds to a frame of input features that has already been recorded and recognized. Thus, using F future states creates a delay of F frames between the time a set of features from the upper part of the body is input in the system and the time its corresponding set of lower features is output. Also note that state sequences recognized a few seconds in the past have generally no impact on the result of the computation of the current features, therefore large values of P are unnecessary.

In the case at hand, we want to minimize the delay between the input and the output to be as reactive as possible. Therefore, we set $P = 20$ and $F = 0$ to set the state queue at a zero-frame delay.[2] As for C, the relatively low frame rate allows us to make the whole computation one frame at a time and we set $C = 1$. Note however that for higher frame rate (or more computationally expensive cases) we may need to compute blocks of several frames per iteration ($C > 1$) instead.

From the $P+C+F$ state models, C frames containing the features of the lower part of the body are computed [6]. Once these C output frames have been computed, the oldest C frames are dropped from the state queue, thus advancing in the state queue with a hopsize equal to C, and the next iteration can begin with insertion of C new state models.

5 FUTURE WORK & CONCLUSIONS

In this paper we have proposed a proof-of-concept prototype for the realtime reconstruction of stylistic gait motion which outputs a subset of motion dimensions given the other dimensions of the same motion as input. The stylistic control of the synthesis is performed thanks to a realtime recognition of the motion and of its state-by-state evolution, combined with a covariance-based stylistic mapping adapted from speech processing. This mapping function modifies the synthesis model parameters in order to correspond to the incoming style. We have shown that an implicit stylistic control is possible since the output motion style could be controlled without any explicit tagging, directly from stylistic input. The present use case involves stylistic gait reconstruction, but we aim at testing it on other use cases in the near future, involving expert gestures, like dance, craftsmanship, musical gestures, and combined modalities such as the control of sound based on a realtime implicit mapping from stylistic gestures. Other use cases such as the synthesis of smooth optical mocap like data driven by noisier Kinect skeleton data are also envisioned. Future work also involves the testing of different realtime approaches to Viterbi decoding, as well as the testing of different mapping strategies. We are also planning the realisation of user studies.

ACKNOWLEDGEMENTS

N. d'Alessandro is funded by a regional fund called Région Wallonne FIRST Spin-Off. J. Tilmanne and T. Ravet are supported by the European Union, 7th Framework Programme (FP7-ICT-2011-9), under grant agreement no 600676 (i-Treasures project). M. Astrinaki is supported by a PhD grant funded by UMONS and Acapela Group.

REFERENCES

[1] Animazoo. IGS-190. `http://www.animazoo.com`, 2008.

[2] K. Barczewska and A. Drozd, 'Comparison of methods for hand gesture recognition based on Dynamic Time Warping algorithm', in *Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on*, pp. 207–210, (2013).

[3] F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, and N. Rasamimanana, 'Continuous realtime gesture following and recognition', *Gesture in embodied communication and human-computer interaction*, 73–84, (2010).

[4] M. Brand and A. Hertzmann, 'Style machines', in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 183–192, (2000).

[5] B. Caramiaux and A. Tanaka, 'Machine Learning of Musical Gestures', in *Proceedings of the 13th Conference on New Interfaces for Musical Expression (NIME'13)*, (2013).

[6] N. d'Alessandro, J. Tilmanne, M. Astrinaki, T. Hueber, R. Dall, T. Ravet, A. Moinet, H. Cakmak, O. Babacan, A. Barbulescu, V. Parfait, V. Huguenin, E.S. Kalaycı, and Q. Hu, 'Reactive Statistical Mapping: Towards the Sketching of Performative Control with Data', *IFIP Advances in Information and Communication Technology (AICT)*, to appear, (2014).

[7] T. Hueber, G. Bailly, and B. Denby, 'Continuous Articulatory-to-Acoustic Mapping using Phone-Based Trajectory HMM for a Silent Speech Interface', in *Proceedings of Interspeech, ISCA*, (2012).

[8] M. Lau, Z. Bar-Joseph, and J. Kuffner, 'Modeling spatial and temporal variation in motion data', in *ACM SIGGRAPH Asia 2009 papers*, pp. 171:1–171:10, New York, NY, USA, (2009). ACM.

[9] S. Mitra and T. Acharya, 'Gesture Recognition: A Survey', in *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, pp. 311–324, (2007).

[10] G. W. Taylor and G. E. Hinton, 'Factored Conditional Restricted Boltzmann Machines for Modeling Motion Style', in *Proc. 26th International Conference on Machine Learning*, pp. pp 1025–1032, (2009).

[11] J. Tilmanne, N. d'Alessandro, M. Astrinaki, and T. Ravet, 'Exploration of a Stylistic Motion Space Through Realtime Synthesis', in *Proceedings of the 8th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP'14)*, (2014).

[12] J. Tilmanne, A. Moinet, and T. Dutoit, 'Stylistic gait synthesis based on hidden Markov models', *Eurasip journal on Advances in Signal Processing*, **2012:72**(1), 1–14, (2012).

[13] J. Tilmanne and T. Ravet. The Mockey Database. `http://tcts.fpms.ac.be/~tilmanne/`, 2010.

[14] Tokuda et al. HMM-Based Speech Synthesis System (HTS). `http://hts.sp.nitech.ac.jp`, 2008.

[15] N. F. Troje, 'Decomposing biological motion: A framework for analysis and synthesis of human gait patterns', *Journal of Vision*, **2**(5), 371–387, (2002).

[16] University of Cambridge. The Hidden Markov Model Toolkit (HTK). `http://htk.eng.cam.ac.uk`, 2009.

[17] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell, 'Hidden conditional random fields for gesture recognition', in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pp. 1521–1527, (2006).

[2] The influence of upcoming states is lost, but using past frames still ensures that the transitions between frames remain smooth, although a small degradation might be observed compared to using non-zero values of F.