Analysis of dance movements using Gaussian processes

Antoine Liutkus, Angélique Drémeau, Dimitrios Alexiadis, Slim Essid, Petros Daras

Abstract

This work addresses the Huawei/3DLife Grand Challenge, presenting a novel method for the analysis of dance movements. The approach focuses on the decomposition of the dance movements into elementary motions. Placing this problem into a probabilistic framework, we propose to exploit Gaussian processes to accurately model the different components of the decomposition. The preliminary results, presented in this paper, are very promising. In particular, two applications are considered, illustrating the relevance of the proposed approach, namely the correction of tracking errors and the smoothing of some movements of the teacher to help toward the dance learning.

Index Terms

Grand Challenge, 3DLife, dance analysis, interactive environments, Gaussian process.

I. Introduction

The Huawei/3DLife Grand Challenge[1] focuses on a dance class scenario, where dance lessons are given online by an expert Salsa-dance teacher. In particular, both the teacher's and the students' performances are to be automatically analyzed and rendered using their respective avatars in an online virtual dance room.

Within this scenario, one important task involves recognizing the dance steps and movements executed by a student performing a given choreography. This enables the teacher agent to detect possible mistakes and suggest corrections. This task can be performed using a decomposition of dance gestures into

A. Liutkus, A. Drémeau and S. Essid are with Institut Télécom, Télécom ParisTech, CNRS LTCI, France. D. Alexiadis and P. Daras are with Centre for Research and Technology - Hellas, Information Technologies Institute, Thessaloniki, Greece.

[1] http://www.acmmm12.org/3dlife-huawei-challenge-realistic-interaction-in-online-virtual-environments/

elementary motions, in order to e.g. facilitate their classification among a reference set of choreographies. Some relevant contributions have been recently presented in the literature. In [11], [10], the authors introduce a music-based temporal segmentation of the motion, resulting in a sequence of "primitive motions". In [7], a frequency-based representation of the dance motion is obtained by means of a wavelet decomposition. Finally a segmental singular-vector-decomposition (SVD) is used in [2] to perform a hierarchical decomposition of the dance.

In this paper, we show how recent works on Gaussian Processes (GP) [4] can be directly applied to decompose a dance motion as a sum of separate *latent components*. Each of these latent components stands for a movement having its own spatio-temporal characteristics. In the context of dance performance analysis, one of these could for example correspond to a repetitive pattern at the measure level in the music, while another would account for a repetitive pattern at the beat level. Compared to the existing approaches, this framework offers some desirable features. Firstly, it allows to easily include both temporal and spatial structures into the modeling. Secondly, it does not assume exact periodicity of the components nor does it rely on parametric models for the estimation. Thirdly, it readily permits to consider new components into the model, thus demonstrating the significant expressive power of the approach. Indeed, as many other components as desired may be added, for example accounting for a slowly varying average position of the body or for temporally unpredictable movements. Finally, the proposed methodology provides a Minimum Mean Squared Error (MMSE) estimate of the components given the observations.

The Huawei/3DLife Grand Challenge is accompanied by a rich multimodal data set, recorded by a network of video cameras, multiple microphones, Microsoft Kinects, inertial measurement units, etc. [3]. While it can be interesting to jointly exploit all sensors, it is also of primary importance to consider the case where students do not have access to a complete acquisition platform. In this paper, we focus on the exploitation of the Microsoft Kinect data emphasizing the fact that the Kinect sensor is becoming increasingly popular.

The rest of the paper is organized as follows. In Section II, we present the data considered. The GP model, proposed for data analysis, is described in detail in Sections III and IV. Then, in Subsection V-A we present how sensible hyperparameters for this model can be chosen based on both musicological considerations and audio analysis methods. In Subsection V-B we demonstrate that the proposed approach allows to successfully decompose complex performance as a sum of simpler movements. Finally, we draw conclusions as well as some perspectives for future work in Section VI.

II. Data and general model

A. 3D Skeletons

The Microsoft Kinect depth maps are exploited by means of the OpenNI SDK[2]. The API provides the ability to track 17 3D skeletal joint positions (Head, Neck, Torso, Collar bones, Shoulders, Elbows, Wrists, Hips, Knees and Feet) for each video frame, along with the corresponding tracking confidence level.

A limiting feature of the OpenNI skeleton tracking is that it requires user calibration beforehand. This is achieved by having the user standing in a specific pose for several seconds while an individual skeleton is calibrated for the user. In a real world context however, it is not assured that a dancer will perform this calibration pose correctly, for the required time or may not perform it at all. To overcome this limitation, pre-computed custom calibration data was used by manually sourcing and calibrating persons with body characteristics similar to each dancer in the data set.

B. General model formulation

Let j denote one of the $N_j = 17$ joints, c one of the $N_c = 3$ spatial coordinates and t one of the N_t frames. The value of the c^{th} coordinate of joint j at time t is written $z\left(j,c,t\right) \in \mathbb{R}$. Gathering all sets of variables (j,c,t) into an *input space* $\mathcal{X} \triangleq \{x | x = (j,c,t)\}_{j,c,t}$, one can express $z\left(x\right)$ as $z\left(j,c,t\right)$. In this study, we propose to model the observed signal z as the sum of an *average position* $z_0\left(j,c\right)$ of the body and a *movement* term $\bar{z}\left(j,c,t\right)$, understood as the deviation from the equilibrium z_0:

$$\forall \left(j,c,t\right) \in \mathcal{X}, \quad z\left(j,c,t\right) = z_0\left(j,c\right) + \bar{z}\left(j,c,t\right), \tag{1}$$

where $z_0\left(j,c\right) \triangleq \mathbb{E}_t\left[z\left(j,c,t\right)\right]$.

Furthermore, we assume that the movement term $\bar{z}\left(j,c,t\right)$ is well modeled as the sum of M signals y_m called *latent components*:

$$\forall \left(j,c,t\right) \in \mathcal{X}, \quad \bar{z}\left(j,c,t\right) = \sum_{m=1}^{M} y_m\left(j,c,t\right). \tag{2}$$

Considering this model, we focus on the estimation of the latent components y_m given the observation of z. This problem is reminiscent of those encountered in the field of *source separation* [1]. In this study, we will concentrate on a GP model for the latent components, which has recently been introduced in source separation as a convenient way to include spatio-temporal knowledge on the sources into the modeling [4], [5].

[2] `http://www.openni.org`

III. Probabilistic framework

A. Gaussian processes

We consider that all latent components y_m introduced in (2) are independent Gaussian processes. A GP [6], [8], [9], [12] is completely determined by its mean and covariance functions. Let μ_m and k_m denote the mean and covariance functions of the GP y_m, respectively. Since the y_m are assumed to model deviations from equilibrium, we will consider their means to be 0, that is $\mu_m(x) = 0$, $\forall x \in \mathcal{X}$. To define k_m, we focus on one particular simple hyperparameterization , which proved to be very efficient for our purpose. We consider all k_m to be *separable*, that is:

$$k_m\big((j,c,t),(j',c',t')\big) = k_m^{(S)}\big((j,c),(j',c')\big)\, k_m^{(T)}\big(t,t'\big), \tag{3}$$

where $k_m^{(S)}$ is a *spatial* covariance function which involves only the (j,c) dimensions while $k_m^{(T)}$ is a *temporal* covariance function involving only time.

Since the deviation $\bar{z}(j,c,t)$ from body equilibrium $z_0(j,c)$ is modeled as a sum of independent GPs, it is itself a GP whose covariance function is the sum of the covariance functions of the components:

$$\bar{z} = \sum_{m=1}^{M} y_m \sim \mathcal{GP}\left(0, \sum_{m=1}^{M} k_m\left(x,x'\right)\right). \tag{4}$$

B. Estimation problem

Considering the model (2)-(4), we focus on the problem of estimating any component y_{m_0} of interest given the mixture z (or equivalently $\bar{z}$).

Let $\mathbf{y}_m = [y_m(x_1)\cdots y_m(x_n)]^\top$ and $\bar{\mathbf{z}} = [\bar{z}(x_1)\cdots\bar{z}(x_n)]^\top$ with $n = N_j \times N_t \times N_c$. We then define the covariance matrix of component $\mathbf{y}_m$, say K_m, as $[K_m]_{pl} = k_m(x_p, x_l)$. The joint distribution of the component of interest $\mathbf{y}_{m_0}$ and the mixture $\bar{\mathbf{z}}$ is expressed as:

$$\begin{bmatrix} \bar{\mathbf{z}} \\ \mathbf{y}_{m_0} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sum_{m=1}^{M} K_m & K_{m_0} \\ K_{m_0} & K_{m_0} \end{bmatrix}\right). \tag{5}$$

Given (4)-(5), a classic result (see [8]) is that the distribution $p(\mathbf{y}_{m_0} \mid \bar{\mathbf{z}})$ of $\mathbf{y}_{m_0}$ given $\bar{\mathbf{z}}$ is Gaussian with mean $\overline{\mathbf{y}}_{m_0}$ such as[3]:

$$\overline{\mathbf{y}}_{m_0} = K_{m_0}\left[\sum_{m=1}^{M} K_m\right]^{-1} \bar{\mathbf{z}}. \tag{6}$$

[3]In the case of a singular covariance matrix $\sum_{m=1}^{M} K_m$, numerical methods such as Moore-Penrose pseudo-inversion may be used.

Since this posterior distribution is Gaussian, the minimum mean square error (MMSE) estimate of $\mathbf{y}_{m_0}$, say $\hat{\mathbf{y}}_{m_0}$, is thus $\hat{\mathbf{y}}_{m_0} = \overline{\mathbf{y}}_{m_0}$.

It is interesting to note that the main computationally expensive operation in (6), the inversion of $\sum_m K_m$, is only required once for the extraction of all the latent components. Naively implemented, the computational cost of this extraction method is $\mathcal{O}\left(n^3\right)$. Still, efficient techniques as presented for example in [4] may be used to significantly reduce the complexity.

IV. Temporal and spatial covariance functions

In this section, we specify the expressions of the temporal and spatial covariance functions, resp. $k_m^{(T)}$ and $k_m^{(S)}$, used in (3).

A. Temporal covariance function

We choose for the temporal covariance function a form similar to the one used in [4], [5]:

$$k_m^{(T)}\left(t,t'\right) = \sigma_m^2 \exp\left(-\frac{2\sin^2\frac{\pi(t-t')}{T_m}}{l_m^2} - \frac{(t-t')^2}{2\lambda_m^2}\right). \tag{7}$$

This pseudo-periodic function allows modeling of complex periodic signals with only four hyperparameters:

- σ_m, representative of the *magnitude* of the considered component,
- T_m, the *period* of the function,
- l_m, the *intra-period lengthscale* which controls the smoothness of the signal within each period,
- λ_m, the *absolute lengthscale* which guarantees the independence of very distant samples and thus enables the assumed periodicity assumption to only hold locally.

A remarkable advantage of this model is that it allows for an intuitive interpretation while being of simple expression.

B. Spatial covariance function

The spatial covariance $k_m^{(S)}\left((j,c),(j',c')\right)$ in (3) indicates the structural covariance of two joint-coordinates (j,c) and (j',c'). Some considerations could be introduced here to set this spatial covariance sensibly with respect to a dynamical model involving the connections between the joints. We choose here a different strategy, based on the use of the MMSE estimates of the components.

The spatial covariance function for the component y_m between (j, c) and (j', c') can indeed be simply estimated as the empirical correlation coefficient between $y_m(j, c, t)$ and $y_m(j', c', t)$:

$$k_m^{(S)}\big((j,c),(j',c')\big) = \frac{\mathbb{E}_t\left[y_m(j,c,t)\, y_m(j',c',t)\right]}{\sqrt{\mathbb{E}_t\left[y_m(j,c,t)^2\right]\mathbb{E}_t\left[y_m(j',c',t)^2\right]}}. \tag{8}$$

We thus propose a short iterative procedure, which, at each iteration, computes $\hat{y}_m$ according to (6), and updates the corresponding spatial covariance functions. In practice, we set the number of iterations to 2: at the first iteration, the latent components are estimated assuming all (j, c) are independent, that is $k_m^{(S)}((j,c),(j',c')) = \delta_{jj'}\delta_{cc'}$; given these first estimates, the second iteration updates $k_m^{(S)}$ as in (8) and leads to the final estimates of y_m.

V. Experiments

A. Hyperparameters assessment

In this section, we discuss the choice of the hyperparameters presented in Section IV-A, σ_m, T_m, l_m and λ_m $\forall m$.

We assume a decomposition of the dance motions into four latent components: the first two are repeating patterns, the others correspond respectively to slowly varying movements of the body and to unpredictable, non-repetitive movements. For each of these components, one particular set of hyperparameters has to be defined.

Since the movement $\bar{z}$ corresponds to a dance, it is highly likely that some kind of synchronization exists between the movement and the music. In order to correctly assess the period of the pseudo-periodic components, we use the manual annotations of the music beats provided in the Huawei/3DLife Grand Challenge data set and assume that the first component would have a period of $T_1 = 8$ beats, while the second one would have a period of $T_2 = 4$ beats. This assumption is sensible in the case of salsa movements. For the last two non-periodic components, we set $T_{\{3,4\}} = \infty$, in order to eliminate the sine part.

The intra-period lengthscale l_m has a natural interpretation as the stability of the signal within each period. It is useless in the case of non-periodic components, *i.e.* the 3rd and the last ones, for which $T_{\{3,4\}} = \infty$. For the repeating patterns, we experimentally set $l_1 = l_2 = 0.1$.

The absolute lengthscale λ_m permits to introduce pseudo-periodicity instead of strict periodicity that may badly fit the data. For the two pseudo-periodic components, we simply expressed it as a multiple of the period, thus assuming that each periodic component may only be self-similar for some consecutive

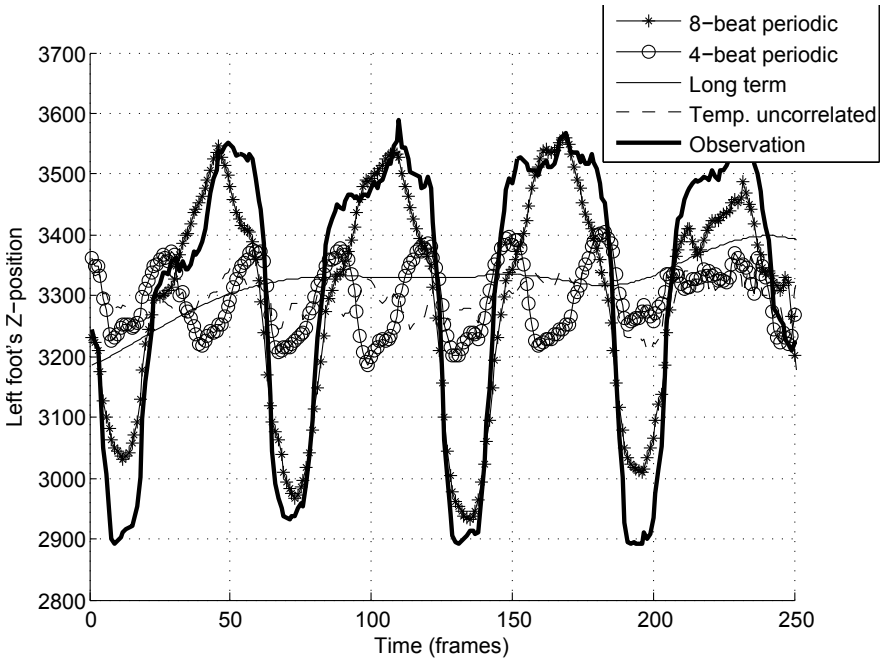

Fig. 1. Decomposition of the Z-position of "Thomas' left foot" in choreography c3 into four components: 8-beat periodic, 4-beat periodic, long-term and temporally uncorrelated components.

periods. We set in particular $\lambda_m = 2 \times T_m$, $m \in \{1, 2\}$. For non-periodic components, this parameter was set to some seconds for the slowly varying component and was on the contrary brought to zero for the component accounting for unpredictable moves. In that last case, the temporal covariance function reduces to:

$$k_4\left(t, t'\right) = \sigma_4^2 \delta_{tt'}. \tag{9}$$

Finally, the σ_m parameter corresponds to the expected magnitude of the m^{th} component within the movement. In this study, we simply set all components to have the same magnitude, except for the unpredictable component, having a covariance function as in (9), whose magnitude was supposed to be less than that of other components. More precisely, we set $\sigma_m = 2$, $m \in \{1, 2, 3\}$ and $\sigma_4 = 0.5$.

B. Decomposition of the dance motion

As we previously mentioned, the proposed decomposition is interesting, among others, because of the intuitive interpretation of the constituting components.

To illustrate this assertion, Figure 1 presents the decomposition of the Z-position of "Thomas's left foot" for the ten first seconds of choreography c3. In this figure, we can clearly distinguish the contributions

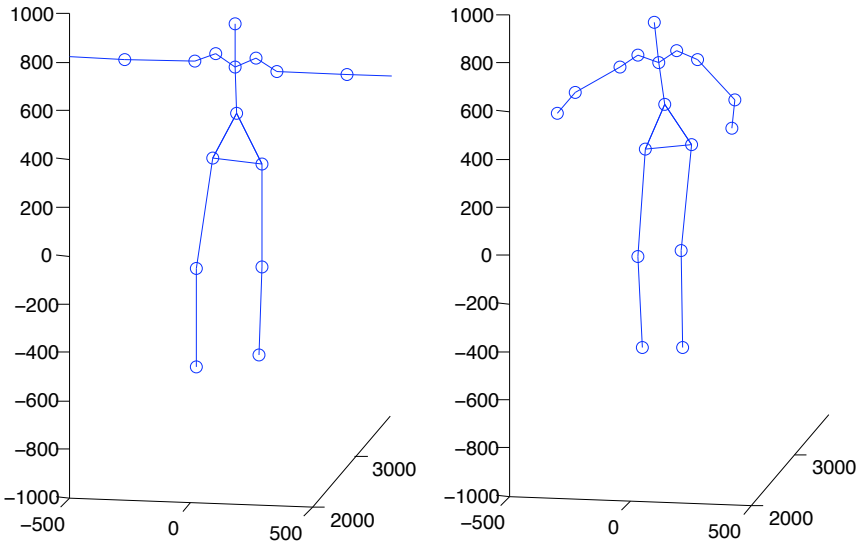

Fig. 2. Frame of the 3D skeleton video of Thomas for choreography c3: left, original data and right, approximated data.

of the four components: as expected, the 8-beat periodic component accounts for the periodicity of the salsa steps, that is, it follows the large periodic fluctuations of the movements; this first component is coupled with the 4-beat periodic component which catches the small periodic fluctuations; the long term component can be interpreted as the general displacements of the considered joint, while the last component stands for the "uncorrelated" information. The video sequences of the different components are available online[4].

Such a decomposition can be favourably exploited in many tasks of dance analysis. For the proof of concept, we consider here the problem of smoothing the "superfluous points" of the observations. Such points may occur due to tracking errors or mistakes made by the dancer him/herself. Taking them into account, interpretation mistakes can happen in the gesture recognition task, possibly resulting in an inadequate feedback from the teacher's agent. On the other hand, smoothing some movements of the teacher can help the student to learn a choreography more easily.

Considering the proposed decomposition, the last component, qualified as "temporally uncorrelated", can be seen as a noise added to a "perfect" movement. Thus, substracting this component from the initial dance motion leads to a "denoised" motion. Figure 2 presents one frame of the original and approximated 3D skeleton sequences of "Thomas" for choreography c3. The complete video is available online[4]. On

[4] http://3dlife-huawei-gc-submission.blogspot.fr/

this video, we can observe that the tracking errors have been smoothed over all frames. In particular, Thomas' arms are presented in narrowed, more realistic positions.

As mentioned, we can also consider a different application, where some movements of the teacher are simplified to help the student. In this application, the objective is not to obtain denoised, corrected dance movements, but to deliberately reduce the complexity of some movements. To this purpose, we set $l_1 = l_2 = 10$ to force the stability of the periodicities of the first two components. In the website[4], we give the example of "Anne-Sophie-k" whose arms' movements have been decomposed and "smoothed". Note that this example illustrates the possibility to consider different decompositions, depending on the joints of interest and/or the intended application.

VI. Conclusion

In this paper, we have proposed a method for decomposing dance movements into elementary motions. The approach relies on Gaussian processes allowing for a flexible representation, from extremely coarse to detailed, capturing the periodicities of the dance movements.

The preliminary results are promising, offering some nice perspectives. In particular, we intend to address the task of gesture recognition, which could benefit from the good approximations obtained by considering only three of the four components of the decomposition.

VII. Acknowledgements

This research was supported by the European Commission under contract "FP7-247688 3DLife".

References

[1] P. Comon and C. Jutten, editors. *Handbook of Blind Source Separation: Independent Component Analysis and Blind Deconvolution*. Academic Press, 2010.

[2] L. Deng, H. Leung, N. Gu, and Y. Yang. Recognizing dance motions with segmental svd. In *Int'l Conference on Pattern Recognition*, pages 1537–1540, 2010.

[3] S. Essid, X. Lin, M. Gowing, G. Kordelas, A. Aksay, P. Kelly, T. Fillon, Q. Zhang, A. Dielmann, V. Kitanovski, R. Tournemenne, A. Masurelle, E. Izquierdo, N. E. O'Connor, P. Daras, and G. Richard. A multi-modal dance corpus for reseach into interaction between humans in virtual environments. Accepted for publication in Journal on Multimodal User Interfaces, Special Issue on Multimodal Corpora, Springer, 2012.

[4] A. Liutkus, R. Badeau, and G. Richard. Gaussian processes for underdetermined source separation. *IEEE Transactions on Signal Processing*, 59(7):3155 –3167, July 2011.

[5] A. Liutkus, R. Badeau, and G. Richard. Multi-dimensional signal separation with gaussian processes. In *Proc. of IEEE Conf. on Statistical Signal Processing (SSP2011)*, Nice, France, 2011.