PROBABILISTIC DANCE PERFORMANCE ALIGNMENT BY FUSION OF MULTIMODAL FEATURES

Angélique Drémeau, Slim Essid

Institut Telecom, Telecom ParisTech, CNRS-LTCI, Paris, France

ABSTRACT

This paper presents a probabilistic framework for the multimodal alignment of dance movements. The approach is based on a Hidden Markov Model (HMM) and considers different feature functions, each corresponding to a particular modality, namely motion features, extracted from depth maps, and audio features, extracted from audio recordings of dancers' steps. We show that this approach allows performing accurate dancer alignment, while constituting a general framework for various multimodal alignment tasks.

Index Terms— Multimodal alignment, Hidden Markov Model, dance gestures

1. INTRODUCTION

This work is concerned with the analysis of a challenging type of human activity that is dance performance. More particularly, we consider here a virtual dance class scenario (inspired by the 3DLife ACM Multimedia Grand Challenge [1]) where dance lessons are given online by an autonomous virtual agent who acts as a dancer teacher and is expected to be able to automatically detect possible mistakes of the students and suggest corrections. Hence, in such a scenario, one important task involves recognizing the dance steps executed by a student mimicking the teacher's choreography, and, intimately linked to this, aligning the dance movements of the student with those of the teacher. In this paper, we propose a novel dance alignment method allowing for the fusion of different streams of features extracted from modalities of different nature.

Among the numerous works interested in the alignment problem, only a few deal with the particular application of gesture alignment and even fewer focus on multimodal gesture alignment. Thus, in [2] and [3], multidimensional Dynamic Time Warping (DTW, introduced in [4]) is used for gesture recognition and classification, respectively. In these works, the term "multidimensional" refers to the size of the feature vectors (namely 3D positions), extracted from the same modality (visual and depth cameras, respectively). The multidimensional DTW is further improved in [5] for the gesture recognition using 3D skeletons stemming from depth maps. In that work, the authors argue that not all skeletal joints do participate equally to the alignment, some of them would not be relevant, even leading to alignment mistakes. They thus propose to weight each joint contribution within the DTW framework. Multimodality is finally considered in [6] where the authors propose a general approach to align data composed of two asynchronous multidimensional modalities.

Parallel to these DTW-based deterministic methods, probabilistic approaches find an equal success. Along this line of research, most contributions consider, in particular, Hidden Markov Models (HMM). As examples, we can cite [7] where the authors are interested in the alignment of image sequences within the context of human action recognition, and [8] which, as a probabilistic counterpart of [6], introduces a method to align asynchronous multidimensional and multimodal sequences with a view to audio-visual speech recognition. Motivated by their common application (recognition), both contributions exploit HMMs to describe different classes of observations. Thus one HMM is trained for each targeted action, then the alignment is performed through a model detection.

In this paper, we propose to directly model the observations (*i.e.,* the student's dance sequence) as functions of reference data (*i.e.,* the teacher's dance sequence).

This paper is organized as follows. In Section 2, we present the data considered in our alignment problem. We then describe our probabilistic framework in Section 3, before showing some results illustrating the good performance of the approach.

2. DANCE PERFORMANCE DATASET

In this paper, we consider the 3DLife ACM Multimedia Grand Challenge 2012 dataset [1] which consists of 15 multimodal recordings of Salsa dancers performing between 2 to 5 fixed choreographies, captured by a variety of sensors. Each multimodal recording contains multi-channel audio, video from multiple viewpoints, Wireless Inertial Measurement Unit (WIMU) sensor data, Microsoft Kinect depth maps and original music excerpts. For each of these modalities, a time-stamp is available, allowing us to synchronize them in a pre-processing step (we thus here differ from the approaches [8, 6] where the considered signals are asynchronous).

We propose here to focus on the use of two different

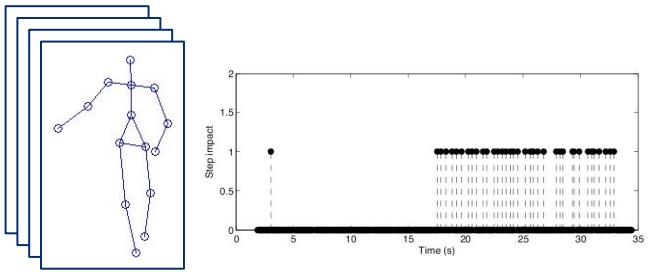


Fig. 1. Multimodal data considered for the alignment of the dancers: 3D skeletons extracted from Kinect depth maps (left) and step impacts extracted from piezoelectric transducers (right).

modalities, namely the Kinect depth maps and the piezoelectric transducers (audio channels 1, 2, 17 and 18 of the data). Note however that the presented approach can easily be extended to many other modalities (see Section 3).

The Microsoft Kinect depth maps are exploited by means of the Matlab SDK developed by Dirk-Jan Kroon and available on the Matlab website [9]. The code permits to track 15 3D skeletal joint positions (head, neck, torso, left and right shoulders, elbows, hands, hips, knees and feet) for each video frame.

From the onfloor piezoelectric transducers we detect the step impacts following the same procedure as in [10], by applying a one-class Support Vector Machine (SVM) [11] to feature vectors made up of the concatenation of onset detection functions. The step detection results are then subsampled at the Kinect frame rate (*i.e.,* 30 frames per second).

Figure 1 illustrates the data considered for the proposed alignment method.

3. PROBABILISTIC FRAMEWORK

We aim at aligning the movements of two dancers, a teacher and his/her student. To this end, we propose to resort to a particular probabilistic model. We expose this model in Subsection 3.1 and formalise the alignment problem as a maximum a posteriori detection problem in Subsection 3.2.

3.1. Model

We define the following notations. Let $\{\mathbf{K}_m\}_{m\in\{1,\dots,M\}}$ and $\{s_m\}_{m\in\{1,\dots,M\}}$ denote respectively the set of 3D skeleton positions and step impacts, of the *student* over time (M is thus the total number of data frames related to the student). More precisely, $\forall m \in [\![1, M]\!]$, $\mathbf{K}_m = [k_m(i,j)]_{(i,j)}$ is a real-valued 3×15 matrix corresponding to the 3 coordinates of the 15 skeleton joints, while s_m is equal to 1 if a step impact is detected, 0 otherwise; $\mathbf{K}_m$ and s_m are gathered into the variable $\mathbf{y}_m$[1]. The same variables are also considered for the *teacher*, but distinguished by the exponent "ref". We finally denote by N the number of frames of the teacher data. Note that, in general, $M \neq N$.

$\forall m \in [\![1, M]\!]$, the observation $\mathbf{y}_m$ is assumed to obey a particular model indexed by the variable $x_m \in [\![1, N]\!]$:

$$p(\mathbf{y}_m|x_m = n) = p(\mathbf{K}_m|x_m = n)\; p(s_m|x_m = n), \quad (1)$$
$$\text{with} \quad p(\mathbf{K}_m|x_m = n) \propto \exp(-\mu_1 f_1(\mathbf{K}_m, \mathbf{K}_n^{\text{ref}})),$$
$$p(s_m|x_m = n) \propto \exp(-\mu_2 f_2(s_m, s_n^{\text{ref}})),$$

where $\mu_1, \mu_2 > 0$. The functions f_1 and f_2 are feature functions. They are here defined by

$$f_1(\mathbf{K}_m, \mathbf{K}_n^{\text{ref}}) = ||\mathbf{K}_m - \mathbf{K}_n^{\text{ref}}||_F, \quad (2)$$
$$f_2(s_m, s_n^{\text{ref}}) = |s_m - s_n^{\text{ref}}|, \quad (3)$$

where $||.||_F$ stands for the Frobenius norm.

$\mathbf{K}_m$ and s_m are thus assumed to be independent given x_m. We assume that the labels x_m are linked through a first-order Markov chain with N states. Hence, the chain states are defined to be the set of all teacher's frames and each observation $\mathbf{y}_m$ is assumed to be generated by one of these N states. The transition probabilities are parametrised as follows: $\forall m \in [\![2, M]\!]$, $\forall (i, j) \in [\![1, N]\!]^2$,

$$p(x_m = j|x_{m-1} = i) = \begin{cases} \lambda_0 & \text{if } j = i, \\ \lambda_1 & \text{if } j = i + 1, \\ \lambda_2 & \text{if } j = i + 2, \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

$$p(x_1 = i) = 1/N, \quad (5)$$

where $\lambda_0, \lambda_1, \lambda_2 > 0$ and $\lambda_0 + \lambda_1 + \lambda_2 = 1$.

According to model (1)-(5), the student's movements are seen as noisy versions of the teacher's movements, considered as reference. The parameters μ_1 and μ_2 then stand for the deviation of the student's movements from those of the teacher. Parameters λ_0, λ_1 and λ_2 are merely the transition probabilities of the HMM. As per (4), only transitions to the next following two states are allowed, that is the sequence of student's movements is expected to be consistent with the one of the teacher. To allow for more degrees of freedom in the model, we could have considered a fourth parameter to express the possibility to "jump two states". However, this would lead to an unnecessary increase of the complexity in light of the performance achieved with the proposed model (see Section 4).

3.2. Detection problem

Within model (1)-(5), the alignment between the student's and teacher's movements is expressed as the estimation of

[1] When clear from the context, we will use indifferently the terms realization and variable.

	Cross-correlation	DTW	Model (1)-(5) using Kinect only	Model (1)-(5)
Well-aligned dance steps (%)	63	78	81	**84**
Average inaccuracy a (%) for well-aligned dance steps	24	**15**	17	**15**

Table 1. Average results of different alignment methods

the HMM states sequence. Formally, let $\mathcal{X}$ define the set of states, $\mathcal{X} \triangleq \{x_m\}_{m\in\{1,\dots,M\}}$, and $\mathcal{Y}$ the set of observed data, $\mathcal{Y} \triangleq \{\mathbf{y}_m\}_{m\in\{1,\dots,M\}}$. We focus on the following maximum a posteriori detection problem

$$\hat{\mathcal{X}} = \underset{\mathcal{X}}{\operatorname{argmax}}\, p(\mathcal{X}|\mathcal{Y}). \tag{6}$$

Problem (6) can be efficiently solved using a particular instance of the well-known sum-product algorithm, namely the Viterbi algorithm [12], of complexity $\mathcal{O}(MN)$.

The proposed approach offers some desirable properties:

- Through the insertion of additional feature functions, it allows naturally taking novel modalities into account without modifying the general formalism (see (1)). This property is interesting in situations where the dancers are not always recorded with the same set of capture devices (as envisaged for instance in the online virtual dance class scenario [1]). The proposed model can then easily adapt to different capture setups.

- It makes some model parameters (such as μ_1 and μ_2) explicit which can be tuned to match the data. If μ_1 and μ_2 are defined to be varying with respect to the frame number n, we recognize the idea of some *weighted* deterministic approaches, as the Weighted DTW proposed in [13].

4. EXPERIMENTS

In this section, we evaluate the performance of the proposed approach.

4.1. Evaluation dataset

Among the dance sequences made available in the 3DLife ACM Multimedia Grand Challenge 2012 dataset [1], we focus our experiments on the male dancers for which we performed a manual annotation of the dance steps[2] (see Subsection 4.3). We have thus 4 different choreographies ("c2" to "c5") at our disposal, each containing between 10 and 12 dance steps. This leads to a set of 534 dance steps to align with those of the teacher, "Bertrand".

[2]In future work, these ground-truth annotations will be extended to the female dancers as well.

4.2. Description of the compared alignment methods

We evaluate and compare four different alignment methods:

- "Cross-correlation" estimates the time-shift between the dancing sequences to compare, by finding the time-lag $\hat{m}$ that maximizes the cross-correlation between teacher and student joint sequences. Formally it writes

$$\hat{m} = \underset{m}{\operatorname{argmax}} \sum_i \sum_j \sum_{n=0}^{N-m-1} k_{n+m}(i,j)\; k_n^{\text{ref}}(i,j).$$

- "Dynamic Time Warping" corresponds to a multidimensional Dynamic Time Warping (DTW) (similar to [2]), using the Frobenius norm as distance measure.

- "Model (1)-(5) using Kinect data only" relies on the probabilistic model (1)-(5) in which the variables corresponding to the step impacts are not taken into account. Equation (1) is thus rewritten as

$$\begin{aligned} p(\mathbf{y}_m|x_m = n) &= p(\mathbf{K}_m|x_m = n), \\ &\propto \exp(\mu_1 f_1(\mathbf{K}_m, \mathbf{K}_n^{\text{ref}})). \end{aligned}$$

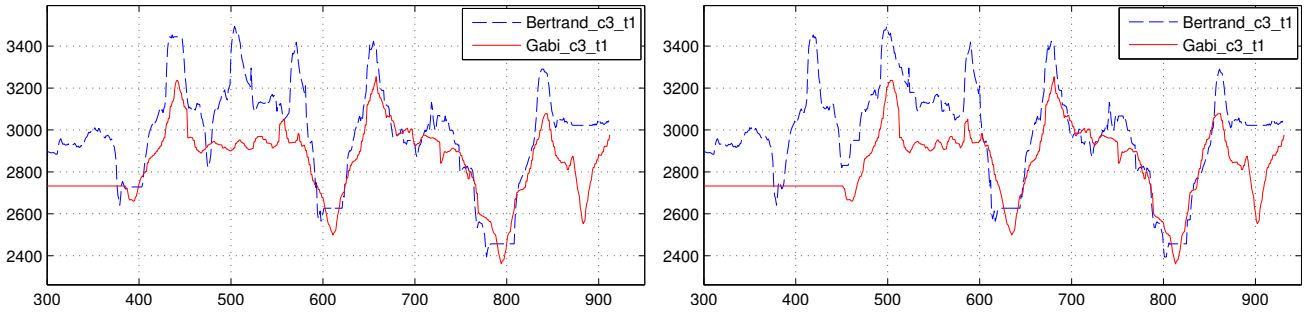
- "Model (1)-(5)" corresponds to the proposed approach, exploiting both Kinect and piezoelectric data.

Three of these methods are multidimensional methods and deal with Kinect data only; the last one corresponds to the model (1)-(5) (exploiting both Kinect and piezoelectric data). For the two latter methods, we arbitrarily set $\lambda_0 = \lambda_1 = \lambda_2 = 1/3$, and, after preliminary testing, $\mu_1 = 10^{-3}$ and $\mu_2 = 1$, in order to strengthen the relevance of the piezoelectric data.

4.3. Objective evaluation of the alignment accuracy

In order to achieve an objective evaluation of the alignment accuracy, all the dance recordings considered have been manually annotated. The step annotations are labels defined by a start time (ST) and an end time (ET). After alignment, the start time of each student step annotation is compared to the nearest start time of the teacher annotation set. *If the annotations match*, we consider that the current step has been well-aligned and we compute a measure of inaccuracy a of the alignment as

$$a = \frac{100 \times d}{d + dist(\text{ST student annot., ST 2nd nearest teacher annot.})},$$



(a) Model (1)-(5) using Kinect data only

(b) Model (1)-(5)

Fig. 2. Alignment between the right knee Z-positions of the teacher "Bertrand" (dotted line) and the student "Gabi" (continuous line) in choreography "c3", with two different methods. The x-axis corresponds to the frame number of the aligned sequences.

where $d = dist(\text{ST student annot., ST nearest teacher annot.})$ and $dist$ stands for a distance expressed in number of frames. Note that a is zero in the best case (perfect alignment), and equal to 50% in the worst case (for the steps considered as "well-aligned"). We thus have at our disposal a relative measure allowing for an objective comparison of the different alignment methods.

Table 1 presents the percentage of well-aligned dance steps and the inaccuracy a of the alignment, averaged on the dataset, for the four considered alignment methods. We can here observe that, among the four considered methods, the best alignment is performed by the proposed multimodal approach. Note that DTW adopts the strategy of missing some steps but accurately aligning the remaining ones: it presents the same inaccuracy measure as the proposed approach.

4.4. Discussion on the multimodality approach

Figure 2 presents an example of aligned sequences with "Model (1)-(5) using Kinect data only" (Figure 2(a)) and "Model (1)-(5)" (Figure 2(b)), to "visually" illustrate the relevance of the multimodal approach. On this example, the student "Gabi" does not perform the entire choreography "c3" (some dance steps are missing at the beginning). His movements are compared to those of the teacher "Bertrand" for the same choreography. To make easier the visualization of the aligned sequences, we proceed as follows: if the student presents a delay with regard to the teacher, the teacher "waits" for the students, *i.e.,* new "virtual" frames are added in his movement sequence by repeating his latest position, until the student catches up; conversely, if the student is in advance, he "waits" for the teacher, *i.e.,* his latest position is repeated until the teacher catches up. Consequently, the frame numbers of the aligned sequences are the same but can differ from an alignment method to another. On this example, the proposed approach (Figure 2(b)) presents a perfect alignment, while its "monomodal" counterpart misses one step. Thanks to the integration of the step impacts into the alignment process, the multimodal approach significantly improves the performance. Note that this improvement comes with a lesser increase of the computational complexity since the only additional cost to consider lies in the evaluation of the features (see equations (1)-(3)).

5. CONCLUSION

In this paper, we have presented a probabilistic framework for the multimodal alignment of dance gestures. The proposed approach exploits on the one hand a HMM and on the other hand, feature functions able to take into account different modalities. This approach is proved to be efficient on a large dataset of multimodal recordings of different Salsa dancers with different expertise. Moreover, its structure makes it particularly flexible: more modalities can be easily taken into account by considering other feature functions.

6. ACKNOWLEDGEMENTS

This research was supported by the European Commission under contract "FP7-287723 REVERIE".

7. REFERENCES

[1] "3dlife/huawei acm mm grand challenge 2012," http://perso.telecom-paristech.fr/∼essid/3dlife-gc-12/.

[2] G.A. ten Holt, M.J.T. Reinders, and E.A. Hendriks, "Multi-dimensional dynamic time warping for gesture recognition," in *Proc. Conference of the Advanced School for Computing and Imaging*, 2007.

[3] M. Raptis, D. Kirovski, and H. Hoppe, "Real-time classification of dance gestures from skeleton animation," in *Proc. SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*, 2011.

[4] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 26, pp. 43–49, 1978.

[5] M. Reyes, G. Dominguez, and S. Escalera, "Feature weighting in dynamic time warping for gesture recognition in depth data," in *Proc. IEEE Int'l Conference on Computer Vision Workshops (ICCV)*, 2011.

[6] M. Wöllmer, M. Al-Hames, F. Eyben, B. Sculler, and G. Rigoll, "A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams," *Journal Neurocomputing*, vol. 73, pp. 366–380, 2009.

[7] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 1992.

[8] S. Bengio, "An asynchronous hidden markov model for audio-visual speech recognition," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2003.

[9] Dirk-Jan Kroon, "Kinect matlab code," http://www.mathworks.com/matlabcentral/fileexchange/30242-kinect-matlab.

[10] S. Essid, D. Alexiadis, R. Tournemenne, M. Gowing, P. Kelly, D. Monaghan, P. Daras, A. Drémeau, and N. E. O'Connor, "An advanced virtual dance performance evaluator," in *Proc. IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[11] B. Shölkopf and A. J. Smola, *Learning with kernels*, The MIT Press, Cambridge, MA, 2002.

[12] A. Viterbi, "Error bounds for convolutional code and an asymptotically optimum decoding algorithm," *IEEE Trans. On Information Theory*, vol. 13, pp. 260 – 269, 1967.

[13] Y.-S. Jeong, M. K. Jeong, and O. A. Omitaomu, "Weighted dynamic time warping for time series classification," *Pattern recognition*, vol. 44, pp. 2231–2240, 2011.