AN ADVANCED VIRTUAL DANCE PERFORMANCE EVALUATOR

Slim Essid[*], *Dimitrios Alexiadis*[†], *Robin Tournemenne* [*], *Marc Gowing*[‡], *Philip Kelly*[‡]
David Monaghan[‡], *Petros Daras*[†], *Angélique Drémeau* [*] *and Noel E. O'Connor*[‡]

ABSTRACT

The ever increasing availability of high speed Internet access has led to a leap in technologies that support real-time realistic interaction between humans in online virtual environments. In the context of this work, we wish to realise the vision of an online dance studio where a dance class is to be provided by an expert dance teacher and to be delivered to online students via the web. In this paper we study some of the technical issues that need to be addressed in this challenging scenario. In particular, we describe an automatic dance analysis tool that would be used to evaluate a student's performance and provide him/her with meaningful feedback to aid improvement.

Index Terms— Human activity analysis, dance analysis, multimodal processing, interactive environments.

1. INTRODUCTION

In this paper, we describe techniques for the automatic analysis of dance performances in view of the development of tutoring tools for online dance tutorial sessions. Unlike traditional gaming scenarios, when the motion of a user must be kept in synchronisation with a pre-recorded avatar that is displayed on the screen, the technique described in this paper targets online interactive scenarios where dance choreographies can be set, altered, practiced and refined by users. In such a scenario, a dance teacher is free to illustrate to online users choreography steps of their choice. After viewing the sequence at a later date, a student can attempt to mimic the steps, and obtain feedback from the system to help refine his/her dance moves. At any time, the teacher can alter the choreography or introduce extra steps when they receive notification that the student has reached a certain level of competency. As such, there is real online interaction between users.

Although it can be considered as a special instance of human activity analysis [1], dance performance analysis entails specific challenges owing to its artistic nature, which comprises complex cultural and cognitive dimensions, such as aesthetics and expressiveness [2]. Moreover, (for most dance styles) the analysis of a dancer's movements cannot be abstracted from the related music, as the steps and movements of the choreography are expected to be responses to particular musical events. This makes the evaluation of a dance performance a difficult problem that can benefit from multimodal approaches.

We find in literature some contributions dealing with this problematic. They distinguish from the proposed approach by two main aspects, namely the purpose of the dance analysis and the nature of the analysed signals. Most contributions consider dance analysis from an archiving and retrieving point of view (see *e.g.*, [3, 4]). The required information can thus be very different. As an example, in [3], the authors propose a method to convert a dance motion to symbolic descriptions (called "primitive motions"), in order to make easier its reconstruction. Our purpose is more ambitious. It postulates that we are able to extract informations susceptible to help dance students and assess their performances with regard to the teacher's one. The analysed signals are mostly made up of motion informations (kinematic sensors or video signals) sometimes helped by musical signals, namely the dance musics. In [5], only kinematic sensors are used to estimate the rhythm of the motion. A refinement of this method is proposed in [3] based on the consideration of musical rhythm. Music and video are jointly analysed in [6] to extract periodicities in the dance movements; the method combines then movement tracking and extraction of beat markers. Here, we make use of different signals, namely inertial measurements, Kinect depth maps and audio signals.

We consider the 3DLife ACM Multimedia Grand Challenge 2011 dataset [7] 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, synchronised video from multiple viewpoints, Wireless Inertial Measurement Unit (WIMU) sensor data, Microsoft Kinect depth maps and original music excerpts. In addition, the data corpus provides ground truth dancer reference ratings, as graded by expert dance teachers. In this work, our main aim is to automatically produce scores that are consistent with these reference ratings.

The rest of the paper is organised as follows: Section 2 provides both a high level overview of the approach and a description of the individual components. A student's dance recital is evaluated using two distinct quantitative scores; the first assesses the student's choreography, as described in Section 3; while the second metric focuses on providing feedback with respect to the students timing, as outlined in Section 4. Quantitative evaluation of the proposed scoring techniques are provided in Section 5. Finally, Section 6 provides conclusions and plans for future work.

2. SYSTEM OVERVIEW

A system flow chart of the overall system is presented in Figure 1. A dance "Score", used to assess an overall dance recital, is based on both choreography and timing metrics. In this work, Kinect sensors are used to acquire a "choreography" dance rating. This metric evaluates both the quality of bodily movements and the accuracy in that the student dancer achieves in executing the predefined dance steps. The second metric, "timing" is obtained using combined WIMU and audio data streams and assesses the dancers ability to keep in step with the dance teacher. Synchronisation between these two "timing" modalities are achieved by maximising the cross-correlation between specific WIMU and audio features around the hand-clap event that occurs at the beginning of each recording in the data corpus [8]. With respect to both scores, an assessment of a student's dance recital is realised by a comparison of their performance to that

[*]Institut Telecom / Telecom ParisTech, CNRS-LTCI - Paris, France
[†]Centre for Research and Technology - Hellas, Informatics and Telematics Institute, Thessaloniki, Greece
[‡]CLARITY: Centre for Sensor Web Technologies, Dublin City University, Ireland

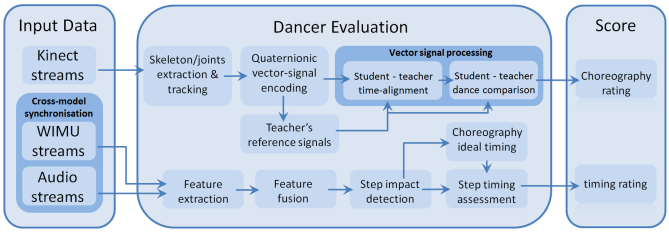

Fig. 1. System overview.

of the dance teacher, which is assumed to be the *gold standard* performance.

3. CHOREOGRAPHY ANALYSIS

The "Choreography" rating is acquired using tracked dancer 3D skeletons. The skeletal movements of the student and teacher are temporally aligned and their body joints are compared to reveal an underlying score on how well their performances match. Body joints from 3D skeletons are detected and tracked in this work by utilising the OpenNI SDK (`http://www.openni.org`) and Kinect depth-maps. The API provides the ability to track 17 skeletal joints 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 stand 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 dataset. This tailored calibration data resulted in robust skeletal tracking and consequently accuracy of the automatic evaluation methods.

3.1. Vector signals encoding

Let the position of the j-th joint at the time instance t (as produced by the skeleton tracking module) be denoted as: $[X_j(t), Y_j(t), Z_j(t)]^{\mathrm{T}}$, $j = 1, 2, \ldots, J$. The instantaneous velocities of the joints, denoted as $[Vx_j(t), Vy_j(t), Vz_j(t)]^{\mathrm{T}}$, are calculated from the convolution of the discrete-time input position data with a 1st order Derivative of Gaussian (DOG) kernel, in order to account for the noise on the input data.

By applying appropriate vector-signal and 3D-data analysis techniques it is possible to (a) register/align two dancers with respect to time and (b) to infer quantitative information about the "similarity" of their dances. In order to simplify the simultaneous processing of the three coordinate variables X, Y and Z, we make use of quaternions [9]. Quaternion theory constitutes a generalization of complex numbers theory, where we have three imaginary units **i**, **j** and **k** and instead of a scalar imaginary part, a 3-D "vector" imaginary part is considered. For the necessary theory, the reader is referred to [9]. We use "pure" quaternions (i.e. quaternions with zero real part) to encode the position and velocity of each joint during time, as quaternionic signals: $\mathbf{p}_j(t) = \mathbf{i} \cdot X_j(t) + \mathbf{j} \cdot Y_j(t) + \mathbf{k} \cdot Z_j(t)$, and $\mathbf{v}_j(t) = \mathbf{i} \cdot Vx_j(t) + \mathbf{j} \cdot Vy_j(t) + \mathbf{k} \cdot Vz_j(t)$.

3.2. Dancer alignment

Given two separate dancing sequences to compare, it is improbable that they both have the same number of frames. In addition, the time-instance at which the tracking module detects the dancer and starts tracking is unlikely the same for all dancing sequences. Therefore, an appropriate preprocessing step is initially taken, which (a) removes all frames before the dancer is detected, and (b) appropriately pads the shortest of the two sequences.

Alignment of dancers is achieved by finding the time-lag that maximizes the quaternionic cross-covariance [9] of each joint's position signal vector. More specifically, the lag $\tau_{\max}$ which maximizes the following function: $C_{\text{total}}(\tau) = \left| \sum_j \sum_t \mathbf{P}^{\text{ref}}(j,t) \cdot \overline{\mathbf{P}^{\text{eval}}}(j, t-\tau) \right|$, constitutes the estimate of the time-shift between the dancing sequences to compare. In this equation, $\mathbf{P}(j,t) = [\mathbf{p}_1(t), \mathbf{p}_2(t), \ldots, \mathbf{p}_J(t)]^{\mathrm{T}}$ is a quaternionic array with the j-th row containing the 3D position of the j-th joint during time. The superscripts "ref" and "eval" refer to the reference and evaluated dancing sequence, respectively. The overline denotes quaternion conjugate. An example of aligned signals is illustrated in the upper diagram of figure 2.

3.3. Choreography Score calculation

Initially, a score for each joint is calculated by considering the modulus of the Quaternionic Correlation Coefficient (QCC) for each pair of joint position signals: $S_{1,j} = \frac{|\sum_t \mathbf{p}_j^{\text{ref}}(t) \cdot \overline{\mathbf{p}_j^{\text{eval}}(t)}|}{\sqrt{\sum_t |\mathbf{p}_j^{\text{ref}}(t)|^2} \cdot \sqrt{\sum_t |\mathbf{p}_j^{\text{eval}}(t)|^2}}$. A total score S_1 is computed as a weighted mean of the separate joint scores, i.e.: $S_1 = \frac{\sum_j w_j \cdot S_{1,j}}{\sum_j w_j}$. The weights could be either heuristically selected, based on the significance of each joint in the dancing performance, or automatically calculated using a global optimization, realizing training with a subset of the ground-truth ratings. For the presented experiments, the weights were selected equal to unity.

Using a similar methodology, an overall score S_2 is extracted based on the velocities of the joints $\mathbf{v}_j(t)$, instead of their positions.

Finally, considering a different approach, the velocities of the joints are considered as 3D motion (flow) vectors. Inspired by the relevant 2D optical flow literature [10], a third score is produced by considering the normalized 3D velocity vectors in homogeneous coordinates (we drop t for simplicity): $\mathbf{s}(\mathbf{v}_j) = \frac{[Vx_j, Vy_j, Vz_j, 1]^{\mathrm{T}}}{\sqrt{|\mathbf{v}_j|^2 + 1}}$. The flow angular error between $\mathbf{s}(\mathbf{v}_j^{\text{ref}})$ and $\mathbf{s}(\mathbf{v}_j^{\text{eval}})$ is defined as the cosine arc of their inner product. A normalized score metric, $S_3 \in [0, 1]$, is obtained using: $S_{3,j} = \frac{1}{2}\left(\mathbf{s}\left(\mathbf{v}_j^{\text{ref}}\right) \cdot \mathbf{s}\left(\mathbf{v}_j^{\text{eval}}\right)^{\mathrm{T}} + 1\right)$.

For each frame t, the median along j is used to calculate an all-joints score $S_3(t)$. The median is utilised in this fashion in order to reject statistical outliers. A total score S_3 for the whole choreography is then given as the median along t.

The three scores S_1, S_2 and S_3 are combined with the use of appropriate weights (set equal to unity for the results in this paper), in order to obtain the final overall choreography score presented.

Instantaneous scores: The approach described so far produces a score for the whole dancing sequence. Instantaneous scores can be calculated by applying the relevant methodologies within a time-sliding window, around each time instant. Additionally, the methodologies can be applied considering different subsets of body joints, in order to extract scores for different body parts and highlight the defects/differences of the student's with the teacher's performance.

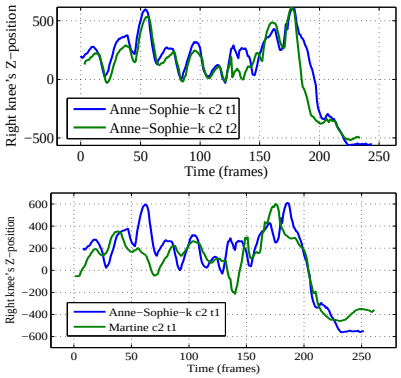

Fig. 2. Upper row: The Z-position of Anne-Sophie-k's right knee in choreographies c2-t1 and c2-t2. The signals are almost identical. Bottom row: The corresponding signals, comparing an amateur dancer with Anne-Sophie-k. The signals differ significantly, resulting in to a low score.

4. DANCER'S "TIMING" ANALYSIS

The quality of a dancer's movements "timing" is assessed by controlling whether the dance steps are executed at the correct time instants as given by the time-coded choreography-ground-truth [7]. We first explain how we detect the steps and classify them into right or left-foot steps. Then we describe the procedure followed to get a "musical timing" rating.

Step detection exploits both the audio and WIMU modalities that are first processed independently one from another before a heuristic fusion approach is used to reach a decision regarding step segmentation and classification.

4.1. Audio processing

Step impacts are detected using the signals captured by the onfloor piezoelectric transducers (audio channels 1, 2, 17 and 18). The impacts are located by:
i) first, extracting onset detection functions $f_c(t)$ from every signal, where c is a channel number ($c \in \{1, 2, 17, 18\}$ or $c \in \{8, 19, 20, 21, 22, 23, 24\}$), and t is a time index;
ii) then forming feature vectors $\mathbf{x}_t$ which are, at every time instant t, merely the concatenation of the coefficients of every onset detection function, that is $\mathbf{x}_t = [f_1(t), f_2(t), f_{17}(t), f_{18}(t)]$;
iii) applying a one-class Support Vector Machine (SVM) [11] to those feature vectors (as a way of achieving a fusion of the onset detection functions) and taking the SVM output values which are below a negative threshold to be indicators of step impact instants.

The use of single-class SVMs is motivated by the fact that they are able to detect extreme points [11] of the feature vector sequence that are to be mapped with step impacts. These extreme points are found by looking at the sign of the SVM prediction values that are then negative.

Figure 3 shows the result of the automatic step impact detection on the recording referenced as `bertrand_c1_t1`, along with ground-truth annotations. It can be seen that the approach successfully detects all steps on this example. The errors are often false alarms due to situations where "double-impact" steps occur, that is when a dancer's foot slowly enters in contact with the floor, typically with a heel impact occurring before a toe impact.

4.2. WIMU processing

WIMU based step detection is performed using the accelerometer signals of the ankle sensors (devices 5 and 6). As with the clap event, a large energy burst will occur on all three axes when the foot strikes the floor. However, the force of the impact can vary greatly, and the natural motion of the dancer's feet introduces noisy accelerations in the step detection process. In order to filter the signal, all three axes are combined and high passed to remove any acceleration values that are not the result of vibration. The signal is then normalized and integrated as described in [12] giving the probabilities of step impacts.

4.3. Decision fusion and step classification

Figure 3 illustrates the way in which audio-based and WIMU-based detectors play complementary roles in spotting the dancers steps. The audio output allows us to locate the steps with a high temporal accuracy (thanks to a higher sampling frequency), while the WIMU output serves as a validation for the step occurrences thus detected, with the potential of eliminating possible false detections, as has been observed in other examples. This is achieved by dropping all step candidates (detected using the audio) for which the WIMU detection functions are always below a threshold (fixed to 0.1) in a 0.1-s width window around those candidates.

Further, by pairing the peaks of the WIMU and audio detection functions (that are closest in time), we straightforwardly obtain a classification of steps detected into left or right-foot steps, given that the mapping between the devices and the feet is known.

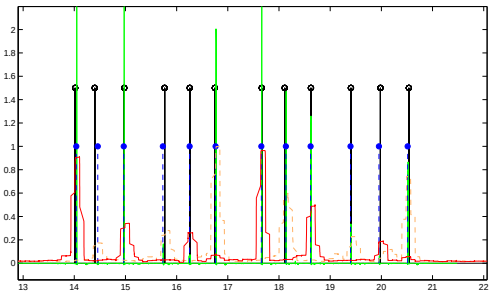

Fig. 3. Step detection results on recording `Bertrand_c1_t1`. Negative SVM decision values across time (in green) along with detected steps (dashed blue), WIMU-based detection functions (in red and orange respectively for the left and right feet), and ground-truth annotations in black.

5. EXPERIMENTAL VALIDATION

5.1. Dancer's choreography score

Scores S_1 and S_2 are correlation-based. Therefore, they constitute a "similarity" measure of the "dance signals" being compared. For example, the signals in the upper diagram of figure 2 are almost identical, revealing that the two dances of Anne-Sophie-k are very similar. Therefore, the corresponding scores are very high. On the other hand, the signals in the bottom row of figure 2 have significant differences (resulting into a low score), since the amateur dancer does not reproduce accurately the choreography. According to our experiments, score S_3 also presents a similar behavior. Assuming that the

ground-truth ratings for the reference dance are "excellent", it is essential to consider that the automatically extracted scores reflect the "Choreography" performance, namely the body movement quality / accuracy in executing a specific sequence of dance steps. In order to demonstrate the above fact, we consider a simple simulation scenario which assumes that the dancer stops dancing at a specific time instance, he/she remains almost still for a small "dead" time interval and finally continues dancing. The greater the dead time interval is, the lower the "choreography" rating will be.

5.2. Dancer's timing score

This score characterises the overall timing precision of the dancer with reference to the ideal step timing, relative to the musical timing, that is provided in the time-coded ground-truth annotation of the choreographies. Such a characterisation cannot be done without accounting for the fact that across the performance, a dancer who would be out of musical synchronisation at a certain point might still be able to re-synchronise their selves, hence end-up receiving a high overall timing rating. Therefore the scores are computed as follows.

First, we match every *expected step* s_r (given in the ground-truth) with the closest automatically *detected step* s_d. If it is observed (after step classification) that the latter has been executed by the correct foot (again with reference to the ground-truth), and its time distance to the reference step s_r is smaller than the musical beat duration (also given in the music annotation files), then it is registered in the set of correct steps $\mathcal{S}_c$; otherwise it is ignored. Next, two intermediate scores are computed: the first is the arithmetic mean of the time-differences between set $\mathcal{S}_c$ steps and their nearest ground-truth counterparts; the second is the fraction of correctly executed steps relative to the total number of ground-truth steps expected. These two intermediate scores are normalised to lie in the range $[0, 5]$, and the final "timing" score is obtained by computing their mean. This strategy has been found successful as will be seen in the next Section.

5.3. Experimental results

The presented methodologies produce meaningful results, in the following sense: 1) Considering Bertand or Anne-Sophie-k (the professional dancers) in two different captures for the same dance, the calculated scores are high. This is essential, since the dance of a professional dancer is almost identical in two different captures and 2) The ranking of the students does not deviate significantly from the ground-truth ranking.

Selected experimental results (for choreographies c2 and c3) are presented in Table 1, where the overall scores are given in the columns "Ch-Score" and "MT-GT" (normalized in the range [0,5]). The dance recordings identified as `Bertand_c2_t2`, and `Bertand_c3_t1` were considered as the reference recordings for computing the choreography rating. In the same table the ground-truth ratings assigned to the dancers are presented (columns "Ch-GT" and "MT-GT"). It should be noticed that the ranking of the students based on the calculated scores are in accordance with the ranking based on the ground-truth.

Table 1. Ground-truth and extracted scores, where Ch-GT and MT-GT stand for "Choreography" and "Musical timing" Ground-truths respectively.

Dancing	Ch-GT	Ch-Score	MT-GT	MT-Score
jacky c2 t1	5	4.3	3	3.4
thomas c2 t2	5	4.1	4	4.3
habib c2 t1	4	3.9	5	4.7
jacky c3 t2	4	3.8	3	3.1
thomas c3 t1	2	3.2	2	3.6
habib c3 t1	2	1.8	4	3.9

6. CONCLUSIONS AND FUTURE WORK

In this paper we presented a framework which allows for automatic dance analysis and evaluation, and can be used by both a teacher to evaluate a student's performance and a student as a source of meaningful feedback. In future work, we aim at providing an online 3D visualization tool allowing for the temporal aligned dance movements of the teacher and the students, along with the associated evaluation scores, in a virtual 3D gaming environment.

7. REFERENCES

[1] JK Aggarwal and M.S. Ryoo, "Human activity analysis: A review," *ACM Comp. Surveys (CSUR)*, vol. 43, pp. 16, 2011.

[2] Antonio Camurri, Barbara Mazzarino, Matteo Ricchetti, Renee Timmers, and G. Volpe, "Multimodal analysis of expressive gesture in music and dance performances," *Gesture-based comm. in human-comp. interaction*, pp. 357–358, 2004.

[3] S. Takaaki, N. Atsushi, and I. Katsushi, "Detecting dance motion structure through music analysis," in *IEEE Inter. Conf. on Automatic Face and Gesture Recognition*, 2004, p. 857862.

[4] Rajkumar Kannan, Frederic Andres, and B. Ramadoss, "Tutoring System for Dance Learning," in *IEEE International Advance Computing Conference (IACC 09)*, Patiala, India, 2009.

[5] T. Kim, S. I. Park, and S. Y. Shin, "Rhythmic-motion synthesis based on motion-beat analysis," in *Proc. of ACM SIG-GRAPH*, 2003.

[6] L.A. Naveda and M. Leman, "Representation of samba dance gestures, using a multi-modal analysis approach," in *ENACTIVE08 5th Inter. Conf. on Enactive Interfaces*, 2008, p. 68.

[7] 3DLife/Huawei ACM MM Grand Challenge 2011: Realistic Interaction in Online Virtual Environments, "http://perso.telecom-paristech.fr/ẽssid/3dlife-gc-11/," .

[8] M. Gowing, P. Kelly, N. E. O'Connor, E. Izquierdo, V. Kitanovski, X. Lin, Q. Zhang, C. Concolato, S. Essid, J. Le Feuvre, and R. Tournemenne, "Enhanced visualisation of dance performance from automatically synchronised multimodal recordings," in *ACM MM*, USA, 2011, Accepted.

[9] C. Eddie Moxey, Stephen J. Sangwine, and Todd A. Ell, "Hypercomplex correlation techniques for vector images," *IEEE Trans. on Signal Processing*, vol. 51, pp. 1941–1953, 2003.

[10] John L. Barron, David J. Fleet, and Steven Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12, pp. 43–77, 1994.

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

[12] H. Ying, C. Silex, A. Schnitzer, S. Leonhardt, and M. Schiek, "Automatic step detection in the accelerometer signal," in *Proc. of International Workshop Body Sensor Networks*, 2007.