Evaluating a Dancer's Performance using Kinect-based Skeleton Tracking

Dimitrios Alexiadis,
Petros Daras
Informatics and Telematics
Institute, Thessaloniki, Greece
{dalexiad, daras}@iti.gr

Philip Kelly,
Noel E. O'Connor
CLARITY: Centre for Sensor
Web Technologies, Dublin City
University, Ireland
{philip.kelly,
noel.oconnor}@dcu.ie

Tamy Boubekeur
Institut Telecom / Telecom
ParisTech, Paris, France
tamy.boubekeur@telecom-paristech.fr

Maher Ben Moussa
MIRALab, University of
Geneva, Switzerland
benmoussa@miralab.ch

ABSTRACT

In this work, we describe a novel system that automatically evaluates dance performances against a gold-standard performance and provides visual feedback to the performer in a 3D virtual environment. The system acquires the motion of a performer via Kinect-based human skeleton tracking, making the approach viable for a large range of users, including home enthusiasts. Unlike traditional gaming scenarios, when the motion of a user must by kept in synch with a prerecorded avatar that is displayed on screen, the technique described in this paper targets online interactive scenarios where dance choreographies can be set, altered, practiced and refined by users. In this work, we have addressed some areas of this application scenario. In particular, a set of appropriate signal processing and soft computing methodologies is proposed for temporally aligning dance movements from two different users and quantitatively evaluating one performance against another.

General Terms

Algorithms, Experimentation

Keywords

Skeleton tracking, Microsoft Kinect, Signal Processing

1. INTRODUCTION

The future of social networking is gearing towards immersive, content-centric, collaborative environments that support real-time, realistic interaction between humans. The Huawei 3DLife/EMC^2 challenge [2] run as part of the ACM Multimedia Grand Challenge Series 2011 announced a call for demonstrations of relevant technologies that can support real-time online human interaction. The application scenario considers an online dance class provided by an expert Salsa dancer. In such a scenario, a dance teacher (for example) is free to illustrate to online users choreography steps of their choice. After viewing the sequence at a later date, another online user (a student, for example) 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 the student has reached a certain level of competency. As such, there is real online interaction between users.

In this paper, we study some of the technical issues that would need to be addressed in this challenging scenario. In particular, we target the problem of real-time automatic alignment, evaluation and feedback of dance performances. More specifically, we present a system that can automatically align a student dance performance to that of a teacher, calculate an overall and an instantaneous score for his/her performance, and provide visual feedback on the performance compared to the teacher. In order to align and evaluate dance performances, Kinect depth-maps from the associated Grand Challenge dataset are considered. A set of signal processing methodologies, combined with human skeleton tracking from Kinect depth-maps, is proposed for evalaution. In addition, we provide visualization of the temporally aligned dance movement of both teacher and students, along with the associated evaluation scores, in a virtual 3D gaming environment. This visualization tool also provides functionality to slow down or change the orientation of the visualization, allowing users analyze their dance moves from multiple perspectives.

The paper is organized as follows: In section 2 we briefly describe the ACM Grand Challenge dataset. In section 3 we describe hoe the Kinect skeleton tracking module is used for the real-time tracking of the dancers. In section 4 we provide details on the proposed signal processing methodologies for dancer evaluation. Section 5 outlines the visualization tool and its operation. Finally, in section 6 we present some experimental results acquired using the developed software.

2. DATASET

In this work, we have utilized the dataset from the *Realistic Interaction in Online Virtual Environments* Huawei 3DLife/EMC^2 ACM Grand Challenge that includes recordings of Salsa dancers captured with a variety of modalites and equipment, including (among others) Microsoft Kinect sensors. The dataset includes recordings of two professional dancers (a male and a female), corresponding to teachers, and 13 amateur dancers (8 males and 5 females), corresponding to students, in six different Salsa choreographies. Although the dataset contains much more information, in this work we use solely the Kinect modality, as this may be all that is available for many home enthusiasts.

3. KINECT SKELETON TRACKING

The Kinect dataset recordings of dancer performances were captured using the OpenNI [3] drivers/SDK and are OpenNI-encoded (.ONI). The OpenNI SDK provides, among others, a high-level skeleton tracking module, which can be used for detecting the captured user and tracking his/her body joints. More specifically, the OpenNI tracking module produces the positions of 17 joints (Head, Neck, Torso, Left and Right Collar, L/R Shoulder, L/R Elbow, L/R Wrist, L/R Hip, L/R Knee and L/R Foot), along with the corresponding tracking confidence – see panel 7 in Figure 2.

The OpenNI tracking module requires a-priori user calibration in order to infer information about the user's height and body characteristics. More specifically, skeleton calibration requires the captured user to stay still in a specific "calibration pose" for a few seconds. However, this pose was not captured for the dancers and therefore custom skeleton tracking calibration for each dancer is not possible. Fortunately, our experiments show that skeleton tracking is quite effective even if non-exact calibration data is provided to the OpenNI tracking module. In this work, we created custom calibration results for each dancer by manually sourcing and calibrating persons with similar body characteristics to each dancer in the dataset. This tailored calibration data resulted in more robust skeletal tracking and consequently higher accuracy of the automatic evaluation methods.

4. AUTOMATIC DANCER EVALUATION

Based on the calibration procedure described in Section 3, we developed C++ OpenNI-based skeleton tracking software and a MATLAB wrapper to acquire the skeleton tracking output from the ONI recordings. The skeleton tracking module outputs the positions of the dancer's joints for each frame. These positions actually constitute a 3D vector signal. In addition, the dynamics of dancing movements can also be acquired from the outputs of the skeleton tracking module. In this work, these dynamic movements are acquired as the instantaneous 3D velocities of the joints and are calculated from the convolution of the (generally noisy) discrete-time position signals with a 1st order Derivative of Gaussian (DOG). In order to provide a score for each choreography we propose to compare the aligned position and velocity vector signals of an amateur dancer with the corresponding signals of a professional one. Note, that the Grand Challenge dataset signifies which dancers are amateurs and professionals.

4.1 Methodologies

In order to handle the three coordinate variables X, Y and Z in a holistic manner, the adopted signal processing techniques make use of hypercomplex numbers and specifically quaternions [5]. Quaternions, which are popular in the 3D graphics community, have recently been used for signal and image processing [5] and constitute a generalization of complex numbers, where instead of a scalar imaginary part, a 3D "vector" imaginary part is considered. As providing details on quaternion theory is beyond the scope of this paper, interested readers are referred to [5].

4.1.1 Dancer Alignment

Within the Grand Challenge dataset, all dancers align their dance routines with respect to the background music, however this audio is not available with the Kinect data, and the Kinect data stream is not synchronized with respect to the start/end of the background music. As such, the Kinect dance sequences are in effect non-synchronized with respect to the frame numbers. In addition, typically two distinct dancing sequences that we wish to compare are not of the same temporal length, i.e. number of frames. Additionally, the time-instance at which the skeletal tracking module detects the dancer and starts tracking is not common across all sequences. As such, in order to obtain more useful signals for comparison, preprocessing is required to align dance sequences. In this work, alignment is achieved by exploiting solely the Kinect data.

The adopted preprocessing scheme is as follows: 1) For each sequence, discard all the frames before the detection of the dancer, i.e. before the time-instance when at least one joint is detected; 2) Use a flag value, NaN (Not A Number), to represent an undetected/tracked joint; and 3) Pad the shortest sequences with NaN flag values, so that both sequences to be compared have the same temporal length. Note that NaN values are not considered in further calculations. In this work, dance alignment is achieved by exploiting only the Kinect data and using the quaternionic cross-covariance [5]. The time-shift between two dancing sequences is estimated by calculating the modulus of the quaternionic cross-covariance for all joints. An example is depicted in figure 1.

4.1.2 Dancer Evaluation

In this work three different scores are calculated for a dancer's performance, which are subsequently combined to produce an overall score. The proposed scores are calculated as follows:

Score #1 - Joint Positions
A score for each joint is calculated by considering the modulus of the quaternionic Correlation Coefficient (CC) for each pair of joint position signals. A total score S_1 is then computed as the weighted mean of the separate joint scores.

Score #2 - Joint Velocities
Similarly, an overall score S_2 is extracted based on the velocities of the joints, instead of their positions, by considering the quaternionic CC for the joint velocity signals.

Score #3 - 3D Flow Error
For a given frame, the velocities of the joints are considered as 3D motion (flow) vectors. Inspired by the relevant 2D optical flow literature [4], we consider the normalized 3D velocity vectors in homogenous coordinates and calculate the vectors' inner product to obtain a score for each joint. An "all-joints" score is produced from the median of the separate inner products in order to reject outliers, i.e. very wrong estimates due to inaccurate skeleton tracking. A total score S_3 for the whole choreography is calculated from the median across time.

Combined score
Having computed three different scores S_1, S_2 and S_3, a

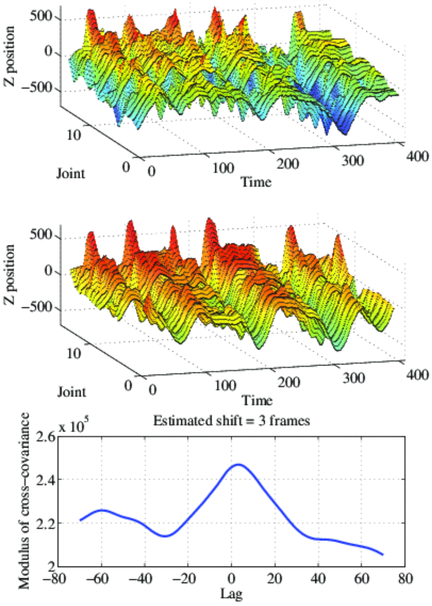


Figure 1: Two upper rows: The Z positions of all joints for the professional dancer (Betrand c3 t1) and an amateur dancer (Gabi c3 t1). Bottom row: The corresponding modulus of quaternionic cross-covariance. The estimated time-shift is 3 frames.

combined score is calculated as the weighted mean of the three. The set of the three weights can be optimally selected using an optimization approach.

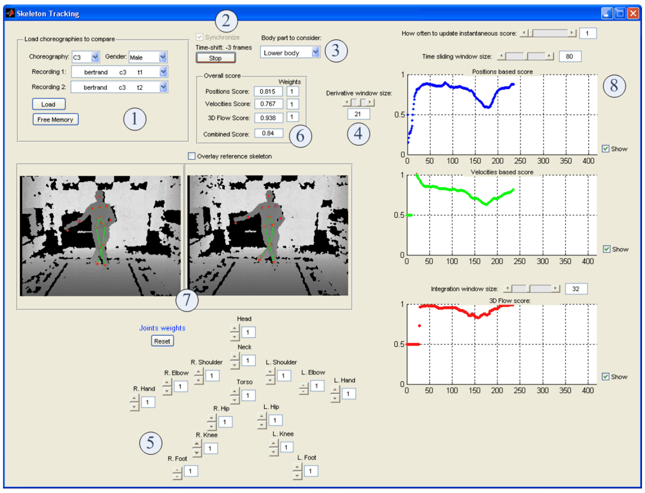
Relations with the ground-truth ratings

The calculated scores are produced by comparing the amateur's dance with the professional's one. Consequently, assuming that the ground-truth ratings for the reference dance is "excellent", the automatically extracted scores reflect more-or-less all five ground-truth criteria ("Upper" and "Lower Body Fluidity", "Musical Timing", "Body Balance" and "Choreography"). However, it is essential to consider that the calculated scores are mainly related to the "Choreography" (accuracy in executing a specific sequence of dance steps) and "musical timing" performance. This can be supported based on the fact that the scores S_1 and S_2 are correlation-based. Therefore, they constitute a measure of "similarity" of the "dancing signals" being compared and the degree to which they are time-aligned. Score S_3 presents also a similar behavior. The above arguments can be demonstrated by simple simulation scenarios experiments, the presentation of which is however beyond the scope of this paper.

Instantaneous scores and separate scores for different body parts

If scoring is required for different sections the dance, rather than a single score for the entire recital, a straightforward methodology to produce instantaneous scoring can be adopted. In this case, instead of considering the vector's signals for the whole time interval, a time-sliding window is employed, whereby the instantaneous score for a time instance, t, is calculated by applying the described methodologies for the time-interval of length, L, around t.

Additionally, the described ideas can be extended in order to provide feedback to the amateur on how he/she can


Skeleton Tracking

Load choreographies to compare
Choreography: C3
Gender: Male
Recording 1: bertrand c3 t1
Recording 2: bertrand c3 t2

☑ Synchronize
Time-shift: -3 frames

Body part to consider: Lower body

Overall score
		Weights
Positions Score:	0.815	1
Velocities Score:	0.767	1
3D Flow Score:	0.938	1
Combined Score:	0.84	

Derivative window size: 21

□ Overlay reference skeleton

How often to update instantaneous score: 1
Time sliding window size: 80
Integration window size: 32

Joints weights
Head: 1
Neck: 1
R. Shoulder: 1
L. Shoulder: 1
R. Elbow: 1
L. Elbow: 1
R. Hand: 1
L. Hand: 1
Torso: 1
R. Hip: 1
L. Hip: 1
R. Knee: 1
L. Knee: 1
R. Foot: 1
L. Foot: 1

☑ Show
☑ Show
☑ Show

Figure 2: Data Acquisition Software GUI.

improve his/her performance. In such a scenario, separate scores can extracted for different body parts, for example the upper and the lower body parts, by separately considering the upper and lower body joints.

5. EVALUATION FRAMEWORK

The evaluation framework is split into two different parts; (A) Data acquisition and (B) Data visualization.

Data Acquisition Software.

A Graphical User Interface (GUI) was developed in Matlab, in order to integrate the developed methods, allow parameter selection and provide initial visualization of the results. The GUI is depicted in figure 2, where an evaluation example is presented. Among the possibilities provided by the GUI are (the enumeration is given with respect to figure 2): 1) Selection of two choreography recordings to compare, 2) Dancer alignment, 3) Selection of the body part to consider, 4) Methodologies' parameter selection, 5) Weights selection, 6) Calculation of three overall scores, 7) Visualization of the skeleton tracking results, 8) Visualization of the three scores over time. Using this GUI all parameters associated with the proposed system can be controlled.

Data Visualization Software.

Although the data acquisition software provides some basic 2D visualization of the dancers, we believe that users will benefit more from a virtual 3D gaming environment. In such an environment, users are free to view their dance (plus the aligned dance of the teacher) from any orientation they wish, allowing them to subjectively critique their movements from a variety of angles. In addition, a user can pause at a particular frame, step forward/backward, or continuously play back at a decreased framerate to facilitate users to accurately analyze their dance moves. For this visualization, we employed the Unity gaming development tool [1], mapping the human motions onto photo-realistic avatars and placing the avatars into a virtual dance studio. In addition, we provide feedback on their movements in terms of scoring their motion against a teacher's motion. Example images from the virtual environment can be seen in Figure 3.

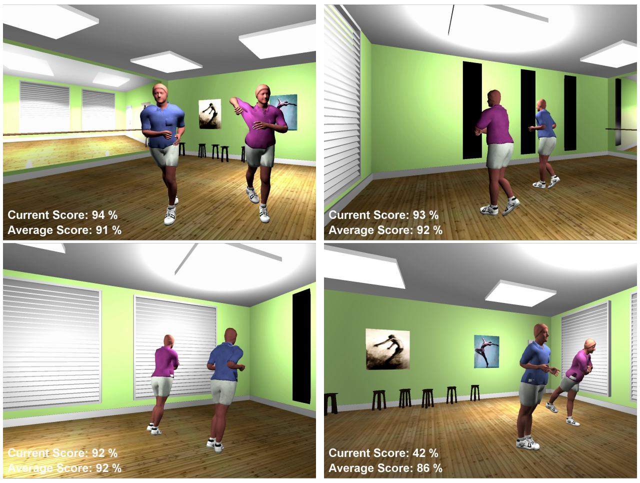


Figure 3: Visualization results from multiple angles and time sequences, notice how images with similar poses have good scores, but the dances diverges in the bottom right and is assigned a low score.

6. EXPERIMENTAL RESULTS

Although it is of course extremely difficult to automatically evaluate something as subjective as dancer performance, according to our experiments, the adopted methodologies do produce meaningful results, in the following sense; (1) Considering the professional dancer (Bertand or Anne-Sophie-k) in two different captures for the same choreography, the calculated scores are high. For example, in figure 2, the professional male dancer Bertand is compared with himself with respect to choreography c3, considering two different captures, t1 and t2. Obtaining a high score in this case is essential for our system, since a professional dancer is able to perform almost identical dancing movements in two different captures; (2) All three scores produce more-or-less the same ranking of the amateur dancers; (3) The ranking does not deviate significantly from the ranking produced using the ground-truth ratings provided in the dataset; (4) The three instantaneous scores have similar behavior (e.g. they have minima or maxima at almost the same time instances, see figure 2); and (5) For not perfectly aligned dancers, the instantaneous scores are lower in time-intervals where synchronization of dance steps is lost.

Some experimental results are presented in table 1, where the overall score for the whole body calculated automatically is given in the last column ("Score") for choreographies c2 and c3. In these experiments, either Bertand or Anne-Sophie-k were used as the professional dancer (teacher), depending on the gender of the amateur dancer. In the same table, the ground-truth ratings assigned to the dancers by the teachers (i.e. a score between 1-5, the higher the better the performance) for "Choreography" (CH), "Musical Timing" (MT) and "Body Balance" (BB) are displayed. The rows of the table are sorted by CH, then by MT and finally by BB. It should be noted that in general there is strong correlation between the ground truth and the calculated scores, in the sense that the ground-truth-based ranking of the students is similar to the one obtained according to the calculated scores. Notice that the low scores of 0.33 and 0.32 for automatic evaluation are due to bad skeleton calibration and loss of the dancer during tracking.

Table 1: Teacher Ratings and Methodology Scores.

Choreography	BB	MT	CH	Score
thomas c2 t2	4	4	5	0.72
anne-sophie-k c2 t2	5	3	5	0.8
jacky c2 t1	5	3	5	0.71
habib c2 t1	5	5	4	0.70
habib c3 t2	5	3	4	0.33
jacky c3 t2	4	3	4	0.65
ming-li c2 t1	4	1	4	0.61
habib c3 t1	5	4	2	0.32
ming-li c3 t2	4	2	2	0.67
thomas c3 t1	3	2	2	0.59

7. CONCLUSIONS

Skeleton tracking from Kinect depth-maps can provide input to a set of appropriate signal processing methods that can support real-time evaluation of dancers in online interactive environments. In this paper, some relevant novel ideas were described and the implementation details of a working system were given. The experimental results were promising, demonstrating the appropriateness of the proposed approach for real-time interaction in an online dance class.

8. ACKNOWLEDGMENTS

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

9. REFERENCES

[1] http://unity3d.com/.

[2] http://www.3dlife-noe.eu/3dlife/2011/04/3dlife-acm-grandchallenge-2011/.

[3] http://www.openni.org/.

[4] J. L. Barron, D. J. Fleet, and S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12:43–77, 1994.

[5] C. E. Moxey, S. J. Sangwine, and T. A. Ell. Hypercomplex correlation techniques for vector images. *IEEE Trans. on Signal Processing*, 51:1941–1953, 2003.