Enhanced Visualisation of Dance Performance from Automatically Synchronised Multimodal Recordings

Marc Gowing[1]
Marc.gowing@gmail.com
Philip Kell[1]
kellyp@eeng.dcu.ie
Noel E. O'Connor[1]
Noel.OConnor@dcu.ie

Cyril Concolato[2]
cyril.concolato@telecom-paristech.fr
Slim Essid[2]
slim.essid@telecom-paristech.fr
Jean Lefeuvre[2]
jean.lefeuvre@telecom-paristech.fr
Robin Tournemenne[21]
robin.tournemenne@wanadoo.fr

Ebroul Izquierdo[3]
ebroul.izquierdo@elec.qmul.ac.uk
Vlado Kitanovski[3]
vlado.kitanovski@elec.qmul.ac.uk
Xinyu Lin[3]
xinyu.lin@eecs.qmul.ac.uk
Qianni Zhang[3]
qianni.zhang@elec.qmul.ac.uk

[1] CLARITY: Centre for Sensor Web Technologies
Dublin City University, Ireland

[2] Institut Telecom/Telecom ParisTech
CNRS/LTCI
75013 Paris

[3] Multimedia and Vision Group
School of Electronic Engineering and Computer Science
Queen Mary, University of London
London, E1 4NS

ABSTRACT

The Huawei/3DLife Grand Challenge Dataset provides multimodal recordings of Salsa dancing, consisting of audiovisual streams along with depth maps and inertial measurements. In this paper, we propose a system for augmented reality-based evaluations of Salsa dancer performances. An essential step for such a system is the automatic temporal synchronisation of the multiple modalities captured from different sensors, for which we propose efficient solutions. Furthermore, we contribute modules for the automatic analysis of dance performances and present an original software application, specifically designed for the evaluation scenario considered, which enables an enhanced dance visualisation experience, through the augmentation of the original media with the results of our automatic analyses.

Categories and Subject Descriptors

I.2.10 [**Vision and Scene Understanding**]: Video analysis.

General Terms

Algorithms

Keywords

Audio, video, synchronisation, multimodal processing.

1. INTRODUCTION

In this work, we address the Huawei/3DLife Grand challenge [1], which consists of multimodal recordings of Salsa dancers, including multi view videos from UniBrain cameras, Microsoft Kinect streams, data from Wireless Inertial Measurement Units (WIMU), and synchronised 16-channel audio capture of dancers' step sounds, voice and music. We focus on developing a system for the subjective evaluation of Salsa dancers. As dancers are captured with various recording equipment, it is essential to temporally synchronise all of the multimodal data when displaying/playing it to the evaluators. It is also essential to synchronise data between different dancers to allow an evaluator to compare the performances of two dancers, typically the tutor's performance, which is considered as a reference, and the student's performance. In this work, our primary concern is to propose efficient solutions to tackle these synchronisation problems. Furthermore, we contribute an original software application, specifically designed for this evaluation scenario, which enables an enhanced dance visualisation experience through the augmentation of the original media with the results of automatic analyses of the dance performances. In particular, we present a novel technique to achieve dancer step segmentation jointly using audio signals captured by the onfloor piezoelectric sensors and signals from the WIMU devices.

The outline of the paper is the following: The next section describes the methods used to address the synchronisation issues, before proceeding to the description of the automatic analysis components in Section 3. We present the software application design in Section 4, followed by conclusions in Section 5.

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

2. MULTIMODAL SYNCHRONISATION

Figure 1 gives an overview of our approach to synchronisation between the heterogeneous streams of data recorded.

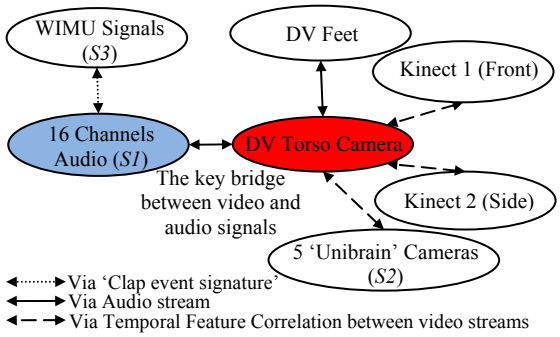

Figure 1 Overview of synchronisation strategy

The details of the different synchronisation components are given hereafter. To understand the rationale behind this synchronisation scheme it is important to keep in mind that some subsets of data streams are already synchronised via hardware [1], these include:

- Subset *S1* consisting of data from 16 audio channels
- Subset *S2* composed of video streams of 5 UniBrain cameras;
- Subset *S3* that is the different WIMU signals.

Therefore, it is sufficient to synchronise instances of each subset with the other data streams to achieve overall synchronisation.

2.1 Audio-based synchronisation

As illustrated in Figure 1, the audio modality is used as a bridge to synchronise other types of modalities. The procedure is as follows:

- Synchronise the videos taken by the feet and torso cameras using audio-to-audio synchronisation between the audio streams of these videos (as described in Section 2.1.1);
- Synchronise one of the audio channels of S1 with either audio streams of the feet or torso cameras, using the same audio-to-audio synchronisation method;
- Synchronise one of the audio channels of S1 with the WIMU signals in S3 as described in Section 2.2.

To complete overall synchronisation, one is left only with the problem of synchronising the videos of feet/torso cameras with the ones captured by the Kinects and UniBrain cameras, which is addressed in Section 2.3.

2.1.1 Audio-to-audio synchronisation method

Audio-to-audio synchronisation is achieved by first estimating the signals energy envelopes, then using a simple cross-correlation measure between these envelopes. The delay between the two signals is deduced as the time-lag that needs to be applied to one data stream in order to obtain maximum cross-correlation.

The audio envelopes are estimated by computing the energy values in 15-ms local audio frames with a 1-ms hop size. The sampling frequency of the envelopes is thus 1000 Hz, hence allowing us to speed-up the process compared to a situation where cross-correlation measures would be taken directly from the original audio signals whose sampling frequencies can be as high as 48 kHz. The other advantage of this approach is that it can cope with the fact that some audio streams are sampled at differing frequencies, for example the audio stream of the foot camera is at 32 kHz while signals from *S1* are sampled at 48 kHz.

Furthermore, it has been found unnecessary to consider the whole signal durations to achieve this synchronisation, rather only the first few seconds of each recording is taken, covering the initial clap event and the start of the music (on recordings with music).

2.1.2 Bridge between different dancer recordings

In the challenge scenario, all dancers are expected to execute the same choreographies and synchronise their movements to the same background music. Therefore synchronising the performances of two dancers is quite straightforward as it solely entails synchronising the recorded music signals relating to each dancer, that is channel 5/6 recordings of a dancer A with channels 5/6 recordings of dancer B. This is done using the previously described procedure.

2.2 Synchronisation of WIMUs

Synchronisation between audio and WIMUs is achieved by maximising the cross-correlation between a specific WIMU and audio features around the clap event. These features are designed to characterise the clap event signature.

The audio feature employed here is the output of an onset detection component [4] applied to the audio signal of channel 20, *i.e.* one of the overhead Shoeps microphones that clearly captures the sound of hands and feet claps.

The WIMU synchronisation feature exploits the accelerometer signal of the both wrist sensors (devices 1 and 2). A clap will appear as a large spike in the accelerometer signal of both wrist WIMU accelerometers simultaneously. To detect this event, all three axes are combined for each sensor. The average, maximum amplitude and its corresponding timestamp are calculated using 150-ms sliding window with 10-ms hops. The window with the largest variance for both WIMUs is identified as the clap signature.

As the sampling frequency of the audio feature is 360 Hz (due to the signal analysis parameters of the onset detection module) and the WIMU feature is 100 Hz, the WIMU frequency is upsampled to that of the audio stream before computing the cross-correlation.

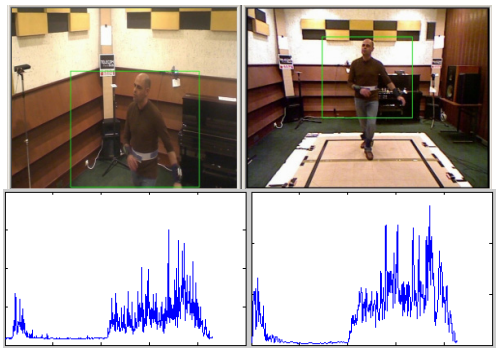

Figure 2: Top row: example for upped body detection; Bottom row: temporal features extracted from the two videos.

2.3 Multi-view video synchronisation

In this section, we describe the approach to uncover the time shift between two videos taken from unsynchronised cameras. The videos can be of different quality and frame rate. Temporal features are extracted for every video frame. Correlation-matching between features from two different videos is then used to obtain the offset between them. We also employed a method to detect the dancer's upper body to locate the area for feature extraction and improve the synchronisation accuracy.

2.3.1 Synchronisation based on co-occurrence appearances

The temporal features used for video temporal alignment are based on appearance changes [2]. This approach is suitable when cameras are static, and it does not require a high level of scene understanding. The total amount of appearance change between two successive frames is calculated by:

$$f_k[n] = \sum_{\Omega} \left(I_{n+1} - I_n \right)^2$$

where I_n is the luminance component of the n^{th} frame, and the sum of squared differences is computed over the image region Ω. In order to achieve sub-frame accuracy, the vector f_k is interpolated on a 1-ms time grid. The time shift between two videos is obtained as the value that maximises the normalised correlation between the temporal features of each video. Region Ω should be chosen so that it contains only the moving objects (or part of moving objects) visible in both videos. As such, for the 3DLife dataset we use an upper-body detection method introduced in the next section to locate the dancer's upper-body movements, as some videos in the used dataset do not capture the whole body. Figure 2 shows temporal features for the DV torso camera and the Kinect camera respectively.

2.3.2 Upper-body detection

In order to improve the synchronisation accuracy, unexpected object movement in the video sequence should be excluded from the temporal features calculation. In the original work [2], this is achieved by splitting video frames into sub-regions of regular size and iteratively excluding sub-regions that have negative impact on the correlation between two video sequences. As this trial and error approach is inefficient in terms of computational cost, we employ a state of art upper-body detection algorithm [3] to facilitate adaptive selection of regions for temporal features calculation as opposed to searching sub-regions with negative impact iteratively. In this work, the algorithm uses trained part-based models combined with an optional face detector to improve the detection performance. Figure 2 (top row) shows the detection results applied to the videos captured using the Microsoft Kinect and the DV camcorder. By applying temporal features and correlation calculations only within detected region, the synchronisation accuracy is improved.

2.3.3 Results

We evaluated the described approach using videos from the dataset [1]. There are total 103 video sets included (each video set includes: 5 UniBrain videos, torsor video and feet video recorded by DV camcorders, and two videos recorded by Kinects). Table 1 shows the accuracy of synchronisation based on one video set performed by teacher dancer. UniBrain camera 1 was used as reference video. It can be seen that the error is mostly around one or two frames (corresponding to 30-80 ms in time).

Table 1: Errors were measured as the differences in number of frames between the ground-truth time shifts and the time shifts calculated using the proposed method. The first column is the videos being synchronised; columns 2-6 are the choreographies performed by teacher dancer

Camera:	CH 1	CH 2	CH 3	CH 4	CH 5
Torso	0	0	0	-1	0
Feet	0	0	0	0	0
Kinect	-1	1	-1	1	0

3. DANCE PERFORMANCE ANALYSIS FOR AUDIOVISUAL AUGMENTATION OF THE CAPTURED MEDIA

As described in the introduction, in this work we strive to provide an augmented-reality presentation of the captured dancer performances aimed to both facilitate the subjective evaluation of a dance recital and to enhance the viewing experience for a user. To this end, two types of analysis are proposed that are briefly described hereafter, namely automatic dance step detection and 3D dancer skeleton and joint extraction. The former is an audio-rendered augmentation, taking the form of beeps sounding every time a new dance step is detected, along with textual subtitles indicating the step classification. The latter is visually rendered additional video, taking the form of a skeleton overlaid onto a depth map of the scene.

3.1 Automatic dance steps detection

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.

3.1.1 Audio processing

Audio-based step segmentation exploits the audio signals recorded by the on-floor piezoelectric sensors (channels 1, 2, 17 and 18). Step impacts on the dance floor are expected to give rise to energy bursts in the corresponding waveforms, which can be suitably characterised using onset detection [4] functions. Concatenating the coefficients of all onset detection functions for every time instant hence forms audio feature vectors, used for step detection. Next, a single-class Support Vector Machine (SVM) classifier [5] is applied to these feature vectors with the aim of achieving a fusion of the information conveyed by the different audio channels. The use of a single-class SVMs is motivated by the fact that they are able to detect extreme points 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.

This step detection strategy turns out to be very efficient as depicted in Figure 3 where it can be seen that the step impacts thus detected (represented by vertical dashed blue lines) accurately match the ground-truth (obtained by manual annotation of the impacts and represented with black vertical lines).

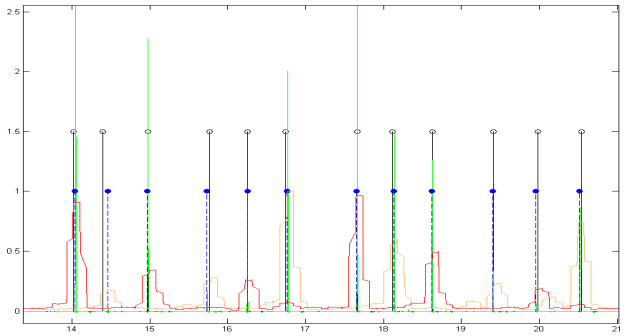

Figure 3: Step detection results on recording Bertrand_c1_t1. Negative SVM decision values across time (in green) along with detected steps (dashed blue), WIMU-based detection functions (in red and orange respectively for the left and right feet), and ground-truth annotations in black.

3.1.2 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 shown in Figure 4, giving the probabilities of step impacts, as described in [6].

3.1.3 Decision fusion and results

Figure 3 illustrates the way in which audio-based and WIMU-based detectors play complementary roles in spotting the dancer's 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) and additionally permits the classification of steps as either *right* or *left* foot steps.

3.2 Dancer's skeleton and joints extraction

The OpenNI framework [7] includes standard tools for extracting the user's joint positions from the Kinect depth map and generating a skeleton overlay. Typically, a user will perform a calibration pose as shown in Figure 5, which allows OpenNI to detect joints positions. However, as none of the dancers in the dataset performed a calibration pose, we captured a user of similar height and build performing a calibration pose and applied it to the recordings. The output is shown in Figure 5 below.

4. DANCER EVALUATION APPLICATION

In this section we present the requirements for the design of a multimedia application designed to enable users to perform a subjective evaluation of dancer performance. Given the time-varying nature of the data, the application has to have a strong focus on synchronization. It has to be able to display various types of media data (audio, video, video with depth information, 2D graphics, 3D graphics, and text) and perhaps several media elements of each type, while keeping synchronization. In terms of layout, the ability to display mixed 2D/3D data, possibly in an augmented reality mode is also deemed important. And finally, in order to propose an interactive application to the evaluators, it has to incorporate some interactive capabilities. Finally we want to design the interaction logic in a flexible manner.

Based on these requirements, we decided to implement the application using declarative languages (with some scripting extensions using JavaScript) and to use the GPAC player [8] to render the application due to its capability to mix declarative languages and media elements of different types, and its streaming and synchronization features, which allows the possibility of streaming in the future. The declarative languages used are MPEG-4 BIFS for the 3D layout and 3D graphics objects, and SVG for the 2D layout and 2D graphics. The interactivity and application logic have been implemented using JavaScript.

Figure 6 shows a snapshot of the application GUI. The interface allows one to evaluate and compare the performances of two dancers, through a choice of viewpoints (displayed in the central columns of the interface). The user can play each dancer separately, or play dancers synchronously (by clicking the chain icon between the two players' controls). Central viewpoints can be dynamically changed during runtime, which is also true for the audio stream to be played. Some viewpoints and audio streams are augmented with overlays and effects, for instance dancer skeleton and joints displays or beeps signalling dance steps (detected automatically as previously explained).

5. CONCLUSIONS

In this work, we have presented a set of methods for the synchronisation of the multimodal recordings captured for the Huawei/3DLife Grand Challenge. The accurately synchronised data was further analysed and augmented for presentation in an original software application for an enhanced dance visualisation experience. The overall system serves as a comprehensive platform for evaluating Salsa dancers' performances.

6. ACKNOWLEDGEMENTS

This research was supported by the European Commission under contract "FP7-247688 3DLife". Affiliations and authors are listed in alphabetical order.

7. REFERENCES

[1] http://perso.telecom-paristech.fr/~essid/3dlife-gc-11/

[2] Ushizaki M., Okatani T. and Deguchi K., Video Synchronization Based on Co-Occurrence of Appearance Changes in Video Sequence, Proceedings of International Conference on Pattern Recognition, 2006

[3] Eichner M. Marin-Jimenez, A. Zisserman, V.Ferrari, Articulated Human Pose Estimation and Search in (Almost) Unconstrained Still Images, ETH Zurich, Technical Report No.272, 2010.

[4] M. Alonso, G. Richard, and B. David. Extracting note onsets from audio recordings. Proc. of IEEE-ICME, 2005.

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

[6] H. Ying, C. Silex, A. Schnitzer, S. Leonhardt, and M. Schiek, Automatic Step Detection in the Accelerometer Signal, International Workshop Body Sensor Networks, 2007.

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

[8] Jean Le Feuvre, Cyril Concolato, and Jean-Claude Moissinac., GPAC: open source multimedia framework. In Proceedings of the 15th international conference on Multimedia '07. ACM, New York, USA, 2007

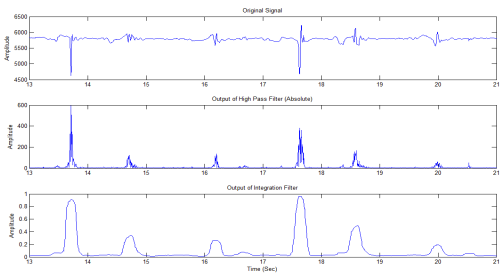

Figure 4: WIMU step detection process

Figure 5: Our Kinect calibration pose (left), applied to the pre-recorded depth map for Anne-c1-t1 (centre), giving a skeletal overlay (right).

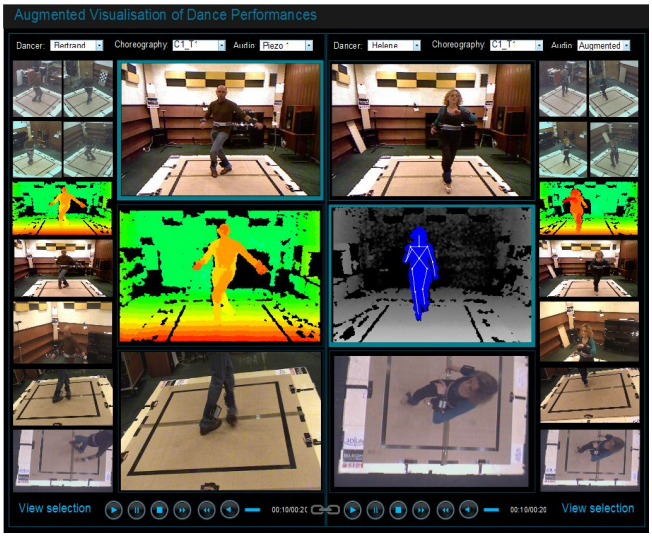

Figure 6: Snapshot of the application GUI