A multi-modal dance corpus for research into real-time interaction between humans in online virtual environments

Slim Essid†, Xinyu Lin⋄, Marc Gowing‡, Georgios Kordelas⋆, Anil Aksay⋄, Philip Kelly‡, Thomas Fillon†, Qianni Zhang⋄, Alfred Dielmann†, Vlado Kitanovski⋄, Robin Tournemenne†, Noel E. O'Connor‡, Petros Daras⋆ and Gaël Richard†

†Institut Telecom / Telecom ParisTech, CNRS-LTCI - Paris, France
⋄Multimedia and Vision Group (MMV), Queen Mary University London, UK
‡CLARITY: Centre for Sensor Web Technologies, Dublin City University, Ireland
⋆Centre for Research and Technology, Informatics and Telematics Institute, Thessaloniki, Greece

ABSTRACT

We present a new, freely available, multimodal corpus for research into, amongst other areas, real-time realistic interaction between humans in online virtual environments. The specific corpus scenario focuses on an online dance class application scenario where students, with avatars driven by whatever 3D capture technology are locally available to them, can learn choerographies with teacher guidance in an online virtual ballet studio. As the data corpus is focused on this scenario, it consists of student/teacher dance choreographies concurrently captured at two different sites using a variety of media modalities, including synchronised audio rigs, multiple cameras, wearable inertial measurement devices and depth sensors. In the corpus, each of the several dancers perform a number of fixed choreographies, which are both graded according to a number of specific evaluation criteria. In addition, ground-truth dance choreography annotations are provided. Furthermore, for unsynchronised sensor modalities, the corpus also includes distinctive events for data stream synchronisation. Although the data corpus is tailored specifically for an online dance class application scenario, the data is free to download and used for any research and development purposes.

1. INTRODUCTION

The 3DLife Network of Excellence [1] is a European Union funded research project that aims to integrate research that is currently conducted by leading European research groups in the field of Media Internet. Within 3DLife we believe that it is time to move social networking towards the next logical step in its evolution: to immersive collaborative environments that support real-time realistic interaction between humans in online virtual and immersive environments.

To achieve this goal 3DLife, partnered by Huawei [5], have proposed a grand challenge to the research community in conjunction with the ACM Multimedia Grand Challenge 2011 [3]. The ACM Multimedia Grand Challenges are a set of problems and issues from industry leaders, geared to engaging the research community in solving relevant, interesting and challenging questions about the industry's 2-5 year horizon. The 3DLife grand challenge calls for demonstrations of technologies that support real-time realistic interaction between humans in online virtual environments. In order to stimulate research activity in this domain the 3DLife consortium have provided a scenario for online interaction and a data corpus to support both the investigation into potential solutions and allow demonstrations of various technical components.

More specifically, the proposed scenario considers that of an online dance class, to be provided by an expert Salsa dancer teacher and delivered via the web. In this scenario, the teacher will perform the class, with all movements captured by a state of the art optical motion capture system. The resulting motion data will be used to animate a realistic avatar of the teacher in an immersive online virtual ballet studio. Students attending the online master-class will do so by manifesting their own individual avatar in the virtual dance studio. The real-time animation of each student's avatar will be driven by whatever 3D capture technology is available to him/her. This could be captured via visual sensing techniques using a single camera, a camera network, wearable inertial motion sensing, or recent gaming controllers such as the Nintendo Wii or the Microsoft Kinect. The animation of the student's avatar in the virtual space will be real-time and realistically rendered, subject to the granularity of representation and interaction available from each capture mechanism.

In this paper, we present the novel annotated data set that accompanies this grand challenge. This free and publicly available data corpus consists of data gathered at two separate site locations, at each site multimodal recordings of 15 Salsa dancers were captured with a variety of equipment, with each dancer performing between 2 to 5 fixed choreographies. The recording modalities captured in each recording setup include multiple synchronised audio capture, depth sensors, several visual spectrum cameras and inertial measurement units. In addition, the publicly available data set contains a rich set if dance choreography ground-truth annotations, including dancer ratings, plus the original music excerpts to which each dancer was performing to. In addition, as not all data stream modalities are synchronised, the corpus incorporates means to synchronise all of the input streams, via distinctive clap motions performed before each dance rendition.

Although created specifically for the ACM Multimedia Grand Challenge 2011, the data corpus is free to be used for other research and development purpose. This could include research into approaches for 3D signal processing, computer graphics, human computer interaction and human factors. This could include, but is not limited to:

- 3D data acquisition and processing from multiple sensor data sources;
- Realistic (optionally real-time) rendering of 3D data based on noisy or incomplete sources;
- Realistic and naturalistic marker-less motion capture;
- Human factors around interaction modalities in virtual worlds;
- Multimodal dance performance analysis, including dance steps/movements tracking, recognition and quality assessment;
- Audio/Video synchronisation with different capture devices;
- Extraction of features to analyse dancer performance, such as the automatic localisation and timing of foot steps or automatic extraction of dancer movement fluidity, timing, precision (to model) and alignment with the music, or another performer;
- Automatic extraction of music information such as tempo, beats time (1, 2, 3, 4), musical structure.

This paper is organised as follows: Section 2 highlights related data corpuses and the major difference in the one presented in this work. Section 3 provides an overview of the data captured and incorporated into the corpus for each dance performance. Section 6 details the hardware setup and capture of all data modalities used within the corpus. Section 4 provides an insight to how each dance performance was captured in terms of rehearsal, performance and capture. The choreographies used in the corpus are detailed in Section 5, while the ground-truth choreography annotations provided with the corpus are outlined in Section 7. In Section 8 we provide details on the data post-processing and release to the community. Finally we provide a discussion on the corpus in section 9.

2. RELATED WORK

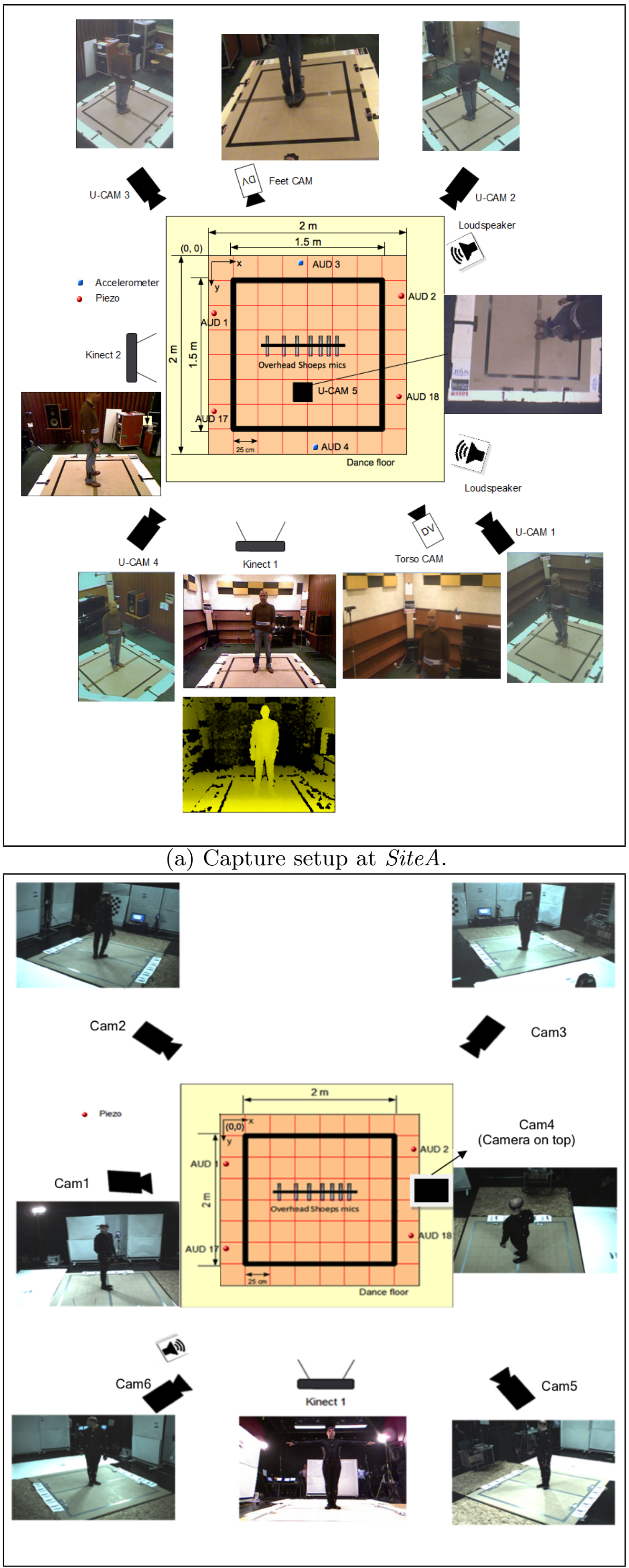
A number of publicly available datasets could be found that contain human actions captured by multiple sychronised cameras, and in some cases also captured by a motion capture rig. However, to the authors knowledge, no released dataset in the community consists of human motions (in all sequences) and inter-human interactions (in some sequences) concurrently by multiple diverse modalities capturing the visual spectrum, audio, inertial motion and depth information. We believe that all these captured modalities are required if research groups in diverse topics wish to work together towards real-time interaction between humans in online immersive virtual environments. For the remainder of this section, we will review of other datasets available to the community. The HumanEva-I dataset [11] contains 7 calibrated video sequences that are synchronized with 3D body poses obtained from motion capture. The database contains 4 subjects performing a 6 common actions (e.g. walking, jogging, gesturing, etc.). The dataset contains training, validation and testing (with withheld ground truth) sets. The i3DPost Multi-view Human Action Dataset [9] is a similar data corpus containing multi-view/3D human action/interaction data. It contains 8 synchronised HD image sequences of 8 people performing 13 common actions (e.g. Walk, Run, Jump, Bend, Hand-wave, etc). The CMU Motion of Body (MoBo) database [10] contains 25 individuals walking on a treadmill in the CMU 3D room. The subjects perform four different walk patterns: slow walk, fast walk, incline walk and walking with a ball. All subjects are captured using six high resolution colour cameras distributed evenly around the treadmill. In [10] is described the capture setup, the collection procedure and the organisation of the database. The Multiple-Camera/Multiple-Video Database [7] of the PERCEPTION group is a database for computer vision and video-based rendering research and experiments. This database contains a set of calibrated and synchronized video sequences. Each dataset in the database comes with the raw videos, the camera calibration files, the silhouettes extracted using background subtraction, as well as the associated 3-D model obtained from these images by using multiple-camera reconstruction software based on visual hulls. In all the previous data corpuses, a static background was assumed but the MuHAVi [12] human action video database has been collected using multi-cameras in a challenging environment. The raw images in the dataset can be used for different types of human action recognition methods as well as a dataset to evaluate robust object segmentation algorithms.

3. DATA CORPUS OVERVIEW

Within the data corpus, dance performances were captured at two separate sites. The setup for each site differs slightly in terms of equipment used and equipment locations, however the following data was captured and provided for each dance choreography regardless of the site – an overview of the two multi-modal capture setups (one for each data capture site) is provided in figure 1. Details of all equipment setup will be described in section 6.

- Synchronised multi channel audio capture of dancers' step sounds, voice and music;
- Synchronised camera video capture of the dancers from multiple viewpoints covering whole body
- Inertial sensor data: captured from multiple sensors on the dancer's body;
- Depth maps for dancers' performances: captured using a Microsoft Kinect;
- Original music excerpts;
- Camera calibration data;
- Different types of ground-truth annotations.

In addition, at one of the two capture sites, dancers were also simultaneously captured using four additional non-synchronised video captures covering a number areas of the dancers body.



(a) Capture setup at *SiteA*.

(b) Capture setup at *SiteB*.

Figure 1: Recording setup.

4. RECORDING PROTOCOL

Each dancer was recorded multiple times performing each time one of 5 pre-defined 5 choreographies. With every new dancer the recording session started by a preparation phase during which the he/she was equipped with the wearable recording devices and given instructions regarding the proceedings of the recordings and the choreographies to be performed (see Section 5). Next, the dancer was given time to rehearse these choreographies until he/she felt ready to be recorded. Only the choreographies that could be mastered by the dancer (after a reasonable rehearsing time that varied from 5 to 30 minutes for each choreography) were hence recorded. For each choreography a number of takes were captured to account for potential defects.

The recording started with the calibration of the camera network which was repeated various times during the whole sessions to ensure that the calibration data was reliably refined over time. It was performed using a 5x4 squared chessboard calibration pattern with square size of 15cm. The square size was set to be large enough so that the chessboard pattern was depicted clearly in the video of the cameras. This pattern was placed on the dancing stage.

While the signals captured by some subsets of sensors are perfectly synchronised, namely all audio channels (except the audio streams of the mini DV cameras), and a number of the camera videos, synchronisation is not ensured across all streams of data. To minimise this inconvenience, all dancers were instructed to execute a ”clap procedure” before starting their performance, where they successively clap their hands and tap the floor with each foot. Hence, the start time of each data stream can be synchronised (either manually or automatically) by aligning the clap signatures that are clearly visible within a 2 second time window from the beginning of every data stream (see for instance audio clap signatures on audio signals).

5. MUSIC AND CHOREOGRAPHIES

Salsa music was chosen for this data corpus as it is a music genre that is centred at dance expression, with highly structured, yet not straightforward rhythmic structures. The music pieces used were chosen from the Creative Commons set of productions to allow us to easily make them publicly available. Three short excerpts from two distinct tracks (of two distinct albums) at different tempos were created and used along with a Son Clave rhythmic pattern [4] in all dance sessions.

Each dancer performed 2 to 5 solo Salsa choreographies among a set of 5 pre-defined ones. These choreographies were designed in such a way to progressively increase in complexity the dance steps/movements as one moves from the first to the last one. They can be roughly described as follows:

C1 4 Salsa basic steps (over two 8-beat bars), where no music is played to the dancer, rather, he/she voice-counts the steps: ”1, 2, 3, 4, 5, 6, 7, 8, 1, ..., 8” (in French or English).

C2 4 basic steps, 1 right turn, 1 cross-body; danced on a Son clave excerpt at a musical tempo of 157 BPM (beats per minute).

C3 5 basic steps, 1 Suzie Q, 1 double-cross, 2 basic steps; danced on Salsa music excerpt labelled C3 at a musical

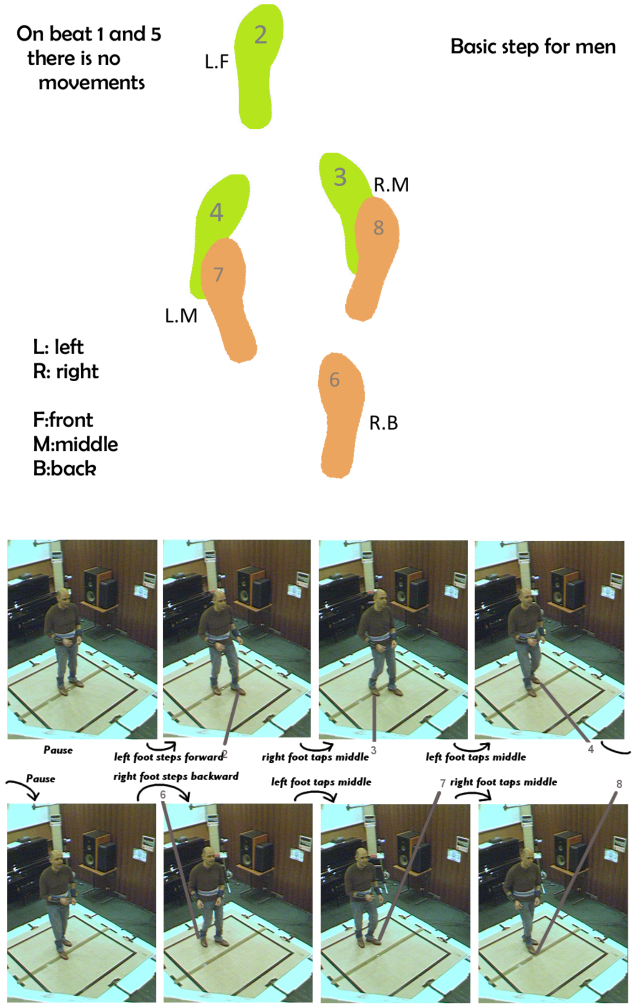


Figure 2: Basic Steps for Men.

tempo of 180 BPM.

C4 4 basic steps, 1 Pachanga tap, 1 basic step, 1 swivel tap, 2 basic steps; danced on Salsa music excerpt labelled C4 at a musical tempo of 185 BPM.

C5 A solo performance mimicking a duo, in the sense that the girl or the boy is asked to perform alone movements that are supposed to be executed with a partner. The movements are: 2 basic steps, 1 cross-body, 1 girl right turn, 1 boy right turn with hand swapping, 1 girl right turn with a caress, 1 cross-body, 2 basic steps; danced on Salsa music excerpt labelled C5 at a musical tempo of 180 BPM. Figure 2 gives visualisations of the timing of basic steps for men.

C6 Whenever possible a real duo rendering of choreography C5 has been captured. It is referred to as C6 in the data repository.

The dancers have been instructed to execute these choreographies respecting the same musical timing, i.e. all dancers are expected to synchronise steps/movements to particular music beats. All the song excerpts are provided in the database at 44,1KHz stereo.

It is also important to note that the dancers have been asked to perform a Puerto Rican variant of Salsa, and are expected to dance ”on two”.

15 dancers (6 women and 9 men) of differing expertise have been recorded at *SiteA* and 11 dancers (6 women and 5 man) at *SiteB*. Bertrand is considered as the reference dancer for men and Anne-Sophie K. as the reference dancer for women, in the sense that their performances are considered to be the ”templates” to be followed by the other dancers.

6. RECORDING EQUIPMENT SETUP

The specifics of each capture modality will be described in detail in the following sections using Figure 1 as reference. It should be noted that all data is recorded and provided in open formats.

6.1 Audio equipment

The audio capture setup was designed to capture the dancer’s voice and step-impact sounds in such a way to allow users of the dataset effectively exploit sound source localisation and separation technologies. The environments at *SiteA* and *SiteB* were recorded using 16 and 14 perfectly synchronised channels respectively. Eight microphones were placed around the dance capture area: seven Schoeps omnidirectional condenser microphones: placed overhead of the dance area; and one Sennheiser wireless lapel microphone positioned to capture the dancer’s voice. In addition, on-floor acoustic sensors were used to focus on the dancer’s step-impact sounds, namely four acoustic-guitar internal Piezo transducers, and only at *SiteA* Bruel & Kjaer 4374 piezo-electric accelerometers (used with a charge conditioning amplifier unit with two independent input channels).

Recording was performed using two Echo Audiofire Pre8 firewire digital audio interfaces controlled by a server based on Debian with a real-time patched kernel that run an open-software solution based on Ffado, Jack and a custom application for batch sound playback and recording. Accurate synchronisation between multiple Audiofire Pre8 units was ensured through Word Clock S/PDIF.

All the channels are encoded in separate files in mono at 48kHz with a 24-bit precision (but the sample encoding in the corresponding files is 32-bit Floating Point PCM). The on-floor positions of the Bruel & Kjaer and Piezo sensors, as well as the spacing between the Shoeps microphones are provided in the data corpus. The music was played to the dancers by a PC through amplified loudspeakers placed in the dance rooms as shown in Figure 1.

6.2 Video equipment

6.2.1 *Synchronised Video Equipment*

For the capture at *SiteA*, 5 firewire CCD cameras (Unibrain Fire-i Color Digital Board Cameras) were connected to a server with two FireBoard-800 1394b OHCI PCI adapters installed. Three cameras were connected to one PCI FireBoard-800 adapter, and two to the second, thereby allowing the network load to be distributed between the two adapters. The server had the UbCore 5.72 Pro synchronisation software installed, which provided the interface for the centralised control of the connected cameras, including the synchronized video capturing and the adjustment of the capturing parameters. The parameters of the video capture were defined to be 320x160 pixels at 30 frames per second with colour depth of 16 bits. In the data set, the Unibrain camera data was decoded from MJPEG to raw AVI and stored as ZIP archives. However, as the camera synchronisation at *SiteA* was controlled by software and therefore it was not perfectly accurate. As a consequence very slight variations appeared in the total number of the frames recorded by each synchronized camera. This is discussed and corrected in the post-processing stage – see Section 8.1.

The equipment for *SiteB* is different however, with the cameras synchronized via hardware. At *SiteB*, the viewpoints of *U-Cam 1* to *U-Cam 5* and *Kinect 2* being replicated by 6 PixeLink 1.3 mega pixel color PL-B742 cameras, labelled *Cam1* to *Cam6* in Figure 1(b). The PixeLink cameras were synchronized using a common triggering signal, which was a square waveform signal generated by a digital function generator and a triggering frequency set to be 15Hz. Each cycle triggered the capture of single image frame for each camera. All captured frames using thee cameras are stored in BMP format in the data set.

6.2.2 *Non-synchronised Video Equipment*

For the *SiteA* data capture, two standalone, non-synchronised, digital video cameras (both with audio) were used to capture the dancers from differing angles. The first shooting the dancers' feet (audio specification: PCM S16 Stereo, 16 bits, 32000 Hz), with the second DV camera shooting the torso (audio specification: PCM S16 Stereo, 16 bits, 48000 Hz). In addition, at *SiteA* two additional non-synchronised video data streams were also acquired using Microsoft Kinect cameras. The first Kinect camera was angled to cover the whole of the dancer's body from the front, while the second was angled to the upper-body of the dancer and taken from the side. In *SiteB* only one of the four non-synchronised streams was replicated, with the first Kinect camera angle being recaptured. In this dataset both the Kinect cameras were captured at circa 30Hz and stored using the OpenNI-encoded (.ONI) data format (see next section). The videos from both the digital cameras were first stored on tapes before being transferred to a PC using a proprietary application. They were encoded using the cameras native DV video codec with 720 × 576 pixels at 25 frames per second, with the audio streams encoded as PCM S16 stereo at 32kHz and 48kHz respectively for the feet and torso cameras.

6.2.3 *Kinect Depth Stream*

In both of the data capture sites a Kinect depth data stream was acquired from *Kinect 1* (see figure 1(a)). This data stream was synchronised with the Kinect video stream (described in the previous section) and both were simultaneously captured and stored using the OpenNI drivers/SDK and the OpenNI-encoded (.ONI) data format [6]. 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 track- ing 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. A overlay of the extracted skeleton (using the OpenNI SDK) on the Kinect depth stream can be seen in Figure 3.

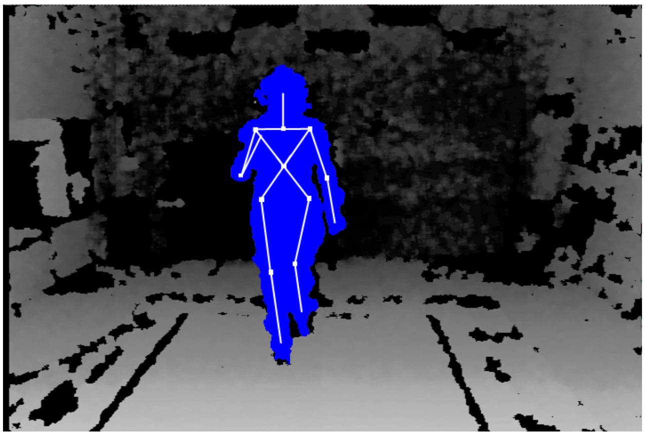


Figure 3: Skeleton tracking for the dancer *Helene*.

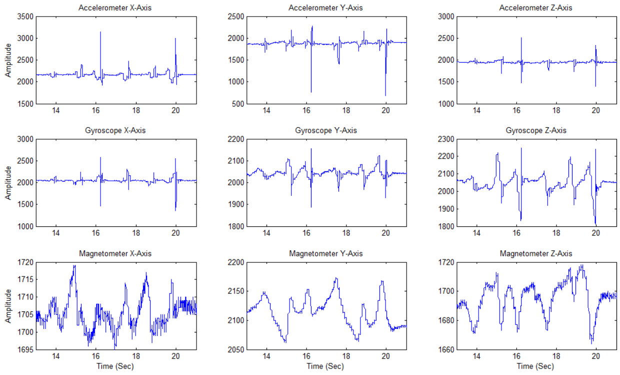


Figure 4: Inertial sensor data for right ankle of dancer *Bertrand*.

6.3 Inertial measurement units

Data from inertial measurement units (IMUs) were also captured with each dance sequence. Each sensor streamed accelerometer, gyroscope and magnometer data at approximately 80 - 160 Hz. Five IMUs were placed on each dancer; one on each dancer's forearm, one on each dancer's ankle, and one above their hips. Each IMU provides time-stamped accelerometer, gyroscope and magnetometer data for their given location at 80 - 160 Hz for the duration of the session and stored as raw ASCII text. A sample of the IMU data is shown in Figure 4.

7. GROUND-TRUTH ANNOTATIONS

Various types of ground-truth annotations are provided with the data corpus, namely:

- Manual annotations of the music in terms of beats and measures, performed by a musician familiar with the salsa rhythm, given in Sonic Visualiser [8] (.svl) format and ASCII (.cvs) format;

- Annotations of the choreographies with reference steps time codes relative to the music also given in Sonic Visualiser (.svl) format and ASCII (.cvs) format, these annotations were acquired using the teachers' input and that indicate the labels of the salsa movements

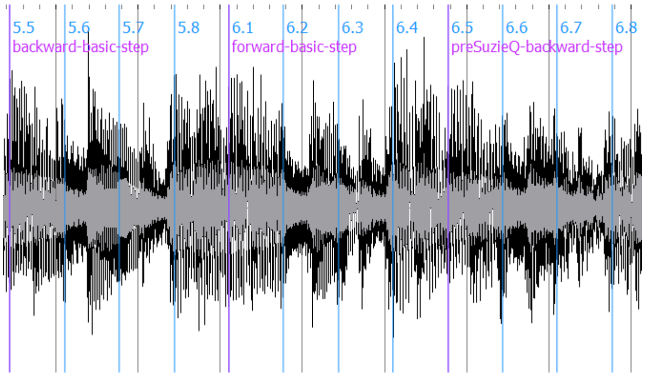


Figure 5: Beat, measures and Choreography annotations.

to be performed with respect to the musical timing. An example of this type of annotation is depicted in Figure 5;

- Ratings of the dancers' performances assigned to dancers by the teachers.

The dancers' ratings are given as an integer score between 1 and 5, 1 being poor and 5 excellent, across five evaluation axes:

Upper-body fluidity evaluates the fluidity of the dancer's upper-body movements;

Lower-body fluidity evaluates the fluidity of the dancer's upper-body movements;

Musical timing evaluates the timing of the executed choreography movements/steps with respect to the music timing, the ideal timing being given in the choreography annotation files given in the music/ folder;

Body balance evaluates the state of balance or quality of equilibrium of the dancer's body while he/she executes the choreography;

Choreography evaluates the accuracy of the executed choreography; a rating of 5 is attributed to a dancer as soon as he/she accurately reproduces the sequence of figures/steps of the choreography, quite independently from the quality of execution of each single figure.

8. DATA PREPARATION AND RELEASE

A number of post-processing stages were undertaken in order to ease the use of the data corpus. Firstly, only valid recording takes were incorporated into the corpus, where we considered as valid any take during which the dancer could finish the execution of the whole choreography (without stopping in the middle), and all modalities could be captured properly (without any technical defects). Secondly, the various streams of data were visually inspected and data manually edited to crop out irrelevant content ensuring the clap event (described in Section 4) would occur within two seconds from the beginning of each recording modality. As such, although some of the data streams are not fully synchronised, the maximum offset of any one modality to another is set to two seconds, allowing users to more easily use multiple sets of unsynchronised data modalities.

8.1 Unibrain Capture Post-processing

As outlined in section 6.2.1, the camera synchronisation at *SiteA* was controlled by software and therefore is not perfectly accurate. As a consequence very slight variations appeared in the total number of the frames recorded by each synchronized camera. These variations were caused by the delays in the time required to propagate the commands of starting and stopping the synchronized capturing to each camera. Hence, the following post-processing procedure was applied to compensate for the recording start and stop delays across the camera network.

Let us assume that the total number of the recorded frames by each of 5 cameras (*U-Cam 1* to *U-Cam 5*) is N_1 to N_5, while N_3 is the minimum number of frames that will be used as a common basis to equalise the number of frames recorded by the rest of the cameras. For instance, in order to compensate delay in the video recorded by *U-Cam 1*, so as to have the same number of frames with the video recorded by *U-Cam 3*, when $N_1 - N_3$ is an even number then $(N_1 - N_3)/2$ frames are removed from the beginning of *U-Cam 1*'s frame sequence and $(N_1 - N_3)/2$ frames from the end of the sequence. Otherwise in case $N_1 - N_3$ is odd, then $(N_1 - N_3 + 1)/2$ frames are removed from the start and $(N_1 - N_3 - 1)/2$ frames from the end of the sequence. The same procedure is applied to frame sequences recorded by *U-Cam 2*, *U-Cam 4* and *U-Cam 5*, respectively. Afterwards, the post-processed recordings have equal number of frames.

This is statistically the most possible solution to deal with the delay compensation problem since it is most likely that the redundant frames per captured video are equally split between the start and the end of each capturing sequence.

8.2 Data Release

Since May 2011, the data corpus for the 3DLife ACM Multimedia Grand Challenge 2011 have been made publicly available through a website [2], allowing anyone to download it through FTP. Researchers are also free to submit work for publication to any relevant conferences/journals/etc. outside of ACM Multimedia 3DLife Grand Challenge 2011, as long as the publication date occurs after the grand challenge has been completed (December 1st 2011).

9. DISCUSSION

In this work, we presented a new multimodal corpus for research into, amongst other areas, real-time realistic interaction between humans in online virtual environments. Although the data set is tailored specifically for an online dance class application scenario, the data corpus provides scope to be used by research and development groups in a variety of areas. As a research asset the corpus provides a number of features that make it appealing including; it is free to download and use; it provides both synchronised and unsynchronised multichannel and multimodal recordings; the novel recording of dancer sound steps amongst other specific sound sources; depth sensor recordings; incorporation of wearable inertial measurement devices; a large number of performers; a rich set of ground-truth annotations, including performance ratings. Due to these features, and others, we believe that the provided data corpus can be used to illustrate, develop and test a variety of tools in a diverse number of technical areas.

Acknowledgments

The authors and 3DLife would like to acknowledge the support of Huawei in the creation of this dataset. In addition, warmest thanks go to all the contributors to these capture sessions, especially:

The Dancers Anne-Sophie K., Anne-Sophie M., Bertrand, Gabi, Gael, Habib, Helene, Jacky, Jean-Marc, Laetitia, Martine, Ming-Li, Remi, Roland, Thomas.

The Tech Guys Alfred, Dave, Dominique, Fabrice, Gael, Georgios, Gilbert, Marc, Mounira, Noel, Phil, Radek, Robin, Slim, Qianni, Sophie-Charlotte, Thomas, Xinyu, Yves.

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

10. REFERENCES

[1] 3dlife. http://www.3dlife-noe.eu, 2011.

[2] 3dlife dance dataset description website. http://perso.telecom-paristech.fr/ẽssid/3dlife-gc-11/, 2011.

[3] 3dlife/huawei grand challenge 2011. http://www.3dlife-noe.eu/3DLife/emc2/grand-challenge/, 2011.

[4] Clave (ryhthm). http://en.wikipedia.org/wiki/Clave_rhythm, 2011.

[5] Huawei. http://www.huawei.com, 2011.

[6] Openni. http://www.openni.org/, 2011.

[7] Perception group database. http://4drepository.inrialpes.fr/, 2011.

[8] C. Cannam, C. Landone, and M. Sandler. Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the ACM Multimedia 2010 International Conference*, pages 1467–1468, Firenze, Italy, October 2010.

[9] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas. The i3dpost multi-view and 3d human action/interactions. In *6th Conference on Visual Media Production (CMVP)*, pages 159–168, 2009. http://kahlan.eps.surrey.ac.uk/i3dpost_action/.

[10] R. Gross and J. Shik. The cmu motion of body (mobo) database. Technical report, 2001. http://www.mendeley.com/research/the-cmu-motion-of-body-mobo-database-1/.

[11] L. Sigal and M. Black. Humaneva: synchronized video and motion capture dataset for evaluation of articulated human motion. Technical report, 2006. http://vision.cs.brown.edu/humaneva/.

[12] S. Singh, S. Velastin, and H. Ragheb. Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In *Advanced Video and Signal Based Surveillance*, pages 48–55, 2010. http://dipersec.king.ac.uk/MuHAVi-MAS/.