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

Slim Essid, Xinyu Lin, Marc Gowing, Georgios Kordelas, Anil Aksay, Philip Kelly, Thomas Fillon, Qianni Zhang, Alfred Dielmann, et al.

Journal on Multimodal User Interfaces

ISSN 1783-7677
Volume 7
Combined 1-2

J Multimodal User Interfaces (2013)
7:157-170
DOI 10.1007/s12193-012-0109-5

Springer

J Multimodal User Interfaces (2013) 7:157–170
DOI 10.1007/s12193-012-0109-5

ORIGINAL PAPER

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

Slim Essid · Xinyu Lin · Marc Gowing · Georgios Kordelas · Anil Aksay · Philip Kelly · Thomas Fillon · Qianni Zhang · Alfred Dielmann · Vlado Kitanovski · Robin Tournemenne · Aymeric Masurelle · Ebroul Izquierdo · Noel E. O'Connor · Petros Daras · Gaël Richard

Received: 1 March 2012 / Accepted: 16 July 2012 / Published online: 24 August 2012

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 is locally available to them, can learn choreographies with teacher guidance in an online virtual dance studio. As the dance 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 performs a number of fixed choreographies, which are 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. The total duration of the recorded content is 1 h and 40 min for each single sensor, amounting to 55 h of recordings across all sensors. Although the dance corpus is tailored specifically for an online dance class application scenario, the data is free to download and use for any research and development purposes.

Keywords Dance · Multimodal data · Multiview video processing · Audio · Depth maps · Motion · Inertial sensors · Synchronisation · Activity recognition · Virtual reality · Computer vision · Machine listening

1 Introduction

The 3DLife Network of Excellence 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 augmented environments.

To achieve this goal 3DLife, partnered by Huawei, has proposed a grand challenge to the research community in conjunction with the ACM Multimedia Grand Challenge 2011. The ACM Multimedia Grand Challenges are a set of problems and issues from industry leaders, geared to engaging the research community in adressing relevant, interesting and challenging questions about the industry's 2–5 years 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 has provided a scenario for online interaction and a corpus to support both the investigation into potential

S. Essid (✉) · T. Fillon · A. Dielmann · R. Tournemenne · A. Masurelle · G. Richard
Institut Telecom/Telecom ParisTech, CNRS-LTCI, Paris, France
e-mail: slim.essid@telecom-paristech.fr

X. Lin · G. Kordelas · A. Aksay · Q. Zhang · V. Kitanovski · E. Izquierdo
Multimedia and Vision Group (MMV),
Queen Mary University, London, UK

M. Gowing · P. Kelly · N. E. O'Connor
CLARITY, Centre for Sensor Web Technologies,
Dublin City University, Dublin, Ireland

G. Kordelas · P. Daras
Centre for Research and Technology-Hellas,
Informatics and Telematics Institute, Thessaloníki, Greece

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, and/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 dataset that accompanies this grand challenge. This free and publicly available dance corpus consists of data gathered at two separate site locations. At each site multimodal recordings of Salsa dancers were captured with a variety of equipment, with each dancer performing between 2 and 5 fixed choreographies. 15 dancers (6 women and 9 men) of differing expertise have been recorded at *SiteA* and 11 dancers (6 women and 5 men) at *SiteB*. The recording modalities captured in each recording setup include multiple synchronised audio capture, depth sensors, several visual spectrum cameras and inertial measurement units. The total duration of the recorded content is 1 h and 40 min for each single sensor, amounting to 55 h of recordings across all sensors. In addition, this publicly available dataset contains a rich set of dance choreography ground-truth annotations, including dancer ratings, plus the original music excerpts to which each dancer was performing to.

Moreover, 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. These clap motions can be used to determine the delays between the different streams as will be described in Sect. 8.2. Such delays as found by a reference automatic system are provided along with the dataset.

Although created specifically for the ACM Multimedia Grand Challenge 2011, the corpus is free to be used for other research and development purposes. This could include research into approaches for 3D signal processing, computer graphics, computer vision, human computer interaction and human factors:

- 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, as a particular case of human activity 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 and beat analysis or musical structure analysis.

This paper expands upon the initial publication [10] by describing a multimodal synchronisation scheme that allows us to provide the delays between the unsynchronised streams of data, and by further developing potential applications for the dataset proposed.

The rest of this paper is organised as follows: Sect. 2 highlights related 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 Sect. 5, while the ground-truth choreography annotations provided with the corpus are outlined in Sect. 7. In Sect. 8, we provide details on the data post-processing and release to the community. Section 9 outlines possible fields of application for the current dataset. Finally we provide concluding remarks on the corpus in Sect. 10.

2 Related work

In this section, we review the datasets available to the community in the research fields related to our work.

The KTH database [19] contains 6 types of human actions performed by 25 people in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. The Weizmann human action dataset [6] contains 90 low-resolution video sequences showing 9 different people, each performing 10 natural actions. The backgrounds

are static and the foreground silhouettes are included in the dataset. KTH and Weizmann databases have been extensively used for the evaluation and comparison of single-view video human action recognition algorithms. The Assisted Daily Living (ADL) dataset [16] is another single-view database that contains high resolution videos of ten complex daily activities.

Regarding multi-view and 3D activity databases, several researches have created their own databases for verifying their implemented approaches.

The IXMAS [26] dataset contains 11 actions each performed 3 times by 10 actors. The multi-view acquisition is achieved using five Firewire cameras, with static background and illumination settings. Each dataset contains the raw videos, camera calibration files, extracted silhouettes using background subtraction, as well as the associated 3-D models obtained from these images by using multi-view camera reconstruction software based on visual hulls.

The HumanEva dataset [22] consists of two parts: HumanEva-I and HumanEva-II. HumanEva-I contains data from four subjects performing a set of six predefined actions in three repetitions (twice with video and motion capture, and once with motion capture alone). Each sequence is recorded by three colour and four grayscale cameras and a synchronized motion capture system that provides the 3D body poses of the subjects. HumanEva-II contains only two subjects performing an extended sequence of actions. However, for the capturing a more sophisticated hardware system is used than HumanEva-I, that consists of four colour cameras and a better quality motion capture system.

The i3DPost multi-view human action dataset [11] is a corpus containing multi-view 3D human action/interaction data. This database contains videos of 8 persons and 12 actions captured from 8 high resolution cameras. Moreover, there are sequences that capture the basic facial expressions of each person. The multi-view videos have been further processed to produce a 3D mesh at each frame describing the respective 3D human body surface.

MuHAVi human action video database [23] has been created using eight cameras in a challenging environment. The dataset includes 17 action classes performed by 14 actors. A subset of action classes has been used to manually annotate the image frames and generate the corresponding silhouettes of the actors. Annotated silhouettes provide a useful ground truth for scientists to evaluate their algorithms.

The CMU motion capture database [13] mainly aims at advancing research on human gait as a biometric. The database contains 25 individuals performing 4 different walking patterns on a treadmill. All subjects are captured using six high resolution colour cameras distributed evenly around the treadmill.

CASIA action database [24] is a collection of sequences of human activities captured outdoors by cameras from different angle of view. The sequences include 8 types of actions performed by 24 subjects and 7 types of 2 person interactions performed by 2 subjects. Videos sequences are recorded simultaneously with three static non-calibrated cameras from different viewing angles.

WARD (Wearable Action Recognition Database) [27] consists of continuous sequences of human actions measured by a network of wearable motion sensors. The wireless sensors are instrumented at five body locations: two wrists, the waist and two ankles. There are 20 human subjects that produce a set of 13 action categories that covers some of the most common actions in a human's daily activities.

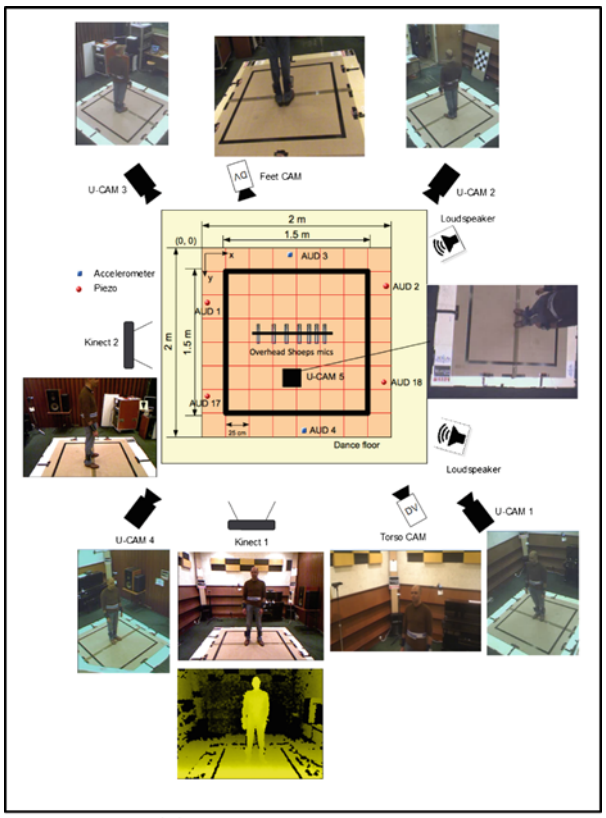
The majority of the aforementioned databases contain simple human actions captured by multiple synchronised cameras. However, to the best of the authors' knowledge, there has been no previous research datasets recorded concurrently through multiple diverse modalities capturing the visual spectrum, audio, inertial motion and depth information; nor has there been multimodal datasets focusing on complex types of human activites such as dance.

3 Corpus overview

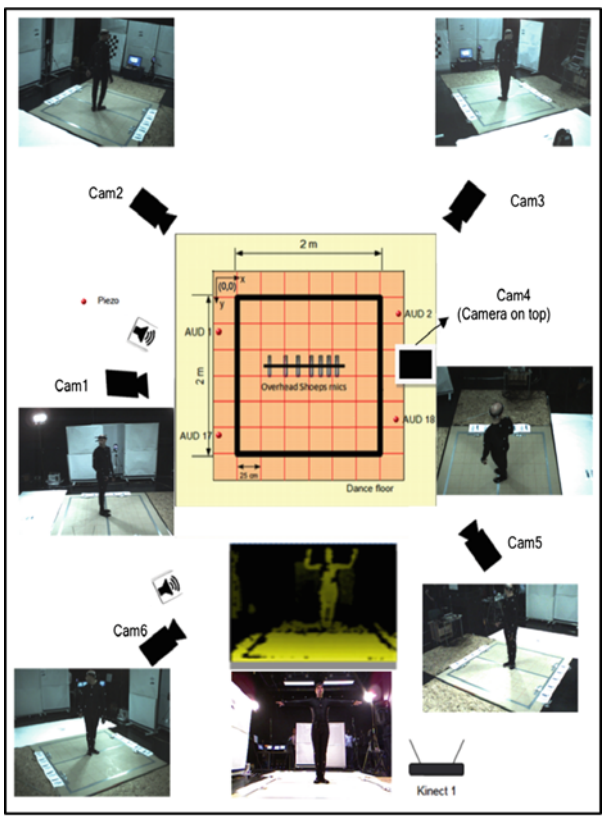
The dance corpus we present provides both synchronised and unsynchronised multi-channel and multi-modal recordings of Salsa dance students and teachers. Within the corpus, dance performances were captured at two seperate sites. An overview of the two multi-modal capture setups (one for each data capture site) is provided in Fig. 1. Details of all equipment setup will be described in Sect. 6.

The setup for each site differs slightly in terms of equipment specifications and equipment locations—however, the following data was recorded regardless of the recording site:

- Synchronized multi-channel audio of dancers' step sounds, voice and music. The environments at SiteA and SiteB recorded 16 and 14 channels, respectively, consisting of 7 overhead microphones, 1 lapel microphone worn by the dancer and the remaining channels recorded by on-floor piezo-electric transducers.
- Synchronised camera video capture of the dancers from multiple viewpoints covering whole body: SiteA and SiteB used five and six cameras, respectively.
- Inertial sensor data: captured from five sensors worn on the dancer's body: both wrists, both ankles and around the waist.
- Depth maps for dancers' performances: captured using a Microsoft Kinect.
- Original music excerpts: three short excerpts sampled from two distinct Salsa tracks.
- Camera calibration data.



(a) Capture setup at *SiteA*.



(b) Capture setup at *SiteB*.

Fig. 1 Recording setup

– Ground-truth annotations, including: musical annotation (tempo and beats), choreography annotation (step labels and ideal timing) and performance ratings awarded to each dancer by the teacher.

In addition, at capturing SiteA, dancers were also simultaneously captured using four additional non synchronised video captures covering a number of areas of their bodies.

The modalities that were synchronised during capturing include 16 channels audio data, multi-view videos captured with unibrain cameras (SiteA) and PixeLink cameras (SiteB). During post processing, synchronisation is achieved between audio and WIMU data, audio and different subsets of video data. Synchronization details will be more thoroughly discussed in Sect. 8.2.

A total of 26 subjects were recorded performing 2–5 predefined Salsa choreographies (depending on their level of ability). Multiple takes for each choreography are included in the corpus, with performances lasting approximately 20–40 s.

4 Recording protocol

Each dancer was recorded multiple times performing each time one of five pre-defined choreographies. For every new dancer, the recording session started with a preparation phase during which 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 Sect. 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 min for each choreography) were hence recorded. For each choreography a number of takes were captured to account for potential defects. The number of takes recorded varied from one dancer to another depending on their time availability. The goal was to try hard to obtain, for each choreography, at least two takes where the dancer would finish the whole choreography (without stopping in the middle).

The recording started with the calibration of the camera network, which was repeated at various times during the entire session to ensure that the calibration data was reliably refined over time. It was performed using a 5×4 squared chessboard calibration pattern with square size of 15 cm. 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 for the audio streams of the mini DV cameras), 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 at the beginning of every data stream.

5 Music and choreographies

Salsa music was chosen for this dance 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 extracted and used along with a forth excerpt consisting of a Son Clave rhythmic pattern [1] in all dance sessions. All the song excerpts used are provided in the database at 44.1 kHz stereo.

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

- C1 Four 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 Four basic steps, one right turn, one cross-body; danced on a Son clave excerpt at a musical tempo of 157 BPM (beats per minute).
- C3 Five basic steps, one Suzie Q, one double-cross, two basic steps; danced on Salsa music excerpt labelled C3 at a musical tempo of 180 BPM.
- C4 Four basic steps, one Pachanga tap, one basic step, one swivel tap, two 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: two basic steps, one cross-body, one girl right turn, one boy right turn with hand swapping, one girl right turn with a caress, one cross-body, two 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. 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".

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. The videos of *Bertrand* and *Anne-Sophie* were actually played to the student dancers during their training, asking them to mimic the performance of the reference dancers. It is worth noting that dance steps for men and women are not identical as they are designed to complement each other in the partnered dance routines.

6 Recording equipment setup

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

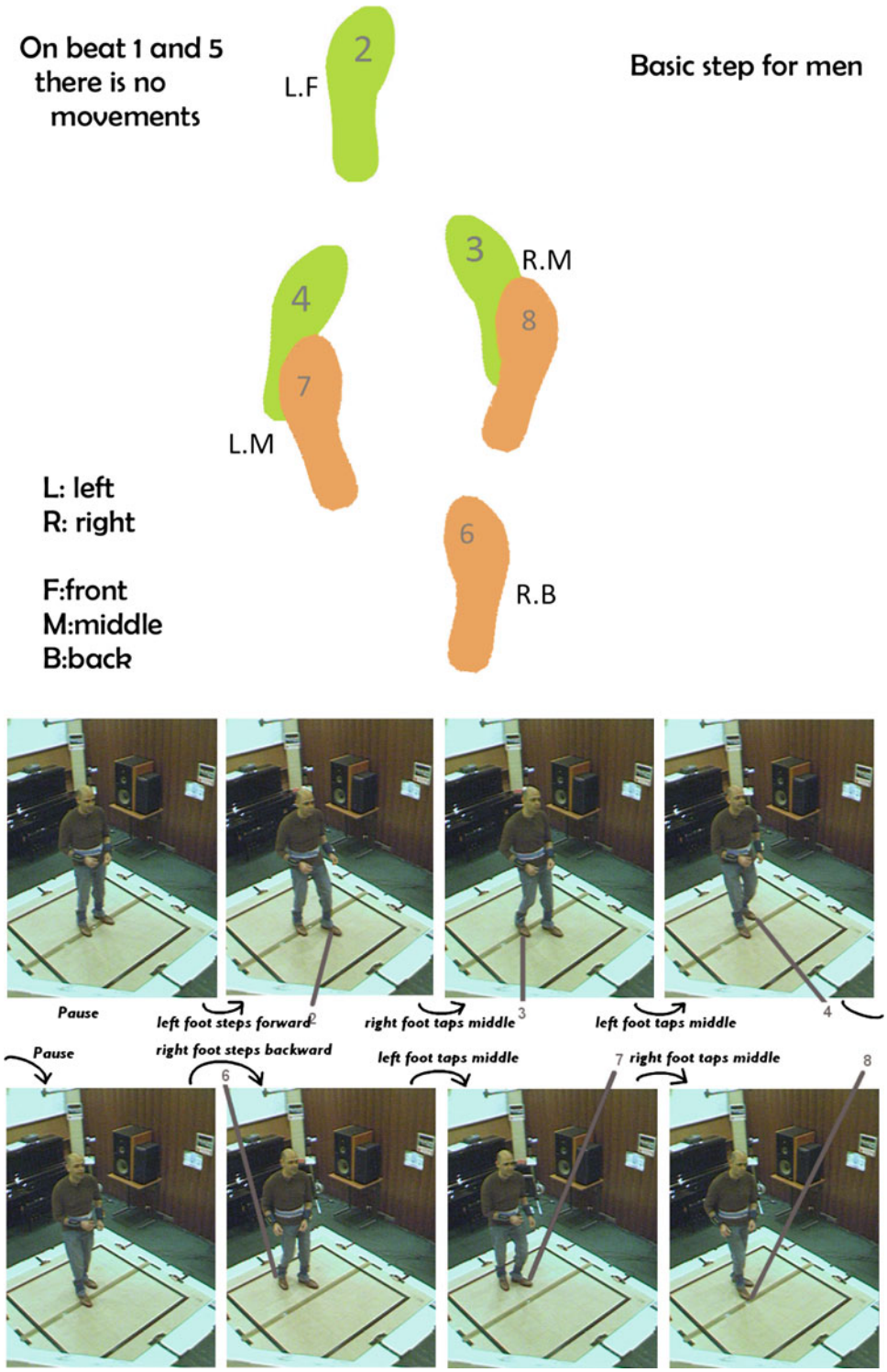
6.1 Audio equipment

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 to 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 piezoelectric accelerometers (used with a charge conditioning amplifier unit with two independent input channels). The position of the microphones and acoustic sensors is given in Fig. 1.

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 runs 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 were encoded in separate files in mono at 48 kHz with a 24-bit precision (but the sample encoding in the corresponding files is 32-bit Floating Point PCM

Fig. 2 Basic steps for men



in order to facilitate reading the generated audio files using standard audio software). The on-floor positions of the Bruel & Kjaer and Piezo sensors, as well as the spacing between the Shoeps microphones are provided in the corpus. The music was played to the dancers by a PC through amplified loudspeakers placed in the dance rooms as shown in Fig. 1.

Audio calibration During the recording at *SiteA*, audio calibration was performed in order to provide a way to localize the positions of the dancer's feet on the dance floor using step-impact sounds. This calibration procedure consisted of hitting the floor with a metal tip at different known locations on the dance floor board. For *SiteA*, the dance floor was

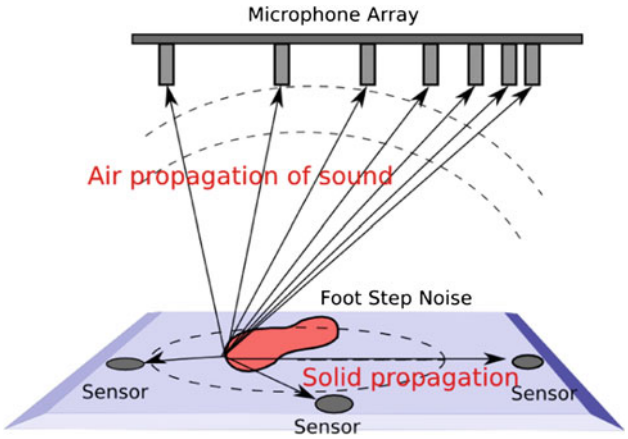


Fig. 3 Sound propagation of the foot steps noise through the dancing board to the vibration sensors and through the air to the microphone array

made up with two medium-density fiberboard (MDF) panels of 2 m × 1 m. Each panel was hit several successive times over a regular grid of points. Hence, we provide a set of audio calibration measurements at 92 positions on the 2 boards.

Based on this calibration, one can elaborate localization methods to retrieve the positions of the feet according to the audio signals of the steps on the dancing board. Due to the exact synchronization of the different audio channels and the fixed position of the sensors, the delays of arrival of the sound produced by a step at the different sensors can be used to determine the sound source localization and thus, the position of the foot on the dancing board. Figure 3 shows an illustration of the sound propagation from the sound source (the foot step) to the vibration sensors and to the microphone array. Such localization methods rely on a physical model of the sound propagation in the air and/or in the board material. In that case, the calibration signals can enable one to estimate the physical parameters of the models (speed of sound, delays or positions between the sound source and the sensors).

Alternate localization methods could rely on pattern recognition approaches. For such methods, the calibration signals could be considered as references or used as a training database for either classification or regression methods.

6.2 Video equipment

6.2.1 Synchronised video equipment

For the capture at *SiteA*, five 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 320 × 160 pixels at 30 frames/s with colour depth of 16 bits. In the dataset, the Unibrain camera data was decoded from MJPEG to raw AVI and stored as ZIP archives. However, 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 Sect. 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* were replicated by six PixeLink 1.3 mega pixel colour PL-B742 cameras, labelled *Cam1* to *Cam6* in Fig. 1b. 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 15 Hz. Each cycle triggered the capture of single image frame for each camera. All captured frames using these cameras are stored in BMP format in the dataset.

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, with the second DV camera shooting the torso. 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 30 Hz and stored using the OpenNI-encoded (.ONI) data format (see next section). The videos from both DV 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/s, with the audio streams encoded as PCM S16 stereo at 32 and 48 kHz, 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 Fig. 1a). This data stream

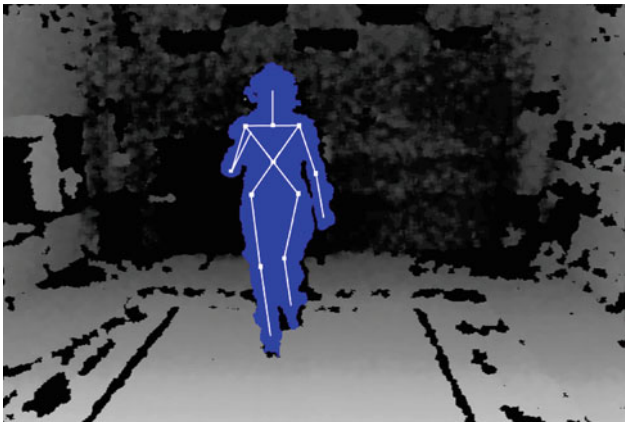


Fig. 4 Skeleton tracking for the dancer *Helene*

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 [2].

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. An overlay of the extracted skeleton (using the OpenNI SDK) on the Kinect depth stream can be seen in Fig. 4.

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 across the duration of the session. These measurements are stored as raw ASCII text. A sample of the IMU data is shown in Fig. 5.

7 Ground-truth annotations

Various types of ground-truth annotations are provided with the 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 [7] (.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; they indicate the labels of the salsa movements to be performed with respect to the musical timing. An example of this type of annotation is depicted in Fig. 6.
- Ratings of the dancers' performances assigned to dancers by the teacher Bertrand.[1]

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 placed 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 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 corpus. Firstly, only valid recording takes were incorporated into the corpus. 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 Sect. 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.

[1] More ratings by other experienced Salsa dancers will be provided in the near future

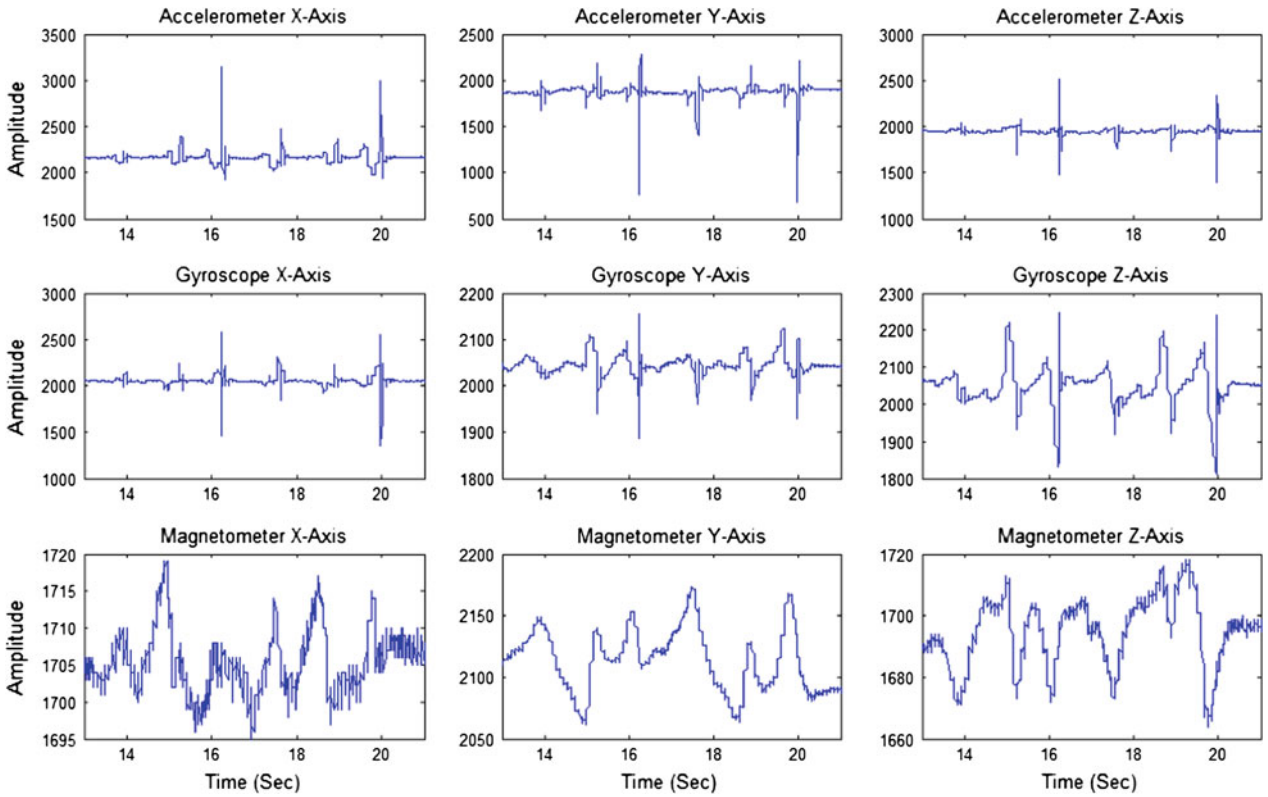


Fig. 5 Inertial sensor data for right ankle of dancer *Bertrand*

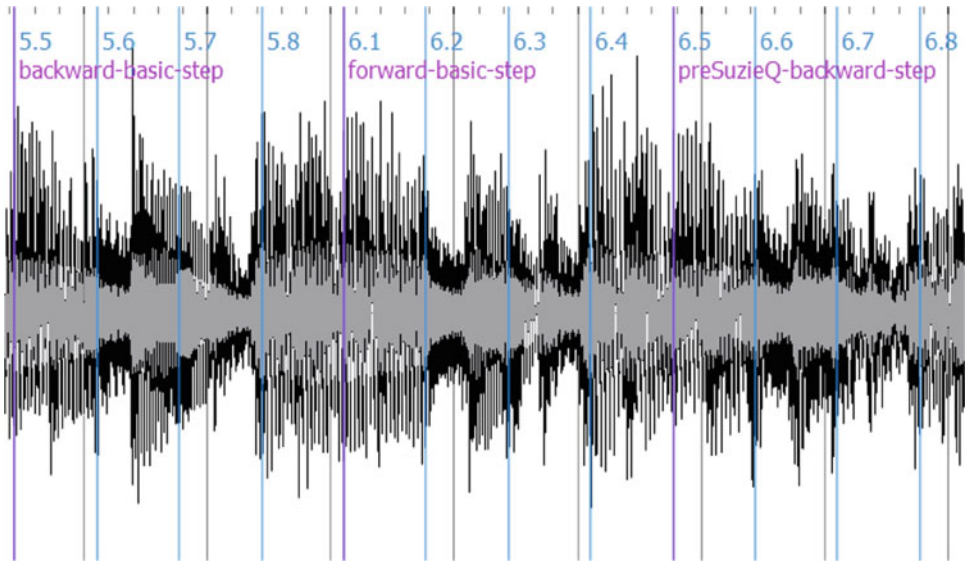


Fig. 6 Beat, measures and choreography annotations

8.1 Unibrain capture post-processing

The camera synchronisation at SiteA was controlled by software (see Sect. 6.2.1), which didn't provide perfect synchronisation. Inaccurate synchronisation was caused by delays in the time required for the software to propagate the commands of starting and stopping the capturing across the camera network. As a result, very slight variations (<12 frames) were appeared in the total number of the frames recorded by each camera. Based on technical specifications, the most likely possibility is that the redundant frames per captured video sequence were equally split between its beginning and its end.

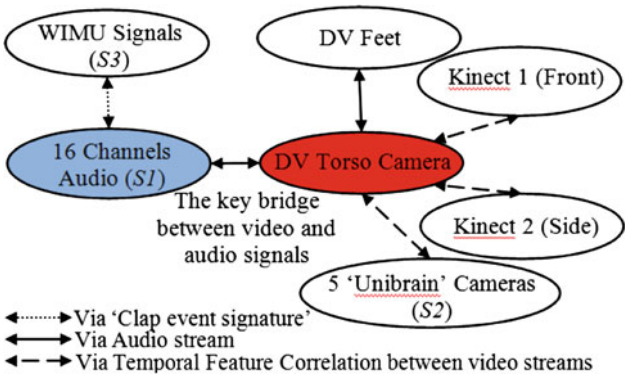


Fig. 7 Overview of synchronisation strategy

Hence, the following post-processing procedure was applied to remove the redundant frames from each captured video sequence.

Let us assume that the total number of the recorded frames by each of five 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, $(N_1 - N_3)/2$ 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, $(N_1 - N_3 + 1)/2$ 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. Thus, the resulting post-processed recordings have equal number of frames.

However, this post-processing procedure does not ensure that the video streams are perfectly synchronized, since it is based on the most possible circumstance that the redundant frames are equally split between the start and the end of the video sequence, which is true in most but not in all cases. Therefore, the synchronisation procedure described in Sect. 8.2 is followed in order to compensate more reliably the time delays that lead to variation in the total number of frames captured by each camera. Aside from video synchronisation, the procedure in Sect. 8.2 is actually used to synchronise all heterogeneous data streams.

8.2 Multi-modal synchronisation

Figure 7 gives an overview of our approach to synchronisation between the heterogeneous streams of data recorded.

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, these include: subset S1 consisting of data from 16 audio channels, subset S2 composed of video streams of 5 UniBrain cameras(not perfectly synchronised as mentioned in Sect. 8.1), and 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. As can be seen from Fig. 7, 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 (described in Sect. 8.2.1).
- 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 (described in Sect. 8.2.2).

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 Sect. 8.2.3.

8.2.1 Audio-based synchronisation

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 5-ms local audio frames with a 1-ms hop size. The sampling frequency of the envelopes is thus 1,000 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). The whole procedure has been validated by listening tests (across all the recordings) where a reference audio stream (from the feet camera) was played along with any of the other delayed streams to confirm that they became synchronous.

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 achieved using the previously described procedure.

8.2.2 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 both wrist sensors. 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 and maximum amplitudes, and their corresponding timestamps 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.

8.2.3 Multi-view video synchronisation

In this section, we describe the approach to calculate 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.

The temporal features used for video temporal alignment are based on appearance changes [25]. 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 for each frame and the values are interpolated on a 1-ms time grid in order to achieve sub-frame accuracy. The time shift between two videos is obtained as the value that maximises the normalised correlation between the temporal features of each video.

The frame region used to calculate the appearance change 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 to locate the dancer's upper-body movements, as some videos in the used dataset do not capture the whole body. Figure 8 shows temporal features for the DV torso camera and the Kinect camera, respectively.

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 [25] 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 [8] 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 8 (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 regions, the synchronisation accuracy is improved.

We evaluated the described approach using videos from the dataset. A total of 103 video sets were included in this evaluation (each video set includes: five UniBrain videos, torso video and feet video recorded by DV camcorders, and two videos recorded by Kinects). The average error of synchronisation was within one frame, which is quite accurate as it corresponds to around 30–40 ms in time.

The time shifts for all videos in the dataset were calculated relative to the videos taken from the feet camera—the reference camera. These time shifts were added to the dataset as an additional useful data to be used in applications requiring synchronised multi-view videos. The accurately synchronised data was further analysed and augmented for presentation in an original software application [12] for an enhanced dance visualisation experience.

8.3 Data release

Since May 2011, the dance corpus for the 3DLife ACM Multimedia Grand Challenge 2011 has been made publicly available through a website, 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.

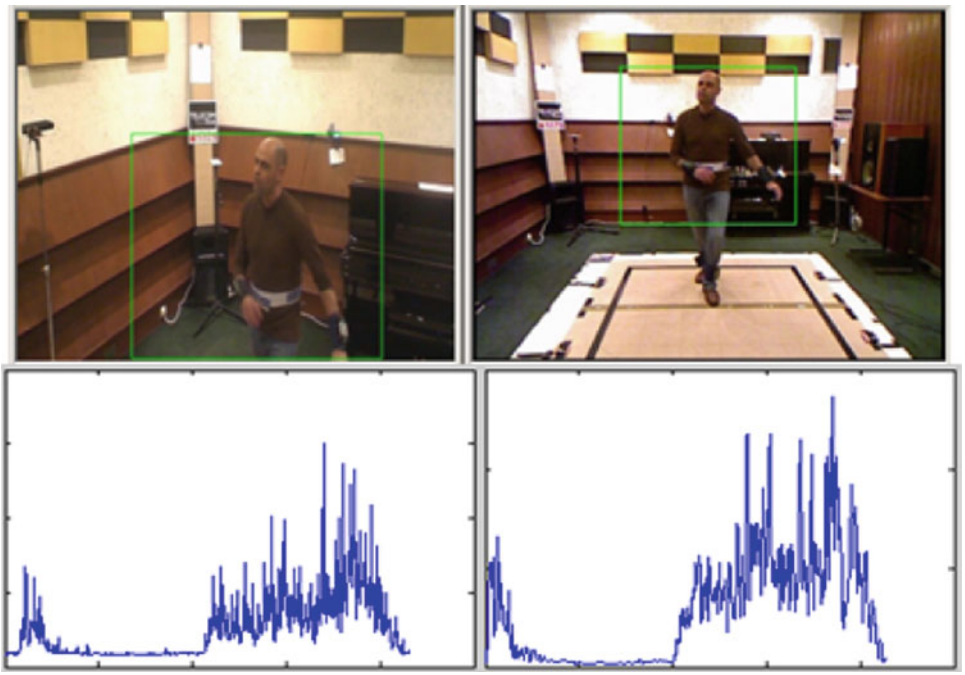


Fig. 8 *Top row* example for upped body detection; *bottom row* temporal features extracted from the two videos

9 Fields of application

The 3DLife dance dataset can be exploited by the scientific community to boost research in various emerging scientific fields. In the following, we refer to some interesting applications that could take advantage of our dataset.

This multi sensor dataset is an adequate testbed to evaluate algorithms relevant to human pose recovery. Human body pose recovery is a fundamental task in human motion analysis and in recent years there has been rapid progress regarding markerless 3D pose recovery of humans from multi-view video data. A recent framework for unconstrained 3D human upper body pose estimation from multiple camera views in complex environments is presented in [14].

Depth sensors provide an alternative technology for human pose recovery. In particular, Microsoft's Kinect is an affordable solution for boosting research and innovation in 3D human pose recovery methods. The work in [21] proposes a novel method for fast and accurate prediction of 3D positions of body joints from a single depth image generated by a single Kinect device. Another solution for handling the pose recovery problem is to use measurements from wearable inertial sensors attached to the person's limbs [20]. The work in [17] proposes a hybrid human pose tracker that combines video data with inertial units measurements to compensate for the drawbacks of each sensor type.

When human poses are estimated over time, the term human motion analysis is adopted. Advances in human motion analysis would benefit a wide spectrum of applications, especially in the domains of virtual reality, human-computer interaction, surveillance and elderly monitoring.

A comprehensive survey on visual view-invariant human motion analysis can be found in [15]. This work presents recent developments in view-invariant human motion analysis with an emphasis on view-invariant pose representation and estimation, and view-invariant action representation and recognition. While video-based human motion analysis has received much interest during the last decades, the launch of Kinect technology may provide a new insight in the human motion analysis field.

The work in [3] describes a novel system that automatically evaluates dance performances against a standard performance and provides visual feedback to the performer in a 3D virtual environment. In order to align and evaluate dance performances, Kinect depth-maps from the current dataset are considered to extract the motion of a performer via human skeleton tracking. The framework for the automatic evaluation of dancers performance is extended in [9] to include, except for Kinect depth-maps, audio and WIMU modalities. In [18] a classification system designed to recognize dancing gestures in real-time and with high accuracy from Kinect depth data is introduced.

The low cost of inertial sensors and their significant technological improvement in terms of size and power consumption provides an alternative option for analysing human motion. An extensive review of the literature related to the techniques for classifying human activities that are performed using inertial sensors is presented in [5]. This study pinpoints that inertial sensor technology can be exploited, among others, in remote monitoring of the elderly, rehabilitation and physical therapy, dance and ballet, sports science and virtual reality.

This multimodal dance corpus stands as a complex human activity database that is challenging for developing and testing human pose recovery and human motion analysis approaches. Moreover, the multimodal nature of the collected data allows the evaluation of methods that use diverse types of input data (i.e. visual input, depth data and inertial measurements), forming a convenient benchmark for comparing algorithms that either use input data from a single type of sensors or fuse data from different types of sensors for improved performance. Concluding, the appealing features of the presented corpus could constitute it as a roadmap for the construction of new databases with rich multi-source content.

10 Concluding remarks

In this work, we have presented a new multimodal corpus for research into, amongst other areas, real-time realistic interaction between humans in online virtual environments. Although the dataset is tailored specifically for an online dance class application scenario, the 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.

We believe that the provided corpus can be used to illustrate, develop and test a variety of tools in a diverse number of technical areas. For instance within our research teams, this dataset is currently being used to develop enhanced frameworks for the automatic analysis and evaluation of human activities (in particular dance performances) from multimodal recordings, including tasks such as (i) human motion analysis, (ii) gesture/dance movement recognition, or (iii) pose estimation using depth and inertial sensors in order to make tracking more robust to self occlusions and subtle joint orientation errors and trying to balance the demands between accuracy and speed in real time human–computer interaction applications.

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.

References

1. Clave (ryhthm) (2011). http://en.wikipedia.org/wiki/Clave_rhythm
2. Openni (2011). http://www.openni.org/
3. Alexiadis D, Kelly P, Daras P, O'Connor N, Boubekeur T, Moussa MB (2011) Evaluating a dancer's performance using kinect-based skeleton tracking. In: ACMR, pp 659–662
4. Alonso M, Richard G, David B (2005) Extracting note onsets from musical recordings. In: IEEE International Conference on Multimedia and Expo. IEEE Computer Society, Los Alamitos, USA. http://doi.ieeecomputersociety.org/10.1109/ICME.2005.1521568. ISBN 0-7803-9331-7
5. Altun K, Barshan B (2010) Human activity recognition using inertial/magnetic sensor units. In: HBU, pp 38–51
6. Blank M, Gorelick L, Shechtman E, Irani M, Basri R (2005) Actions as space-time shapes. In: ICCV, vol 3, pp 1395–1402.
7. Cannam C, Landone C, Sandler M (2010) Sonic visualiser: an open source application for viewing, analysing and annotating music audio files. In: Proceedings of the ACM multimedia 2010 international conference, Firenze, Italy, October 2010, pp 1467–1468.
8. Eichner M, Marin-Jimenez M, Zisserman A, Ferrari V (2012) 2D articulated human pose estimation and retrieval in (almost) unconstrained still images. Int J Comput Vis 99:190–214
9. Essid S, Alexiadis D, Tournemenne R, Gowing M, Kelly P, Monhagan D, Daras P, Dremeau A, O'Connor NE (2012) An advanced virtual dance performance evaluator. In: IEEE International Conference on Acoustics, Speech and Signal Processing, Kyoto, Japan
10. Essid S, Lin X, Gowing M, Kordelas G, Aksay A, Kelly P, Fillon T, Zhang Q, Dielmann A, Kitanovski V, Tournemenne R, O'Connor NE, Daras P (2011) Richard G (2011) A multimodal dance corpus for research into real-time interaction between humans in online virtual environments. In: ICMI workshop on multimodal corpora for machine learning, Alicante, Spain
11. Gkalelis N, Kim H, Hilton A, Nikolaidis N, Pitas I (2009) The i3dpost multi-view and 3d human action/interactions. In: CMVP, pp 159–168
12. Gowing M, Kell P, O'Connor N, Concolato C, Essid S, Lefeuvre J, Tournemenne R, Izquierdo E, Kitanovski V, Lin X, Zhang Q (2011) Enhanced visualisation of dance performance from automatically synchronised multimodal recordings. In: ACMR, pp 667–670
13. Gross R, Shik J (2001) The cmu motion of body (mobo) database. Technical report.
14. Hofmann M, Gavrila D (2012) Multi-view 3d human pose estimation in complex environment. IJCV 96(1):103–124
15. Ji X, Liu H (2010) Advances in view-invariant human motion analysis: a review. SMC Part C 40(1):13–24
16. Messing R, Pal C, Kautz H (2009) Activity recognition using the velocity histories of tracked keypoints. In: IEEE 12th International Conference on Computer Vision, Kyoto, Japan
17. Pons-Moll G, Baak A, Helten T, Mueller M, Seidel H, Rosenhahn B (2010) Multisensor-fusion for 3d full-body human motion capture. In: CVPR, pp 663–670
18. Raptis M, Kirovski D, Hoppe H (2011) Real-time classification of dance gestures from skeleton animation. In: ACM/SIGGRAPH SCA
19. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach. 17th International Conference on Pattern Recognition. Cambridge, UK. 3, pp 32–36
20. Schwarz L, Mateus D, Navab N (2012) Recognizing multiple human activities and tracking full-body pose in unconstrained environments. Pattern Recognit 45(1):11–23

21. Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, Kipman A, Blake A (2011) Real-time human pose recognition in parts from single depth images. In: CVPR
22. Sigal L, Balan AO, Black MJ (2010) Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. Int J Comput Vis 87(1):4–27
23. Singh S, Velastin S, Ragheb H (2010) Muhavi: a multicamera human action video dataset for the evaluation of action recognition methods. In: AVSS, pp 48–55
24. Wang Y, Huang K, Tan T (2007) Human activity recognition based on r transform. In: CVPR, pp 1–8
25. Ushizaki KDM, Okatani T (2006) Video synchronization based on co-occurrence of appearance changes in video sequence. In: ICPR
26. Weinland D, Ronfard R, Boyer E (2006) Free viewpoint action recognition using motion history volumes. CVIU 104(2–3):249–257
27. Yang A, Jarafi R, Kuryloski P, Iyengar S, Sastry S, Bajcsy R (2008) Distributed segmentation and classification of human actions using a wearable motion sensor network. In: CVPRW, pp 1–8