TPT-Dance&Actions : un corpus multimodal d'activités humaines

Aymeric Masurelle, Ahmed Rida Sekkat, Slim Essid, Gaël Richard

LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France
{aymeric.masurelle,ahmed.sekkat,slim.essid,gael.richard}@telecom-paristech.fr

RÉSUMÉ. *Nous présentons une nouvelle base de données multimodales d'activités humaines, TPT - Dance & Actions, s'adressant, parmi d'autres, aux domaines de recherche liés à l'analyse de scènes multimodales. Ce corpus se focalise sur des scènes de danse (lindy hop, salsa et danse classique), de fitness rythmées par de la musique, et inclut également des enregistrements d'actions isolées plus "classiques". Il regroupe* 14 *chorégraphies de danse et* 13 *séquences d'autres activités multimodales effectuées en partie par* 20 *danseurs et* 16 *participants respectivement. Ces différentes scènes multimodales sont enregistrées à travers une variété de réseaux de capteurs : caméras, capteurs de profondeur, microphones, capteurs piézoélectriques et une combinaison de capteurs inertiels (accéléromètres, gyroscopes et magnétomètres). Ces données seront disponibles sur internet librement à des fins de recherche.*

ABSTRACT. *We present a new multimodal database of human activities, TPT - Dance & Actions, for research in multimodal scene analysis and understanding. This corpus focuses on dance scenes (lindy hop, salsa and classical dance), fitness and isolated sequences.* 20 *dancers and* 16 *participants were recorded performing respectively* 14 *dance choreographies and* 13 *sequences of other human activities. These different multimodal scenes have been captured through a variety of media modalities, including video cameras, depth sensors, microphones, piezoelectric transducers and wearable inertial devices (accelerometers, gyroscopes et magnetometers). These data will be available to download for research purpose.*

MOTS-CLÉS : *danse, lindy hop, salsa, danse classique, fitness, actions isolées, données multimodales, audio, vidéos, cartes de profondeur, données inertielles, synchronisation, analyse d'activités multimodales, reconnaissance de gestes et d'actions.*

KEYWORDS: *dance, lindy hop, salsa, classical dance, fitness, isolated actions, multimodal data, audio, video, depthmaps, inertial data, synchronisation, multimodal activity analysis, gestures and actions recognition.*

DOI:10.3166/TS.32.443-475

Extended abstract

We present, in this article, a new multimodal database of human activities, *TPT - Dance & Actions*, for research in multimodal scene analysis and understanding, with various applications,including human-machine interaction and human-human interaction in virtual environment.
Our corpus is composed of recordings of multimodal scenes where a participant executes sequences of body movements following a specific music and/or instructions. We consider dance movements of particular interest as gestures are in this case driven by the corresponding music which should be taken into account during the analysis. Thus we decide to focus this corpus mainly on dance scenes, namely choreographies of lindy hop, salsa and classical dance accompanied by some excerpts of their music genre. The dataset also includes recordings of fitness sequences cadenced by rhythmic music and of "classical" isolated actions to enhance the gesture diversity of this corpus. The sequences of fitness motions and isolated actions have been arranged by ourselves. However, for each dance style, an expert has composed the different choreographies and chosen the respective musical accompaniments. We ask these different experts to arrange short and simple sequences of dance steps in which some steps appear in several choreographies. It allows dancers of differing expertise to perform most of the sequences. Also, we obtain different transitions and velocities for a set of dance steps. To record the different performances, we have built a multimodal capture system : audio rigs, multiple cameras, depth sensors, piezoelectric tranducers on floor and wearable inertial measurement devices. With this setup, 20 dancers and 16 participants were recorded performing respectively 14 dance choreographies and 13 sequences of others human activities.
Thus the variety of data and gesture types gives an interesting baseline for movement analysis. Moreover the correlation between music and dance motion offers new challenges and original research perspectives for multimodal scene analysis.
To the best of the authors' knownledge, there is only one multimodal database providing dance recordings restricted to salsa dance, the 3*DLife Dance dataset*. Our multimodal corpus is quite complete and original in terms of the different kinds of sensors used, of the quality of the recordings and of the recorded gesture diversity.

1. Introduction

Nous présentons, dans cet article, une nouvelle base de données multimodales de danse, de fitness et d'actions, *TPT - Dance & Actions*, s'adressant aux domaines de recherche liés à l'analyse de scènes multimodales et notamment à l'interaction réaliste et en temps-réel entre humains dans des environnements virtuels.

Nous avons construit ce corpus de mouvements du corps humain accompagnés ou non de musique en s'inspirant de deux cas d'usage. Le premier cas d'usage prévoit de permettre à une classe d'élèves accompagnée de leur professeur de visiter virtuellement un monument ou une institution guidée par un agent autonome (*i.e.* entité virtuelle autonome). Dans ce monde virtuel, les utilisateurs sont représentés

par des agents semi-autonomes. Lors de cette visite éducative, le professeur et les élèves doivent pouvoir contrôler leur agent semi-autonome afin de pouvoir interagir ensemble sur des sujets proposés. Ainsi les systèmes d'analyse de gestes proposés doivent pouvoir interpréter des gestes de communications non verbales et des actions simples afin de permettre un échange réaliste.

Le deuxième cas d'usage envisage que les utilisateurs puissent interagir naturellement dans une pièce virtuelle afin de réaliser des activités communes à distance. Par exemple, des participants peuvent prendre part ensemble à une classe de danse virtuelle. Dans ce cas, les participants effectuent des gestes bien plus complexes où toutes les parties de leur corps peuvent être impliquées. De plus, dans des activités comme la danse, la musique joue un rôle structurant important dans les choix de l'interprétation gestuelle d'un danseur. D'ailleurs chaque style de danse est associé à un genre particulier de musique. Alors, les systèmes développés doivent pouvoir effectuer conjointement une analyse précise des mouvements des participants et de l'environnement sonore de la scène multimodale afin de permettre une interaction naturelle entre eux lors d'activités telles que la danse.

Pour cela nous prévoyons que les utilisateurs pourraient avoir différents types de systèmes de captures de données tels que : une caméra RVB face à l'utilisateur, une caméra de profondeur face à l'utilisateur, un réseau de caméra RVB ou de Kinects ou un système de capture de mouvements. De plus pour une activité telle que la danse où l'analyse du mouvement doit être particulièrement précise, l'utilisateur doit pouvoir améliorer son installation d'acquisition de données en portant des capteurs inertiels si nécessaire. Malgré les différents types d'installation, nos systèmes d'analyse doivent permettre aux utilisateurs d'interagir entre eux lors de discussions et d'activités multimodales.

Les différents types d'installations de capture et d'analyse du geste, évoqués précédemment, proposent un environnement adéquat pour l'analyse du mouvement du fait de sa diversité. De plus la corrélation entre musique et mouvement dansé, additionnée à la complexité gestuelle inhérente de ces derniers offrent de nouveaux challenges et des perspectives originales de recherche dans le domaine de l'analyse de scènes multimodales. Nous avons ainsi concentré notre travail sur l'analyse multimodale de scènes de danse.

A notre connaissance, il n'existe qu'une seule base de données multimodales de danse, il s'agit du corpus 3*DLife Dance dataset* (Essid, Lin *et al.*, 2012) se focalisant sur l'analyse de scènes de salsa. Nous avons participé à sa construction et nous l'avons utilisée pour évaluer certains de nos algorithmes. Ainsi nous avons pu constater un manque de diversité dans les données et des problèmes de qualité des enregistrements, que nous détaillons en section 2.2. Riches de cette première expérience et dans une volonté d'explorer de nouveaux champs de recherche, nous avons décidé de créer un corpus original essentiellement constitué de scènes multimodales : des scènes de danse caractéristiques de la salsa, du lindy hop et de la danse classique ; des scènes de fitness pour proposer d'autres types de scènes multimodales ; et des séquences d'actions isolés afin d'enrichir la diversité des gestes de ce corpus. Ce corpus est donc

composé de gestes de complexité variable enregistrés avec plusieurs types de capteurs incluant un réseau de caméras RVB, un réseau de capteurs de profondeur, un réseau de microphones et de capteurs piézoélectriques sur le sol ainsi qu'une combinaison munie de capteurs inertiels.

Ainsi nous avons créé une base de données multimodales assez complète aussi bien au niveau des types de capteurs utilisés qu'au niveau de la diversité des gestes enregistrés. Cependant, des problèmes de fiabilité de l'équipement nous obligent à revoir les méthodes de synchronisation entre les différents media. Ceci retarde la mise en disponibilité de ce corpus et nous empêche de l'exploiter pour l'instant. Une fois ces problèmes résolus, ce corpus pourra potentiellement intéresser différents domaines de recherche tels que :

– l'acquisition et le traitement de données 3D à partir de réseaux de capteurs,

– le rendu réaliste de données 3D,

– la synchronisation multimodale avec différents types capteurs,

– l'analyse multimodale d'actions simples comme la reconnaissance ou la détection de gestes,

– l'analyse multimodale de danse comme le suivi automatique de chorégraphies, l'analyse des similarités entre musique et danse et l'évaluation de performances dansées.

Ceci permet l'évaluation d'une grande diversité d'algorithmes utilisant un ou plusieurs flux de données et ainsi d'avoir un regard homogène et critique sur leur efficacité en comparant leurs performances.

Dans cet article, des travaux connexes à notre corpus sont présentés (*cf.* section 2). Puis une vue d'ensemble de notre corpus est proposée (*cf.* section 3). Ensuite le protocole d'enregistrement (*cf.* section 4), les musiques, les séquences de gestes et les indications sonores (*cf.* section 5), le matériel enregistrement (*cf.* section 6) sont détaillés. A la suite, les annotations des séquences (*cf.* section 7), la préparation des données et leur mise en disponibilité (*cf.* section 8) sont décrites. Enfin, divers champs d'application sont proposés (*cf.* section 9) ce qui nous permettra de conclure (*cf.* section 10).

2. Travaux connexes

Dans cette section nous présentons une sélection de bases de données reliées aux domaines de recherche de nos travaux. Nous articulerons cette présentation en deux points :

– les bases de données monomodales, utilisant un seul type de capteurs ;

– les bases de données multimodales, utilisant plusieurs type de capteurs.

Deux tableaux récapitulatifs des bases de données monomodales et multimodales existantes, reliées aux domaines de la reconnaissance de geste, sont présentés dans l'Annexe A.

2.1. Bases de données monomodales

Au sein des bases monomodales nous pouvons différencier deux groupes :

– les bases de données uni-capteur, utilisant un seul capteur, et,

– les bases de données multi-capteurs, utilisant un réseau de capteurs.

2.1.1. Bases de données uni-capteur

La reconnaissance d'actions humaines est un sujet de recherche populaire particulièrement au sein de la communauté de la vision par ordinateur. D'ailleurs, les deux bases de données certainement les plus utilisées à ce jour sont assez similaires et enregistrées à l'aide d'une seule caméra vidéo.

La base de données *KTH Human Action Dataset* [1] (Schüldt *et al.*, 2004) est une base constituée de données issues d'une caméra vidéo noir et blanc. Elle comprend 6 actions simples (marcher, trottiner, courir, boxer, frapper des mains, agiter les bras) effectuées par 25 personnes, sans mouvement de caméra dans quatre scénari différents (en extérieur, en extérieur avec changement d'échelle (zoom), en extérieur avec d'autres vêtements, en intérieur). L'autre corpus est un peu plus récent, il s'agit de la base de données *Weizmann human action dataset* [2] (Gorelick *et al.*, 2007). Elle est composée de 10 actions simples (marcher, courir, sauter, sauter sur place, se pencher vers le sol, avancer à cloche pied, agiter un bras, agiter les deux bras, pas chassés, jumping jacks [3]) réalisées par 9 personnes dans un seul scénario (en extérieur) et sans mouvement de caméra.

D'autres bases de données d'actions utilisant le même type de modalité ont été constituées à partir d'extraits vidéos issues de films hollywoodiens (Laptev *et al.*, 2008), d'émissions sportives de télévision (Rodriguez *et al.*, 2008) ou d'un site internet d'hébergement de vidéos (Liu *et al.*, 2009).

Cependant ces bases de données sont plutôt destinées aux domaines de l'indexation vidéo et de la vidéo surveillance et assez peu à celui de l'interaction homme-machine. Ainsi nous nous sommes aussi intéressés à des bases de données dédiées à l'interaction homme-machine comme la base de données de langage des signes : *Australian Sign Language (v.*1 *et v.*2*)* [4] (Kadous, 2002). Dans cette base de données, 95 signes différents sont exécutés plusieurs fois par 6 personnes différentes (5 personnes dans la première version (v.1) et une de plus dans la seconde (v.2)). Dans cette expérience, la position de la main dans l'espace est enregistrée à l'aide d'un magnétomètre bas-coût pour la v.1 et d'un magnétomètre de meilleure qualité pour la v.2.

---

1. www.nada.kth.se/cvap/actions
2. www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html
3. un saut partant de la position debout avec les jambes collées et les bras le long du corps arrivant à une position où les jambes sont écartées et les bras au dessus de la tête
4. www.kdd.ics.uci.edu/databases/auslan/auslan.html

2.1.2. Bases de données multi-capteurs

Le corpus multi-vue *INRIA XMAS dataset* [5] (Weinland *et al.*, 2006) contient 11 actions réalisées 3 fois par 10 participants différents avec un arrière-plan et une luminosité fixes. L'acquisition de ces données a été effectuée en utilisant 5 caméras vidéos synchronisées. En plus des données brutes, les fichiers de calibration, les silhouettes de chaque participant ainsi que leurs models 3D associés sont disponibles. D'ailleurs plusieurs bases de données ont été construites de manière assez similaire.

La base *i3DPost Multi-view Human Action Datasets* [6] (Gkalelis *et al.*, 2009) est composée de données issues de 8 caméras statiques haute-résolution représentant 12 actions effectuées par 8 personnes sur un fond bleu et une luminosité constante. Un maillage 3D représentant la surface 3D du corps de chaque participant est également fourni à chaque trame. Le corpus *MuHAVi: Multicamera Human Action Video Data* [7] (Singh *et al.*, 2010) inclut 17 actions réalisées plusieurs fois par 14 participants et enregistrées par 8 caméras statiques.

Néanmoins d'autres bases de données utilisant d'autres types de réseaux de capteurs ont été construites. Par exemple, la base de données *WARD: Wearable Action Recognition Database* [8] (Yang *et al.*, 2009) a été créée en utilisant 5 capteurs inertiels placés sur les poignées, les chevilles et à la taille des participants. Ainsi 20 personnes ont été enregistrées séparément en train d'exécuter 13 actions simples plusieurs fois.

Le grand nombre de bases de données multi-capteurs utilisant des réseaux caméras vidéos reflète le fort intérêt de la communauté de la vision par ordinateur pour l'analyse des mouvements du corps humain.

2.2. Bases de données multimodales

La base de données *HumanEva Database* [9] (Sigal *et al.*, 2010) est constituée de deux parties: *HumanEva-I* et *HumanEva-II*. *HumanEva-I* est composée de données multimodales représentant 6 actions répétées 3 fois par 4 personnes. Ces données sont enregistrées de manière synchronisée en utilisant 7 caméras vidéo ainsi qu'un système de capture de mouvements. *HumanEva-II* est une extension de *HumanEva-I* où l'installation de capture a été améliorée mais où seulement 2 participants effectuent les séquences d'actions.

A notre connaissance, il n'y a que peu de bases de données enregistrées à travers de diverses et multiples modalités afin d'obtenir simultanément des données vidéo, audio, de cartes de profondeur et des données issues de capteurs inertiels, et qui se focalisent sur des activités multimodales aussi complètes et complexes que des scènes de danse.

---

5. www.4drepository.inrialpes.fr/public/viewgroup/6
6. www.kahlan.eps.surrey.ac.uk/i3dpost_action
7. www.dipersec.king.ac.uk/MuHAVi-MAS
8. www.eecs.berkeley.edu/~yang/software/WAR
9. www.vision.cs.brown.edu/humaneva

Une base de données de scènes multimodales de cuisine a été créée : *CMU-MMAC Database* [10] (De la Torre *et al*., 2009). Ce corpus d'activités multimodales unique en son genre comporte les enregistrements de 43 personnes cuisinant 5 recettes différentes. Afin de varier les scenari, un petit ensemble de données contenant des situations anormales (apparition de fumée, coup de téléphone, voleur dans la maison, ...) est aussi fourni dans ce corpus. 6 caméras RVB (5 fixes et 1 embarquée), 5 microphones, un système de capture de mouvement professionnel et 5 capteurs inertiels portés par le participant permettent d'enregistrer au mieux les différentes sources d'information de ce type de scènes multimodales. Ainsi ce corpus constitue un support complet et original pour l'analyse multimodale de scènes de cuisine. Néanmoins les types de gestes effectués au sein de cette base sont principalement des mouvements du haut du corps et en particulier des bras. Aussi, les sons présents dans ces scènes de cuisine sont des sons essentiellement produits par les outils utilisés, la cuisson des aliments et les mouvements du participant. Ils peuvent au mieux aider le participant à distinguer le début et/ou la fin de certaines étapes de préparation ou de cuisson. Les modalités audios des scènes de cuisine n'influencent que peu les mouvements du participant à l'inverse des scènes de danse.

La base de données 3*DLife Actions dataset* [11] (Gowing *et al*., 2014) est une base multimodale regroupant des flux de données synchronisées représentant 22 actions diverses exécutées 5 fois par 17 personnes. Ces données sont enregistrées à l'aide de 5 Kinects (un flux vidéo, un flux de cartes de profondeur et 4 flux audio pour chaque Kinect) et de 5 capteurs inertiels portés par le participant pour la majorité des prises. De plus la Kinect face au participant a été disposée de manière horizontale (contrairement aux autres) permettant une estimation de la position des articulations du participant. Cette base de données a été au moins partiellement utilisée pour nos travaux de reconnaissance automatique d'actions (Masurelle *et al*., 2014). Cela nous a permis d'apprécier l'effort qui a été fourni au niveau de la synchronisation et de la diversité des données (RVB, cartes de profondeur, skeleton [12], audio et inertiel) même si la qualité des flux vidéo et des squelettes estimés est basse. De plus un nombre important de participants a été réuni pour la création de ce corpus. Aussi, plusieurs types d'activités impliquant des mouvements de tout le corps sont représentés: actions simples (faire coucou, frapper des mains, ...), activités sportives (coup de pied, revers de tennis, ...) et des postures statiques (main sur les hanches, bras croisés, ...). Cette base de données multimodales forme un corpus complet dédié essentiellement à la reconnaissance d'actions isolées. En effet, dans les scènes présentées, il n'y a aucune activité de nature réellement multimodale. Aucune modalité n'a d'influence sur l'interprétation des gestes des participants, contrairement aux scènes de danse.

A priori, le corpus 3*DLife Dance dataset* [13] (Essid, Lin *et al*., 2012) est la seule base de données multimodales de danse. Elle est composée d'enregistrements mul-

10. www.kitchen.cs.cmu.edu
11. www.mmv.eecs.qmul.ac.uk/mmgc2013
12. Squelette simplifié obtenu suite à l'estimation de la position des articulations dans l'espace.
13. www.perso.telecom-paristech.fr/~essid/3dlife-gc-11

timodaux de 15 danseurs de niveaux différents effectuant séparément la plupart des 5 chorégraphies de salsa proposées. Les performances des danseurs sont enregistrées via 9 caméras vidéo couvrant la totalité du corps du danseur, 2 Kinects placées en face et sur le côté du danseur afin d'obtenir des flux vidéo et des données de profondeur, 5 capteurs inertiels portés par le danseur, 12 microphones pour capter la musique d'accompagnement et les sons produits par le danseur et 4 capteurs piézoélectriques placés sur la piste de danse afin de capter les sons d'impacts de pas au sol. Nous portons un regard assez critique sur ce corpus car nous avons participé à sa construction et nous nous en sommes aussi servis pour évaluer un de nos systèmes de reconnaissance de pas de salsa (Masurelle *et al.*, 2013). La plupart des enregistrements vidéos sont d'une qualité trop faible pour être exploités. Tous les flux des Kinects n'ont pas pu être enregistrés: aucun flux de cartes de profondeur de la Kinect 2 pour éviter les interférences avec la Kinect 1, ainsi qu'aucun flux audio provenant des microphones internes des deux Kinects car les pilotes, utilisés à ce moment-là, ne le permettaient pas. Les chorégraphies proposées ont été écrites par un danseur de salsa aguerri en s'efforçant d'avoir une structure simple et courte. Ceci a permis un apprentissage rapide des chorégraphies pour la plupart des danseurs quelque soit leur niveau. Il arrive qu'un danseur débutant ne commence pas la chorégraphie à l'instant prévu. Dans ce cas, la musique d'accompagnement ne peut plus complètement jouer son rôle dans la scène observée. De plus, en salsa, les danseuses et les danseurs effectuent des mouvements en miroir ou complémentaires au cours d'une chorégraphie. Même si dans ce cas les danseuses et les danseurs sont seuls à exécuter les chorégraphies, cette particularité a été gardée. Ce choix engendre une grande diversité de mouvements au sein d'une même chorégraphie. Cela peut être problématique pour mener à bien un processus d'apprentissage automatique.

3. Présentation du corpus

Nous avons vu précédemment que les bases de données de gestes existantes sont assez peu variées. En effet, peu d'entre elles sont constituées de données multimodales. La plupart de ces bases sont enregistrées à l'aide d'une seule caméra RVB ou d'un réseau de caméras RVB. Par conséquent un nombre encore plus réduit d'entre elles représente des scènes d'activités multimodales.

Malgré la difficulté que représente la construction d'un tel corpus, nous avons décidé de créer une base de données d'activités multimodales en nous inspirant des corpus 3*DLife Actions dataset* et 3*DLife Dance dataset*, décrits dans la section 2.2. Notre corpus est alors constitué d'une certaine diversité de scènes multimodales détaillées dans le tableau 1. Une minute d'enregistrement regroupant les différents media représente un espace de stockage d'environ 11Go. Afin d'enregistrer ces différentes scènes d'activités multimodales, nous avons conçu un système d'enregistrements multimodaux plus complet et plus performant, illustré à la figure 1. Un récapitulatif des différents types de données que nous avons acquis avec notre système multimodal d'enregistrement est présenté au tableau 2. Ces données multimodales sont enregistrées en utilisant plusieurs réseaux de capteurs dont certains sont synchronisés entre

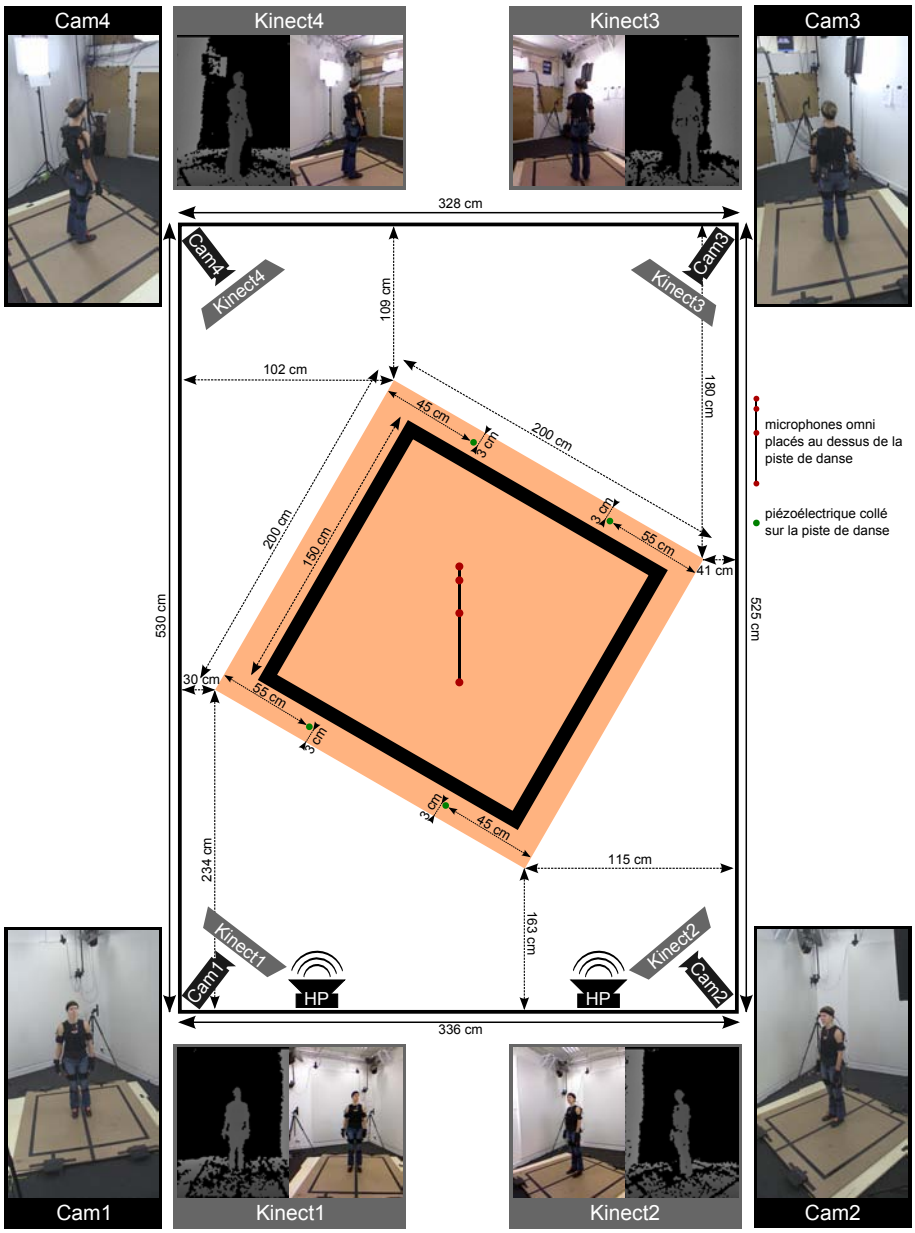

Figure 1. Vue d'ensemble du système multimodal d'enregistrements des données de notre corpus TPT-Dance&Actions

Tableau 1. Tableau récapitulant les trois types de scènes multimodales et cinq parties différentes constituant notre corpus d'activités multimodales ainsi que le nombre de personnes ayant participées à chacune des parties

Scènes de danse	8 chorégraphies de lindy hop	7 danseurs
	3 chorégraphies de salsa	11 danseurs
	3 chorégraphies de danse classique	4 danseuses
Scènes de fitness	8 chorégraphies de fitness	16 participants
Séquences d'actions isolées	5 séquences composées d'actions simples, d'activités sportives rythmiques et de postures.	

Tableau 2. Tableau récapitulatif des différentes données enregistrées avec notre système de capture

Type de données	*Descriptions*
Audio (28 canaux)	4 microphones sont placés au-dessus de la tête des participants.
	Un des microphones de chacune des 4 caméras Canon est utilisé.
	Les 4 microphones internes sont dans chacune des 4 Kinects.
	4 piézoélectriques sont collés sur la piste de danse.
RVB HD	4 caméras Canon entourent le corps du participant.
RVB	4 Kinects sont placées autour du participant.
Cartes de profondeur	
Inertiels	La combinaison Xsens MVN Biomech de 17 capteurs inertiels est portée par le participant.
Skeleton	Le système Xsens MVN Biomech estime la position dans l'espace de 22 articulations.
Calibration	Des données pour calibrer les caméras RVB sont disponibles.

eux. A la fin de chaque prise un événement distinctif est présent dans chacun des flux ce qui permet aux modalités non-synchronisées de pouvoir être post-synchronisées avec les autres flux de données. Des détails sur la synchronisation de ces données peuvent être trouvés à la section 8.2.

Notre corpus regroupe à la fois des actions isolées et des séquences continues de mouvements complexes du corps accompagnées de musique, ce qui justifie pleinement l'utilisation d'un système de capture multimodale aussi complexe. Ceci fait de cette base un corpus complet et unique en son genre.

4. Protocole d'enregistrement

Chaque participant suivant ses compétences choisit un ou plusieurs types de séquences de gestes. Pour cela le participant peut s'aider de vidéos d'exemples des différentes séquences afin d'évaluer le niveau de difficulté des séquences proposées.

Pour chaque participant, une session d'enregistrement commence par une phase

de préparation. Dans un premier temps la personne est équipée d'une combinaison de capteurs inertiels afin qu'elle puisse s'habituer à son contact. Puis les instructions concernant le déroulement des enregistrements ainsi que les séquences de gestes lui sont expliquées et montrées.

Ensuite, nous mettons à la disposition du participant des vidéos d'instructions où des experts effectuent les différentes séquences de gestes. Le participant peut répéter dans de bonnes conditions les séquences de gestes proposés jusqu'à ce qu'il se sente prêt à être enregistré. Seules les séquences de gestes que le participant réussit à maîtriser après un temps raisonnable (5 à 30 minutes) sont enregistrées. Pour chaque chorégraphie un certain nombre de prises réussies est enregistré. Néanmoins le nombre de prises peut varier suivant la disponibilité de chaque participant. Ainsi notre but est d'obtenir, pour chaque séquence de gestes exécutés, au moins deux prises où le participant arrive à exécuter la séquence correctement lorsque ceci est possible. Afin d'aider les participants à effectuer correctement les séquences de gestes, les vidéos d'instructions des différentes séquences sont diffusées lors des différentes prises d'enregistrement.

Régulièrement, les sessions d'enregistrements commencent par la prise d'images de calibration afin de pouvoir assurer que la calibration du réseaux de caméras reste correcte au cours du temps (Zhang, 2000). La calibration est effectuée en utilisant un échiquier de calibration de 5x4 carrés où les côtés des carrés font 15 cm. Cette taille des carrés a été choisie assez grande pour que les motifs de l'échiquier soient représentés clairement dans tous les flux vidéos des différentes caméras. Cet échiquier a été placé devant chaque caméra et aussi sur la piste de danse (*cf.* figure 2).

Lors des enregistrements, nous nous sommes confrontés à des problèmes de fiabilité de l'équipement utilisé. Ainsi, l'ensemble de toutes les données n'a pu être enregistré de manière parfaitement synchronisée. Afin de minimiser ce désagrément, nous avons mis en place des techniques de post-synchronisation et demandé à tous les participants d'exécuter une procédure de clap à la fin de chacune de leur performance. Cette procédure de clap consiste à successivement frapper des mains puis à frapper

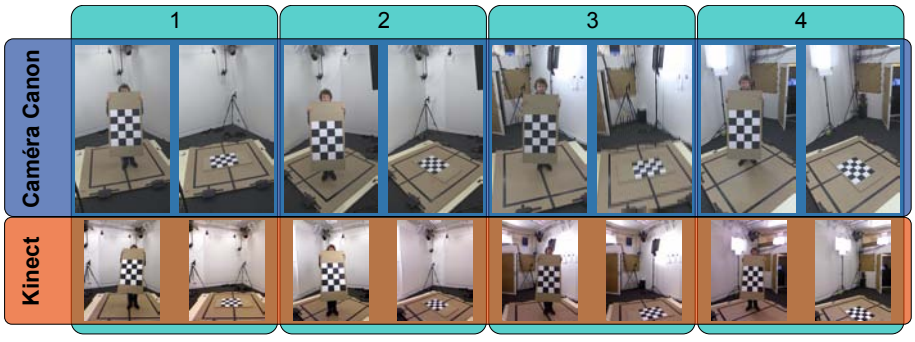

Figure 2. Images extraites des données de calibration où un échiquier a été placé devant chaque caméra (Canon et Kinect) et sur la piste de danse

Figure 3. Frise photographique illustrant la procédure de clap effectuée par chaque participant à la fin de chaque prise

le sol avec chaque pied, *cf.* figure 3. Ainsi tous les flux de données peuvent être synchronisées (soit manuellement ou soit automatiquement) en alignant cette signature clairement perçue à la fin de chaque flux de données.

5. Musiques, séquences de gestes et indications

Comme expliqué précédemment, nous nous sommes essentiellement focalisés sur l'enregistrement de scènes multimodales et en particulier des scènes de danse : salsa, lindy hop et danse classique.

Les séquences de mouvements de fitness et d'actions isolées ont été composées par nos soins. Par contre les choix des musiques et des chorégraphies ont été effectués par un expert de la danse interprétée. En conséquence, le choix des pas de danse, leur emplacement sur la partition et leur vitesse d'exécution respectent la structure et le phrasé musical. Aussi, nous nous assurons que chaque musique d'accompagnement est en accord avec le style d'expression corporelle concernée. Pour enregistrer des danseurs aussi bien débutants que confirmés, nous avons demandé aux experts en danse d'écrire de courtes chorégraphies avec une structure simple et comportant des redondances de certains pas afin qu'ils apparaissent dans plusieurs chorégraphies. Ceci permet d'obtenir différentes transitions et vitesses d'exécution pour un jeu de pas de danse. Pour accélérer l'apprentissage, nous avons produit des vidéos d'instructions pour chacune des séquences avec les différents experts. Dans ces vidéos nous voyons chaque expert de face exécutant les différentes chorégraphies dans des conditions d'enregistrement similaires. Nous avons fait de même pour les séquences de fitness et d'actions isolées. Nous avons rajouté à ces vidéos des indications vocales afin de permettre au candidat de mieux anticiper les mouvements et ainsi l'aider à effectuer correctement les différentes séquences. D'ailleurs, lors des enregistrements ces vidéos d'instructions sont diffusées aux participants.

La salsa et le lindy hop s'appuyent sur des temps forts au sein de structures musicales de huit temps pour intégrer leurs pas de danse. Ce sont aussi des danses de couples. Les mouvements du haut du corps assurent essentiellement la liaison entre les partenaires même s'ils font partie des pas de danse. Les pas de salsa et de lindy hop reposent essentiellement sur une combinaison de mouvements de jambes entraînant le reste du corps, tout en respectant la structure rythmique de la musique. Par

contre la danse classique n'a pas de contraintes au niveau des emplacements des pas par rapport à la structure musicale et les temps ne sont pas autant marqués dans la gestuelle. Les chorégraphies de danse classique suivent plutôt le phrasé musical. Aussi, ce style de danse ne se danse pas forcément en couple ce qui permet d'effectuer des pas composés de mouvements complexes et amples mobilisant l'ensemble du corps d'un danseur. Ainsi ce corpus original de danse offre un support riche pour l'analyse multimodale de scènes. Ceci permet de nouvelles perspectives de recherche sur l'analyse conjointe de séquences continues de mouvements et de leur musique d'accompagnement.

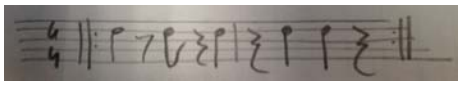

Figure 4. Motif rythmique de clave utilisé dans la chorégraphie de salsa css1

5.1. Musiques d'accompagnement

La plupart des musiques utilisées proviennent de l'ensemble de production Creative Commons, ce qui nous autorise à les partager publiquement. Pour les musiques des chorégraphies de lindy hop, le groupe de musique Jackass Brass Band [14] nous a fourni et autorisé à utiliser des extraits de certains de leurs enregistrements pour ce corpus. L'ensemble de nos 22 extraits musicaux sont tirés de 16 pistes différentes : 8 pistes de musique Swing, 2 pistes de musique salsa, un motif rythmique de clave (*cf.* figure 4), 3 pistes de musique adéquate à la danse classique et 2 pistes de musique électronique dédiées au fitness. Pour chaque type de scènes multimodales, les musiques d'accompagnement ont été aussi sélectionnées afin que différents tempi soient représentés (*cf.* tableaux 3, 4 et 5), cela enrichit notre corpus d'une variété de vitesse d'exécution de gestes. Tous les extraits audios et/ou musicaux utilisés sont disponibles dans la base de données en stéréo à une fréquence d'échantillonnage de 48 kHz.

5.2. Séquences de gestes

Chaque participant suivant ses compétences choisit un ou plusieurs types de séquences de gestes. Lorsqu'il a une habileté particulière dans un des styles de danse proposé, la personne exécute l'ensemble des chorégraphies prédéfinies de ce style si le temps imparti le permet. Sinon les participants n'ayant pas de compétences particulières dans les styles de danse proposés effectuent l'ensemble des séquences d'actions simples et de fitness.

14. www.jackassbrassband.fr

5.2.1. Chorégraphies de danse

5.2.1.1. Lindy hop

Dans un effort de simplicité, les chorégraphies de lindy hop ont la structure suivante : *Pas-de-base* / *Pas-spécifique-A* / *Pas-spécifique-B* / *Pas-de-base*. Afin d'augmenter les variétés de transitions entre les différents pas de danse, nous avons construit des chorégraphies par couple : chorégraphie paire - *Pas-de-base* / *Pas-spécifique-A* / *Pas-spécifique-B* / *Pas-de-base* et chorégraphie impaire - *Pas-de-base* / *Pas-spécifique-B* / *Pas-spécifique-A* / *Pas-de-base*. Ainsi les pas spécifiques A et B sont toujours précédés et suivis par tous les pas présents dans chaque couple de chorégraphies.

De plus pour chaque couple de chorégraphies, nous avons fait en sorte que la différence de tempo entre leur musique d'accompagnement soit maximale, *cf.* tableau 3. Ceci permet d'obtenir un fort contraste de vitesses d'exécution pour chaque pas de lindy hop présent dans ce corpus. Aussi, pour conserver le même minutage à travers tous les danseurs de lindy hop, un compte à rebours vocal indiquant le début de chaque chorégraphie a été rajouté par dessus la musique. Un ensemble de 7 danseurs de niveaux allant de l'amateur à l'expert ont été enregistrés sur ces différentes chorégraphies de lindy hop.

Tableau 3. Listes des différentes chorégraphies de lindy hop de notre corpus TPT-Dance&Actions (bpm : battements par minute)

Nom des chorégraphies	Séquence de pas de lindy hop	durée [s]	Tempo [bpm]
clh1	*pas de base Charleston kick, pas Charleston pose, pas Charleston slide, pas de base Charleston kick.*	12	175
clh2	*pas de base Charleston kick, pas Charleston slide, pas Charleston pose, pas de base Charleston kick.*	19	115
clh3	*pas de base Charleston kick, pas Charleston triple step, pas johnny's drop, pas de base Charleston kick.*	13	170
clh4	*pas de base Charleston kick, pas johnny's drop, pas Charleston triple step, pas de base Charleston kick.*	21	110
clh5	*pas de base Charleston kick, pas bascule, pas triple demi-boucle, pas de base Charleston kick.*	13	160
clh6	*pas de base Charleston kick, pas triple demi-boucle, pas bascule, pas de base Charleston kick.*	21	100
clh7	*pas de base Charleston kick, pas triple step, pas circle, pas de base Charleston kick.*	14	150
clh8	*pas de base Charleston kick, pas circle, pas triple step, pas de base Charleston kick.*	24	90

5.2.1.2. Salsa

Pour la salsa, nous avons réutilisé 3 des 5 séquences chorégraphiques de la base de données 3*D-Life dataset*. Afin de limiter la diversité des pas de salsa, nous avons

Figure 5. Frise photographique illustrant le pas Johnny's Drop présent dans les chorégraphies de lindy hop clh3 et clh4

demandé à tous les participants, femme et homme, d'effectuer les mouvements des pas des hommes. Toujours dans un souci de simplicité, les trois chorégraphies choisies ont été construites en utilisant la structure suivante: *Pas-de-base* / *Pas-spécifique-A* / *Pas-de-base* / *Pas-spécifique-B* / *Pas-de-base*. Ces chorégraphies de salsa commencent et se terminent par deux pas de base de salsa : *forward-basic-step*, *backward-basic-step*. Ainsi ces deux pas de base sont effectués à des vitesses différentes au sein de ces trois chorégraphies. Afin d'améliorer le minutage à travers tous les danseurs de salsa, un compte à rebours indiquant le début de chaque chorégraphie a été rajouté sur la musique.

Ces chorégraphies de salsa ont été réalisées par 11 danseurs de niveaux allant du débutant à l'expert.

Tableau 4. Listes des différentes chorégraphies de salsa de notre corpus TPT-Dance&Actions

Nom des chorégraphies	Séquence de pas de salsa	Durée [s]	Tempo [bpm]
css1	*forward-basic-step, backward-basic-step forward-basic-step, backward-basic-step, right-turn, backward-basic-step, cross-body, backward-basic-step, forward-basic-step, backward-basic-step.*	18	157
css2	*forward-basic-step, backward-basic-step, forward-basic-step, pre-backward-step, suzie-q, suzie-q, forward-basic-step, pre-backward-step, double-cross, double-cross, forward-basic-step, backward-basic-step.*	18	180
css3	*forward-basic-step, backward-basic-step, forward-basic-step, pachanga-tap, pachanga-tap, backward-basic-step, forward-basic-step, backward-basic-step, swivel-tap, swivel-tap-return, forward-basic-step, backward-basic-step.*	19	185

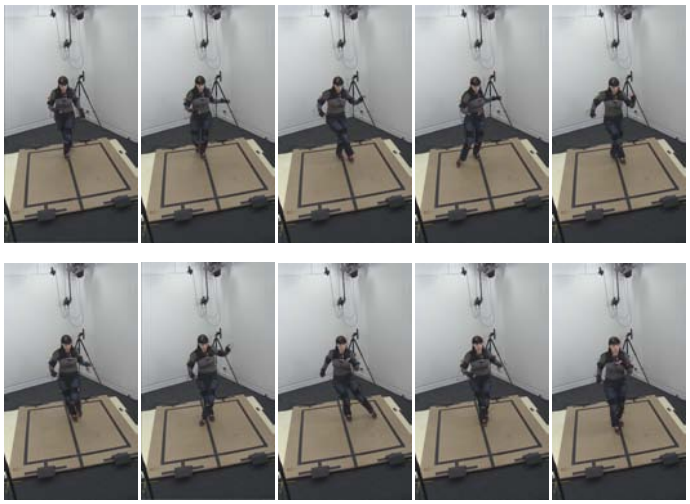

Figure 6. Frise photographique illustrant le pas Double-Cross présent dans la chorégraphie de salsa css2

5.2.1.3. Danse classique

Pour la danse classique, il n'y a pas à proprement parler de pas de base. Néanmoins nous avons demandé à la danseuse chargée d'écrire les différentes chorégraphies de danse classique, de choisir quelques pas qu'elle répète au sein de différentes chorégraphies. Aussi, des musiques de tempi différents ont été choisies, toujours dans un souci d'avoir différentes vitesses d'exécution des différents pas de danse. Cette base plus petite regroupe 4 danseuses classiques.

Tableau 5. Listes des différentes chorégraphies de danse classique de notre corpus TPT-Dance&Actions

Nom des chorégraphies	Séquence de pas de danse classique	Durée [s]	Tempo [bpm]
cdc1	3 *dégagés, temps lié, (idem derrière),* 3 *dégagés* 2*nde, glissade,* 2 *retirés.*	31	70
cdc2	3 *glissades, dégagé* 2*nde, dégagé* 4*ème, raccourcit, développé arabesque, pas de bourrée, développé à la* 2*nde, plié pas de bourrée, grande* 4*ème, grand port de bras, arabesque, rond de jambe.*	35	85
cdc3	*dégagé* 2*nde, plié* 5*ème, relevé retiré, dégagé* 2*nde, posé* 4*ème, tour en dehors* 4*ème, coupé, pas de bourrée, dégagé* 4*ème, grande* 4*ème, tour en dedans, glissade, relevé* 5*ème.*	16	115

Figure 7. Frise photographique illustrant le pas Glissade présent dans chaque chorégraphie de danse classique de notre corpus TPT-Dance&Actions

5.2.2. Séquences d'actions isolées

Ces séquences d'actions sont composées d'actions simples, d'activités sportives et de postures statiques. Elles sont décrites dans le tableau 6. Afin que ces actions soient faites dans le même minutage pour tous les participants, une vidéo de tutoriel de chaque séquence leur est diffusée lors des enregistrements. Dans un premier temps, la position neutre est présentée aux participants, *cf.* figure 8. Cette position est le point de départ et d'arrivée de chaque action. Avant chaque action, un tuteur nomme et montre une interprétation de l'action à exécuter face caméra. Puis s'ensuit une série de trois bip sonores indiquant au participant le début de l'exécution de chaque action et enfin une indication vocale, "position neutre", indique la fin de l'action concernée. Pour créer ce corpus de séquences d'actions isolées, nous avons réuni 16 personnes différentes.

Tableau 6. Listes des différentes séquences d'actions simples isolées de notre corpus TPT-Dance&Actions

Nom des séquences	Séquence d'actions isolées	Durée [s]
cgs1	*faire coucou de la main, position neutre, frapper à une porte, position neutre, frapper une fois des mains.*	23
cgs2	*lancer un objet avec une main, position neutre, pousser avec les deux mains, position neutre, se gratter la tête.*	26
cgs3	*faire semblant de regarder avec des jumelles, position neutre, croiser les bras, position neutre, mettre ses mains sur ses hanches.*	28
cgs4	*faire "oui" de la tête, position neutre, faire "non" de la tête, position neutre, s'incliner comme devant un public.*	21
cgs5	*tendre la main pour saluer, position neutre, faire geste du temps mort, position neutre, donner un coup de pied.*	24

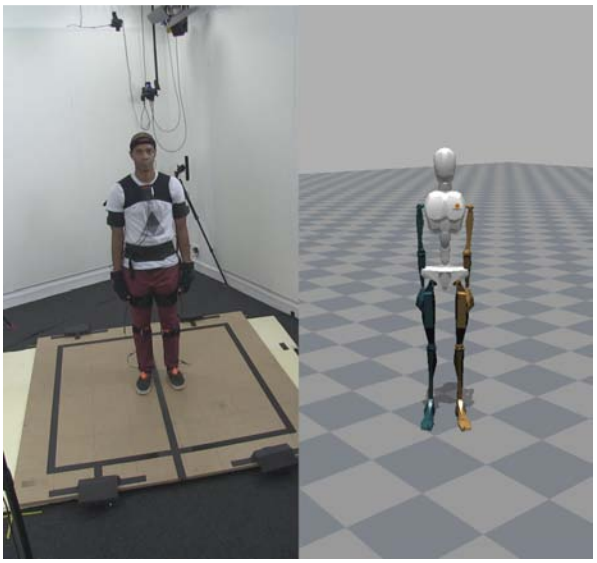

Figure 8. Exemple d'un participant en position neutre (à gauche) et de son avatar produit par le système Xsens (à droite). Le participant se tient debout, les bras le long du corps, les jambes droites, les pieds parallèles et écartés d'une largeur de bassin et le regard droit

Tableau 7. Listes des différentes séquences de fitness de notre corpus TPT-Dance&Actions

Nom des séquences	Séquence de mouvements de fitness	Durée [s]	Tempo [bpm]
cfs1 & **cfq1**	*frapper des mains, position neutre, faire des pas latéraux.*	37 & 25	100 & 130
cfs2 & **cfq2**	*marcher sur place, position neutre, courir sur place, position neutre, donner des coups de poing.*	41 & 35	100 & 130
cfs3 & **cfq3**	*faire les mouvements de bras du crawl, position neutre, faire les mouvements de bras de la brasse.*	39 & 30	100 & 130
cfs4 & **cfq4**	*faire des squats* [a]*, position neutre, faire des jumping jacks* [b]*.*	38 & 28	100 & 130

a. un mouvement cyclique où les genoux se plient jusqu'à former un angle de 90 deg puis se déplient avec les mains au niveau des épaules, les pieds parallèles, écartés d'une largeur de bassin et le dos bien droit

b. un saut partant de la position debout avec les jambes collées et les bras le long du corps arrivant à une position où les jambes sont écartées et les bras au dessus de la tête

5.2.3. Séquences de fitness

Les séquences de fitness sont constituées de différentes activités sportives répétées quatre à cinq fois de suite. Ceci nous permet d'effectuer une seule prise pour chaque séquence de fitness. De plus les séquences de gestes de fitness sont redondantes deux à deux. Les couples de séquences, (cfs1, cfq1), (cfs2, cfq2), (cfs3, cfq3) et (cfs4, cfq4), sont identiques mais avec une musique d'accompagnement différente (et de tempo différent), *cf.* tableau 7. Ceci nous permet ainsi d'obtenir plusieurs vitesses d'exécution de ce type de mouvements. Pour les séquences de fitness, les indications du tutoriel sont un peu différentes de celles des actions isolées. La "position neutre" encadre les répétitions successives des mouvements de fitness. Avant chaque série de gestes, une personne nomme et montre l'activité à exécuter successivement plusieurs fois puis un compte à rebours vocal, "1, 2, 3", indique le début de chaque série et la fin par une indication vocale, "c'est bon". Les personnes ayant réalisé les gestes simples ont également toutes réalisé les séquences de fitness. Elles sont donc au nombre de 16.

6. Matériel d'enregistrement

Au sein de cette partie, nous allons décrire en détail l'installation matérielle (*cf.* figure 1) mise en place pour ces enregistrements. Un schéma récapitulatif des interconnexions entre les différents équipements est présenté à la figure 9.

6.1. Plateforme logicielle d'enregistrement

Afin de contrôler tous les équipements utilisés pour ces enregistrements multimodaux, nous avons développé une plateforme logicielle appelée TPT - Studio. Cette plateforme permet de contrôler, depuis un des ordinateurs, les différentes étapes d'enregistrements via une interface graphique, TPT - Studio maitre, et de programmes esclaves, TPT - Studio esclave. La centralisation du contrôle est nécessaire vu la multitude d'équipements matériels et logiciels employés, *cf.* figure 9. Ainsi, grâce cette interface (*cf.* figure 10), nous pouvons facilement choisir la séquence de gestes à enregistrer, lancer/stopper tous les processus d'enregistrements ainsi que d'annoter directement la qualité des prises (valide, problèmes techniques, erreur du participant) à partir d'un seul ordinateur. En effet, ces différentes informations de contrôle sont propagées localement aux différents équipements via ses programmes esclaves, TPT - Studio esclave.

Nous avons utilisé le langage de programmation `Python 2.7` [15] pour développer cette plateforme logicielle.

15. www.python.org

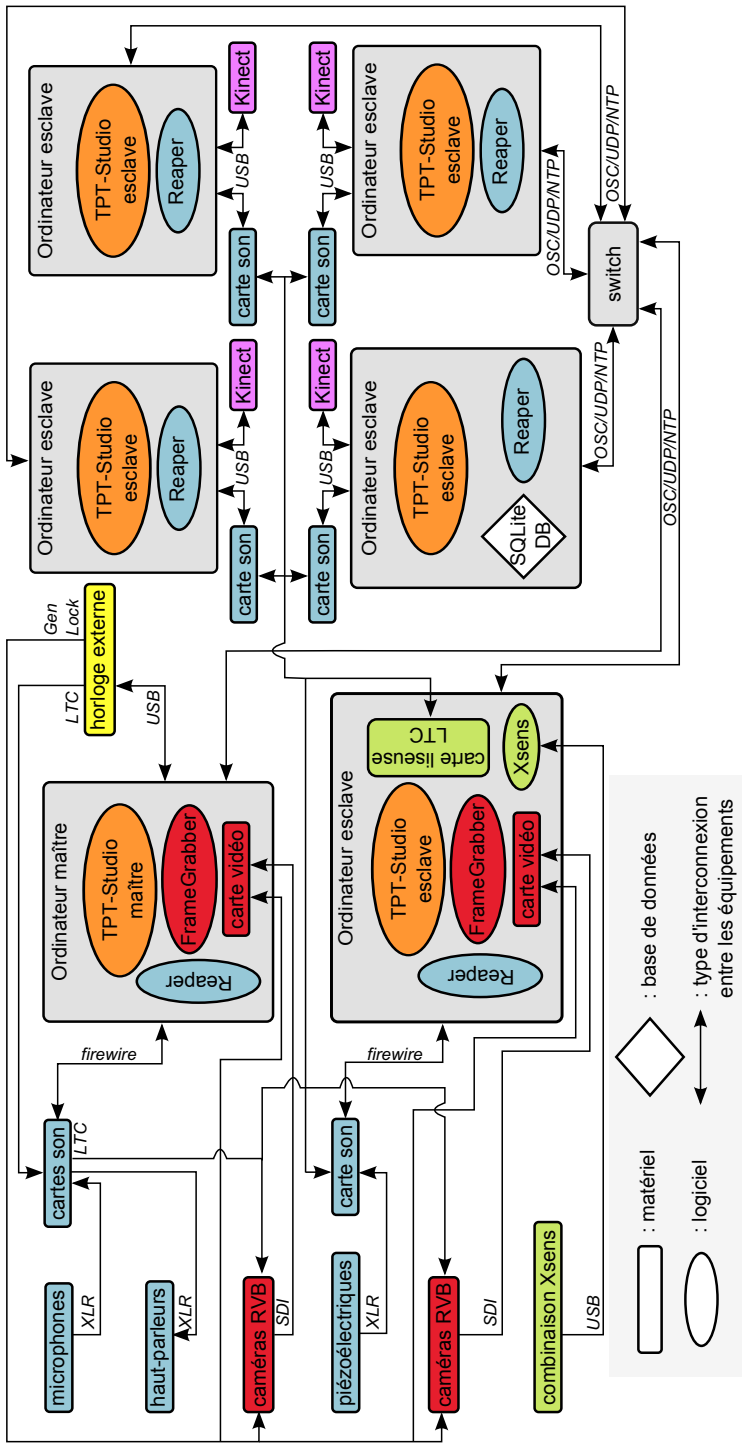

Figure 9. Schéma récapitulatif de l'ensemble des interconnexions entre les différents équipements présents au sein de notre système de capture de scènes multimodales

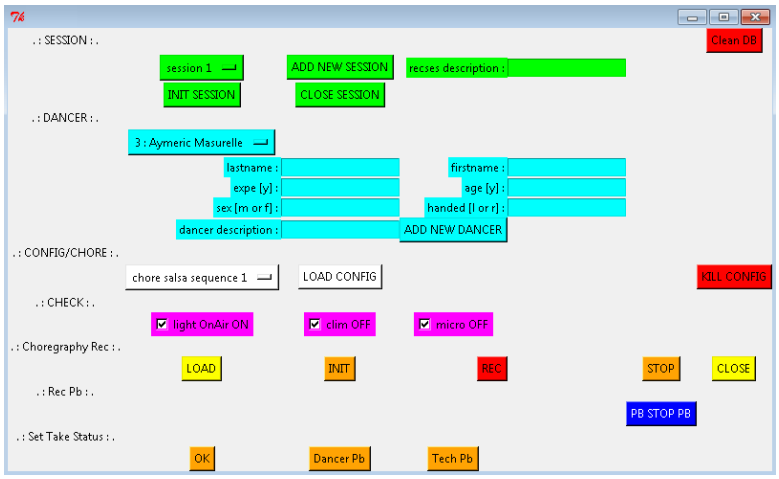

Figure 10. Interface graphique développée et utilisée pour contrôler les différentes étapes de nos enregistrements multimodaux

6.2. Équipement audio

L'équipement audio a été installé pour que la musique d'accompagnement, la voix du participant ainsi que le son de ses impacts de pas au sol puissent être enregistrés. Cette installation comprend 8 capteurs audio synchronisés. Quatre microphones sont placés au dessus de la piste de danse. Ce sont des microphones à condensateur omni-directionnels CMC6-MK2 de la marque Schoeps. Ces microphones nous permettent ainsi d'enregistrer la musique d'accompagnement, la voix et les bruits que les participants produisent. Pour enregistrer le son des impacts de pas au sol, nous avons collé sur la piste de danse quatre transducteurs piézoélectriques de guitare. Le placement de ces capteurs audio est donné dans la figure 1. Ces enregistrements ont été effectués en utilisant trois interfaces audio-numérique Echo Audiofire Pre8. Ces interfaces sont contrôlées par le logiciel Reaper [16] qui est lui-même contrôlé par notre plateforme logicielle de contrôle des enregistrements. Une synchronisation précise entre ces interfaces Audiofire Pre8 est obtenue grâce à un signal Word Clock S/PDIF qu'elles partagent. De plus pour effectuer une synchronisation avec les autres modalités nous enregistrons tous ces flux audio conjointement avec un signal LTC [17] commun délivré par l'horloge externe Rosendahl Nanosyncs HD. Les différents flux audio résultants sont enregistrés dans des fichiers séparés en mode mono. Les enregistrements faits avec ces interfaces Audiofire Pre8 ont une fréquence d'échantillonnage de 48kHz et une précision numérique de 32-bit. La musique d'accompagnement est jouée au moyen de deux haut-parleurs amplifiés placés dans le studio comme le montre la figure 1.

16. version 4.602 - www.reaper.fm
17. LTC (Linear TimeCode) est une encodage de données temporelles SMPTE en signal audio.

6.3. Équipement vidéo

6.3.1. Caméscopes Canon

Par l'intermédiaire de 2 interfaces vidéo Blackmagic Decklink Duo et de notre plateforme logicielle contrôlant ces dernières, les flux vidéo des caméscopes Canon sont transférés par câbles SDI et enregistrés directement sur 2 ordinateurs (PC Win7-64 bits Dell Precision T3600 (RAM 32Go, CPU Intel Xeon E5-1650 0 @ 3.20GHz 3.20GHz)). Etant donné que nous nous intéressons à l'enregistrement de gestes, nous avons choisi d'enregistrer ces mouvements en haute définition avec une résolution de 1 280 x 720 en mode progressif avec un débit d'images égal à 50 trames par seconde. Afin d'éviter de surcharger l'écriture sur les disques durs lors des enregistrements, les flux vidéo ne sont pas enregistrés sous leur forme brute, ils sont encodés en utilisant la norme H.264 à un débit binaire de 2 320kb/s, en préservant la résolution et le débit d'image. En plus du flux vidéo, chaque caméscope comprend 2 microphones internes pouvant être remplacés indépendamment par des entrées lignes. Ainsi nous avons choisi d'utiliser sur chaque caméscope un microphone interne pour capter les sons produits dans le studio et une entrée ligne à des fins de synchronisation. Ces flux audio sont encodés en PCM S16 avec une fréquence d'échantillonnage de 48kHz et une précision de 16-bit. Enfin les flux vidéo et audio de chaque caméscope sont encapsulés au format `.mkv`.

Afin de synchroniser précisément ces 4 caméscopes entre eux nous utilisons l'horloge externe Rosendahl NanoSyncs HD pour leur délivrer un signal GenLock [18]. De plus un signal LTC produit par l'horloge externe est enregistré par l'intermédiaire de l'entrée ligne de chaque caméscope permettant une synchronisation de ces flux avec les autres modalités.

6.3.2. Caméras Kinect

Grâce à chaque Kinect, nous pouvons enregistrer un flux vidéo, un flux de cartes de profondeur et 4 flux audio mono. Mais ces flux de données ne sont pas synchronisés en interne. Utilisant 4 Kinects, nous obtenons ainsi 16 flux audio groupés 4 par 4. Chacun de ces groupes de flux audio est enregistré sur un ordinateur séparé à l'aide de la librairie Kinect SDK [19] et du logiciel Reaper contrôlés par notre plateforme logicielle gérant les enregistrements. Ces flux sont enregistrés en utilisant le codec PCM S24 à une fréquence d'échantillonnage de 16kHz avec une précision de 32-bit. Afin de synchroniser ces flux audio avec les autres modalités, un signal LTC commun, délivré par une horloge externe Rosendahl Nanosyncs HD, est enregistré conjointement via une interface audio sur chaque ordinateur.

Le flux vidéo et le flux de cartes de profondeur de chaque Kinect sont enregistrés

18. Un signal GenLock garantit l'acquisition simultanée des images de plusieurs caméras
19. www.microsoft.com/en-us/Kinectforwindows

sur le même ordinateur que leurs flux audio. [20] Pour ceci, nous utilisons la librairie Kinect SDK contrôlée par notre plateforme logicielle gérant les enregistrements. Alors le flux vidéo et de cartes de profondeur de chaque Kinect sont enregistrés conjointement avec une résolution de 640x480 à environ 30 trames par seconde et sont stockés sous forme de matrices dans des fichiers `hdf5` [21]. Afin de synchroniser au mieux ces flux vidéo et de cartes de profondeur avec le reste des modalités, un serveur de temps NTP a été mis en place à travers tous les ordinateurs utilisés pour qu'ils aient tous la même date et heure système. Ainsi, pour chaque trame de cartes de profondeur, nous inscrivons le temps système lui correspondant. De plus, au début de chaque prise, nous stockons à un instant donné la correspondance de temps entre le temps système et le temps LTC. Cependant, en pratique, les temps NTP sur les différents ordinateurs, bien qu'ils proviennent d'un même serveur de temps, ne sont jamais parfaitement synchronisés. Au mieux, cette méthode nous permet de synchroniser ces flux vidéo et de cartes de profondeur avec une précision de $\pm\, 150$ ms. Nous sommes donc en train d'explorer des techniques de post-synchronisation automatique afin d'améliorer la synchronisation entre ces différents flux.

Il est important de noter que les Kinects ont été disposées de manière verticale pour couvrir au mieux le corps du participant malgré les dimensions de la pièce de capture de notre studio.

6.4. Équipement inertiel

Lors de l'enregistrement des différentes séquences de gestes, des données inertielles sont aussi collectées. Pour ceci, nous utilisons un dispositif Xsens MVN Biomech constitué d'une combinaison de 17 capteurs inertiels et d'un logiciel de traitement des données et de capture de mouvements. Au début de chaque session d'enregistrement, ces capteurs inertiels sont placés sur le corps du participant comme le montre la figure 8. Chacun de ces capteurs nous permet d'obtenir des données issues d'accéléromètres, de gyroscopes et de magnétomètres à une fréquence d'échantillonnage de 120 Hz. De plus, après avoir effectué la calibration de la combinaison, le logiciel de traitement des données nous permet d'obtenir la position dans l'espace de 22 articulations et de 23 segments du corps du participant, *cf.* figure 8.

Afin de pouvoir synchroniser ces données avec le reste des modalités, nous avons transmis le signal LTC de l'horloge externe Rosendahl au dispositif de capture de mouvement en utilisant une interface permettant de lire un signal LTC, Alpermann PCL PCIe LV. Ceci permet au logiciel de capture Xsens de pouvoir inscrire les temps SMPTE aux données inertielles associées. Cependant, en pratique les temps SMPTE obtenus ne correspondent pas aux temps lus par l'interface Alpermann, ce qui semble être lié à un bug du logiciel Xsens. Ainsi une post-synchronisation est nécessaire pour

20. Nous avons dédié un ordinateur par Kinect pour éviter de surcharger l'écriture sur leur disque dur.
21. www.hdfgroup.org/HDF5

aligner les données provenant de ce type d'équipement à celles des autres modalités et peut être effectuée grâce à la "procédure de clap" présente à la fin de flux de données.

7. Annotations des séquences de gestes

Les annotations des différentes séquences de gestes sont disponibles au sein de ce corpus dans des fichiers `.txt`. Chaque action, geste de fitness et pas de danse y est nommé et ses limites temporelles idéales sont exprimées en temps SMPTE :

<temps_SMPTE_début_geste>, <temps_SMPTE_fin_geste>, <nom_geste>

Pour les différentes séquences, les annotations indiquent le label de chaque action, geste de fitness ou pas de danse en respect des indications vocales et/ou de la pulsation musicale dans le cas du fitness et des chorégraphies de danse.

Les annotations des différentes chorégraphies de danse ont été effectuées avec un professeur de chaque style de danse, elles nomment les différents pas de danse réalisés en accord avec la pulsation des musiques d'accompagnement. Aussi, pour les chorégraphies de salsa, des annotations manuelles de la position temporelle des mesures et de la pulsation ont été effectuées par un musicien expert. Elles sont disponibles au sein de fichier `.csv`:

<temps_SMPTE>, <numéro_mesure>, <numéro_temps>

8. Préparation des données et leur mise en disponibilité

8.1. Nettoyage des données

Seules les prises considérées comme valide ont été incorporées au corpus. Nous considérons une prise comme valide lorsque le participant arrive à la fin d'une séquence en respectant l'emplacement temporel et les mouvements des gestes à effectuer et que l'ensemble des modalités ont été correctement enregistrées. Ceci a été effectué en assurant la présence de la "procédure de clap" à la fin de chaque prise.

8.2. Synchronisation multimodale

La figure 11 nous montre une vue d'ensemble de la stratégie de synchronisation des différentes modalités entre elles. Afin de comprendre correctement la logique derrière le schéma de notre stratégie de synchronisation, il est important de garder à l'esprit que certains sous-ensembles de flux de données sont déjà synchronisés grâce à notre installation.

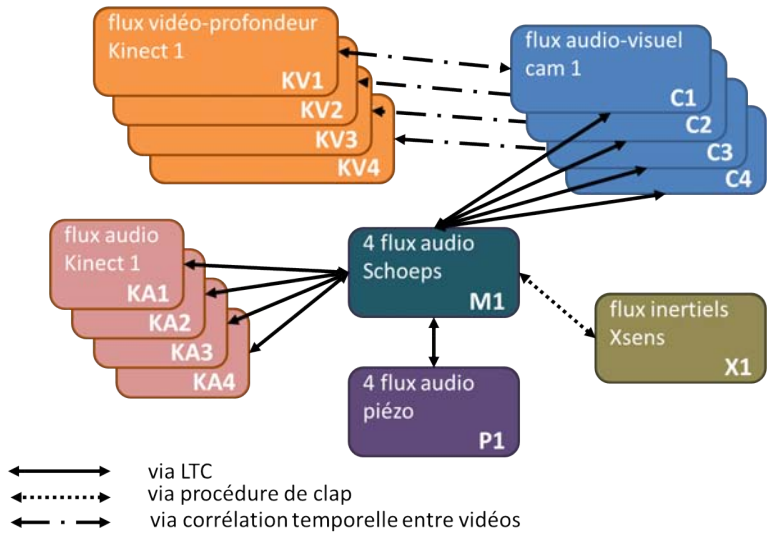

Figure 11. Vue d'ensemble de la stratégie de synchronisation des différentes modalités entre elles

Ces sous-ensembles sont :

– le sous-ensemble M1 formé de 4 flux audio provenant des microphones Schoeps parfaitement synchronisés entre eux, *cf.* section 6.2,

– le sous-ensemble P1 constitué de 4 flux audio provenant des piézoélectriques parfaitement synchronisés entre eux, *cf.* section 6.2,

– les sous-ensembles KA1, KA2, KA3 et KA4 composés chacun de 4 flux audio synchronisés provenant des microphones de chacune des Kinects, *cf.* section 6.3,

– les sous-ensembles KV1, KV2, KV3 et KV4 formés des flux vidéo et de cartes de profondeur synchronisés de chaque Kinect, *cf.* section 6.3,

– les sous-ensembles C1, C2, C3 et C4 constitués des flux vidéo et audio de chaque caméscope Canon parfaitement synchronisés entre eux, *cf.* section 6.3,

– le sous-ensemble X1 composés des flux synchronisés des données des différents capteurs inertiels et de la position des articulations principales du participant, *cf.* section 6.4.

Pour obtenir une synchronisation globale de tous les flux de données, il suffit de synchroniser une des instances de chaque sous-ensemble avec un flux de données d'un autre sous-ensemble. Nous proposons ainsi les trois étapes suivantes afin de réaliser la procédure de synchronisation globale :

– synchroniser les flux audio de Kinects, les flux des piézoélectriques et les flux audio-visuels des caméscopes avec les flux audio des microphones grâce au signal LTC qu'ils ont en commun (*cf.* section 8.2.1) ;

– synchroniser un des canaux audio avec les flux de données inertielles du système Xsens à l'aide de la procédure de clap (*cf.* section 8.2.2) ;

– synchroniser chaque flux vidéo des Kinects avec le flux vidéo d'un des caméscopes à l'aide d'une méthode de corrélation temporelle (*cf.* section 8.2.3).

Les différents écart-temporels résultants de ces processus de synchronisation sont inscrits dans des fichiers `.txt` de la manière suivante :

$$<\textit{nom\_fichier\_media}1> \# <\textit{nom\_fichier\_media}2> \# <\textit{écart\_temporel}>$$

8.2.1. Synchronisation audio avec un signal LTC

Comme expliqué précédemment, un signal audio de référence temporel, un signal LTC, a été distribué et enregistré au sein de chacun des sous-groupes de flux audio (M1, P1, KA1 $\rightarrow$ KA4) et de flux audio-visuel (C1 $\rightarrow$ C4). Pour obtenir une synchronisation entre ces sous-groupes, chaque piste audio LTC de ces sous-ensembles est décodée dans un premier temps. Le décodage de ce signal est réalisé en utilisant la librairie `SyncTools` [22]. Nous obtenons ainsi pour chacun de ces sous-groupes une correspondance entre le code temporel LTC (exprimé en heures, minutes, secondes et images) et leur temps relatif correspondant exprimé en seconde. Ainsi pour connaître l'écart temporel entre les flux de données des différents sous-groupes concernés, il suffit de soustraire leur temps relatif respectif correspondant à un même code temporel. Cette méthode nous permet d'effectuer la synchronisation de ces différents groupes de flux de données quelque soit leur fréquence d'échantillonnage avec une précision inférieure à 40 ms.

8.2.2. Synchronisation avec une procédure de clap

La synchronisation entre les données audio et inertielles est obtenue en maximisant la corrélation croisée entre des caractéristiques audio et inertielles spécifiques extraites au niveau de la procédure de clap au sein des flux concernés. Ces caractéristiques ont été choisies pour décrire au mieux la signature de la procédure de clap. La caractéristique extraite des flux audio est issue d'une fonction de détection d'attaque de notes de musique (Alonso *et al.*, 2005). Dans notre cas, cette fonction est appliquée à un des flux de données provenant d'un des microphones situés au dessus des participants qui clairement capturent le son produit par le clap des mains et des pieds. Cette caractéristique permet ainsi de mettre en exergue les instants où la procédure de clap se produit. Pour l'extraction des caractéristiques inertielles nous utilisons le signal des capteurs situés sur les mains et les pieds des participants. Un clap est représenté par un large pic au sein du signal des accéléromètres des membres concernés. Pour détecter cette procédure de clap, les signaux des trois axes de chacun de ces accéléromètres sont additionnés. La moyenne, le maximum et le temps correspondant sont alors calculés en utilisant une fenêtre glissante de 150 ms avec un pas de 10 ms. Enfin la fenêtre contenant la variance la plus grande est identifiée comme la signature d'un

22. www.musicsensorsemotion.com/2012/02/28/synctools

clap. Dû aux traitements précédents, la fréquence d'échantillonnage de la caractéristique audio est de 360 Hz et celle de la caractéristique inertielle est de 100 Hz. Pour appliquer la corrélation croisée entre ces caractéristiques, nous suréchantillonnons le signal de la caractéristique inertielle. Ainsi, nous obtenons une synchronisation des données audio et inertielles avec une précision de 10 ms. Cependant, des erreurs de post-synchronisation sont obtenues avec ce processus. Nous sommes donc en train d'étudier d'autres méthodes plus robustes afin de minimiser ces erreurs. De plus, une validation manuelle de cette post-synchronisation sera menée pour éviter de potentiels erreurs.

8.2.3. Synchronisation par corrélation temporelle entre vidéos

Dans cette section, nous décrivons l'approche utilisée pour calculer l'écart-temporel entre deux flux vidéo non-synchronisées de qualités et de fréquences d'échantillonnage différentes (vidéo Kinect : 640 x 480 px, 30 fps et vidéo Canon : 1 280 x 720 px et 50 fps). Cet écart-temporel est obtenue en maximisant la corrélation croisée entre des caractéristiques temporelles extraites des flux vidéo des Kinects et des caméscopes Canon. Ces caractéristiques sont obtenues en calculant une fonction de changement d'apparence entre les trames vidéos successives (Ushizaki *et al.*, 2006). Ainsi, cette quantité est obtenue pour chaque trame. Comme les différents flux vidéos ne sont pas de la même fréquence d'échantillonnage, les caractéristiques issues des Kinect sont suréchantillonnées à la fréquence d'échantillonnage de celles des caméscopes Canon. Cette approche est appropriée lorsque les caméras vidéos sont fixes et ne requiert pas un haut degré de compréhension de la scène. Il en résulte une post-synchronisation des flux vidéo avec une précision d'environ 30 ms. Cependant, ce processus de post-synchronisation effectue des erreurs dans le calcul des écart-temporel entre les flux vidéo non-synchronisés. Ainsi, des méthodes plus robustes sont en train d'être mise au point pour réduire ce nombre d'erreurs. De plus, une validation manuelle sera menée afin d'éviter de potentiels erreurs.

8.3. Disponibilité des données

Ce corpus regroupant des actions simples, des gestes de fitness, des chorégraphies de salsa, de lindy hop et de danse classique sera prochainement accessible au public via un site internet [23] pour une utilisation à des fins de recherche.

9. Champs d'applications

Cette base de données peut être exploitée par la communauté scientifique afin d'encourager la recherche dans plusieurs champs émergents de recherche.

Ce corpus constitué de données issues de plusieurs réseaux de capteurs est particulièrement adéquat pour évaluer des algorithmes d'estimation de pose du corps hu-

23. www.tsi.telecom-paristech.fr/aao/?p=1064

main. Ce champs de recherche est fondamental au regard de l'analyse du mouvement du corps humain. Ces dernières années de remarquables progrès ont été faits dans ce domaine et en particulier au niveau de l'estimation de pose du corps humain sans utilisation de marqueurs visuels. Par exemple, Hofmann et Gavrila ont développé un système d'estimation 3D de pose de la partie haute du corps humain en utilisant un réseau de caméras dans des environnements complexes (Hofmann, Gavrila, 2012).

Lorsque les poses du corps humain sont estimées au cours du temps, le terme d'analyse du mouvement du corps humain est employé. Toutes avancés sur ce sujet de recherche bénéficient d'un large éventail d'applications dans les domaines suivants : réalité virtuelle, interaction homme-machine, protection et surveillance.

Ji et Liu ont écrit une vue d'ensemble des récentes avancées dans le domaine de l'analyse des mouvements du corps humain utilisant des méthodes invariantes selon le point de vue de flux vidéos, et en particulier pour la représentation et de l'estimation de pose du corps humain ainsi que pour la représentation et la reconnaissance d'activité humaine (Ji, Liu, 2010). L'analyse des mouvements du corps humain s'est développée de manière significative au sein de la communauté de la vision par ordinateur lors des dernières décennies.

Alexiadis *et al.* ont créé un système d'évaluation de performance de danse utilisant la position d'articulations du corps humain (Alexiadis *et al.*, 2011). Cette évaluation est effectuée par comparaison entre la performance concernée et une performance de référence. Elle est délivrée au participant sous forme de retour visuel dans un environnement virtuel 3D. Essid et al. ont amélioré ce système d'évaluation de performance de danse afin qu'il intègre dans son traitement des modalités audio et inertielle (Essid, Alexiadis *et al.*, 2012).

Le bas coût des capteurs inertiels ainsi que leurs remarquables améliorations technologiques en terme de taille et de consommation d'énergie permettent de considérer ce type de capteur comme une solution alternative pour l'analyse des mouvements du corps humain. Roetenberg et al. nous propose une méthode utilisant une combinaison de capteurs inertiels pour estimer la position dans l'espace des articulations principales du corps humain (Roetenberg *et al.*, 2009). Aussi, une vue d'ensemble des techniques utilisant des flux de données inertielles pour classifier des activités humaines est présentée par Altun et Barshan (Altun, Barshan, 2010). Ces méthodes peuvent être exploitées dans différents domaines d'application tels que la réalité virtuelle, la rééducation, l'étude des sports et de la danse.

Ainsi ce corpus multimodal d'actions, de gestes de fitness et de pas de danse est une base de données de différents types d'activités humaines de différents niveaux de complexités. C'est un support complet et intéressant pour développer et évaluer des approches d'estimation des poses ou d'analyse des mouvements du corps humain. De plus, le fait que les données collectées soient multimodales permet l'évaluation des méthodes utilisant différents types de données d'entrée (*i.e.* données RVB, cartes de profondeur, données inertielles). Ainsi, une comparaison entre les performances de divers algorithmes utilisant un seul type de données ou une fusion de plusieurs modalités peut être réalisée.

10. Conclusions

Dans cet article, nous avons présenté une nouvelle base de données multimodales s'adressant aux domaines de recherche liés à l'analyse de scènes multimodales. Ce corpus est attrayant sous plusieurs aspects :

– son téléchargement et son utilisation sont gratuits ;

– il est constitué d'enregistrements multimodaux de réseaux de capteurs synchronisés et non-synchronisés ;

– ses enregistrements comprennent : l'enregistrement de plusieurs sources audio dont le son des pas des participants, l'enregistrement de flux vidéos de haute et basse définition, l'enregistrement de cartes de profondeurs, l'enregistrement de flux de données inertielles issue d'une combinaison de capteurs portée par les participants, l'enregistrement des positions des articulations principales des participants ;

– un nombre important de participants y ont pris part ;

– de nombreux gestes avec différents niveaux de complexités accompagnés ou non de musique y sont présents ;

– des annotations de la vérité terrain y sont fournies.

Une fois les problèmes de synchronisation résolus, *i.e.* les étapes de post-synchronisation effectuées, cette nouvelle base de données pourra être employée pour illustrer, développer et tester une variété d'outils de divers domaines de recherche. Par exemple, à partir de ces enregistrements multimodaux, ce corpus pourra être utilisé pour développer des systèmes d'analyse d'activités humaines comme l'estimation de pose du corps humain, l'analyse des mouvements du corps humain et la reconnaissance de pas de danse, de gestes sportifs ou bien d'actions. De nouvelles techniques utilisant des combinaisons de diverses modalités afin d'améliorer les performances de systèmes d'interaction homme-machine pourront alors être développées.

Remerciements

Les auteurs remercient chaleureusement toutes les personnes ayant contribuées à la création de ce corpus, et spécialement:

– les participants: Gaël, Bertrand, Anaïs, Ariane, Stéphane, Laurent, Eric, Sylvie R., Rachel, Roland, Floriane B., Zhoungwei, Youcef, Damien, Floriane D., Eve-Marie, Pascal, Sylvie M., Juliette, Albane, Xabier, Hequn, Caroline, Anne-Claire, Alexis, Clément, Nicolas, Frédéric, Thomas, Paul R., François, Paul M., Romain, Simon L., Victor, Mathieu, Nesrine, Florian, Brian, Simon D., Chloé, Alice, et,

– les techniciens: Rida, Slim, Stéphane, Gilbert, Michel, Nesrine, Mounira.

Cette recherche a été financée en partie par le projet européen REVERIE FP7-287723.

Bibliographie

Alexiadis D. S., Kelly P., Daras P., O'Connor N. E., Boubekeur T., Moussa M. B. (2011). Evaluating a dancer's performance using kinect-based skeleton tracking. In *Proceedings of*

conference on multimedia, p. 659–662.

Alonso M., Richard G., David B. (2005). Extracting note onsets from musical recordings. In *Proceedings of conference on multimedia and expo, 2005*, p. 1–4.

Altun K., Barshan B. (2010). Human activity recognition using inertial/magnetic sensor units. In *Proceedings of workshop on human behavior understanding*, p. 38–51.

De la Torre F., Hodgins J. K., Montano J., Valcarcel S. (2009). Detailed human data acquisition of kitchen activities: the cmu-multimodal activity database (cmu-mmac). In *Proceedings of workshop on developing shared home behavior datasets to advance hci and ubiquitous computing research.*

Essid S., Alexiadis D., Tournemenne R., Gowing M., Kelly P., Monaghan D. *et al*. (2012). An advanced virtual dance performance evaluator. In *Proceedings of conference on acoustics, speech, and signal processing*, p. 2269-2272.

Essid S., Lin X., Gowing M., Kordelas G., Aksay A., Kelly P. *et al*. (2012). A multi-modal dance corpus for research into interaction between humans in virtual environments. *Journal on Multimodal User Interfaces: Special issue on multimodal corpora*, vol. 7, n° 1-2, p. 157-170.

Gkalelis N., Kim H., Hilton A., Nikolaidis N., Pitas I. (2009). The i3dpost multi-view and 3d human action/interaction database. In *Proceedings of conference for visual media production*, p. 159–168.

Gorelick L., Blank M., Shechtman E., Irani M., Basri R. (2007). Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, n° 12, p. 2247–2253.

Gowing M., Ahmadi A., Destelle F., Monaghan D., O'Connor N., Moran K. (2014). Kinect vs. low-cost inertial sensing for gesture recognition. In *Proceedings of conference on multimedia modeling*, p. 484–495.

Hofmann M., Gavrila D. (2012). Multi-view 3d human pose estimation in complex environment. *Journal on Computer Vision*, vol. 96, n° 1, p. 103-124.

Ji X., Liu H. (2010). Advances in view-invariant human motion analysis: A review. *Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 40, n° 1, p. 13–24.

Kadous M. (2002). *Temporal classification: Extending the classification paradigm to multivariate time series*. Thèse de doctorat non publiée, School of Computer Science & Engineering, University of New South Wales.

Laptev I., Marszałek M., Schmid C., Rozenfeld B. (2008). Learning realistic human actions from movies. In *Proceedings of conference on computer vision & pattern recognition*, p. 1–8.

Liu J., Luo J., Shah M. (2009). Recognizing realistic actions from videos "in the wild". In *Proceedings of conference on computer vision & pattern recognition*, p. 1996–2003.

Marszałek M., Laptev I., Schmid C. (2009). Actions in context. In *Proceedings of conference on computer vision & pattern recognition*, p. 2929 – 2936.

Masurelle A., Essid S., Richard G. (2013). Multimodal classification of dance movements using body joint trajectories and step sounds. In *Proceedings of workshop on image and audio analysis for multimedia interactive services*, p. 1–4.

Masurelle A., Essid S., Richard G. (2014). Gesture recognition using a nmf-based representation of motion-traces extracted from depth silhouettes. In *Proceedings of conference on acoustics, speech, and signal processing*, p. 1275–1279.

Rodriguez M., Ahmed J., Shah M. (2008). Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In *Proceedings of conference on computer vision & pattern recognition*, p. 1–8.

Roetenberg D., Luinge H., Slycke P. (2009). *Xsens mvn: Full 6dof human motion tracking using miniature inertial sensors*. Rapport technique. Xsens.

Schüldt C., Laptev I., Caputo B. (2004). Recognizing human actions: A local svm approach. In *Proceedings of conference on pattern recognition*, p. 32–36.

Sigal L., Balan A., Black M. (2010). Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *Journal on Computer Vision*, vol. 87, p. 4–27.

Singh S., Velastin S., Ragheb H. (2010). Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In *Proceedings of conference on advanced video and signal based surveillance*, p. 48–55.

Ushizaki M., Okatani T., Deguchi K. (2006). Video synchronization based on co-occurrence of appearance changes in video sequences. In *Proceedings of conference on pattern recognition*, p. 71-74.

Weinland D., Ronfard R., Boyer E. (2006). Free viewpoint action recognition using motion history volumes. *Journal on Computer Vision & Image Understanding*, vol. 104, n° 2-3, p. 249–257.

Yang A., Jafari R., Sastry S., Bajcsy R. (2009). Distributed recognition of human actions using wearable motion sensor networks. *Journal on Ambient Intelligence & Smart Environments*, vol. 1, p. 103–115.

Zhang Z. (2000). A flexible new technique for camera calibration. *Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, n° 11, p. 1330–1334.

Article soumis le 27/02/2015

Accepté le 9/12/2015

Annexe A. Tableaux récapitulatifs des bases de données existantes

Tableau 8. Tableau récapitulatif des bases de données monomodales

	base de données	media	gestes	observations
monomodale / uni-capteur	KTH human actions dataset (Schüldt *et al.*, 2004)	camera N&B	6 actions	25 personnes, en extérieur, en extérieur avec changement d'échelle, en extérieur avec d'autres vêtements, en intérieur.
	Weizmann human actions dataset (Gorelick *et al.*, 2007)	camera RVB	10 actions	9 personnes, en extérieur, sans mouvement de camera.
	Hollywood human actions dataset (Laptev *et al.*, 2008)	camera RVB	8 actions	à partir de 32 films.
	Hollywood-2 human actions and scenes dataset (Marszałek *et al.*, 2009)	camera RVB	12 actions	à partir de 3669 films, 20.1 heures de vidéos.
	Broadcast Television actions dataset (Rodriguez *et al.*, 2008)	camera RVB	9 actions	200 vidéos de chaînes télévisées, divers points de vue et scènes.
	YouTube actions dataset (Liu *et al.*, 2009)	camera RVB	11 actions	1160 vidéos, divers points de vue et scènes.
	Australian Sign Language v.1/v.2 (Kadous, 2002)	magnétomètre bas-coût/professionnel	95 signes	6 personnes/1 personne, 6650 exemples/2565 exemples.
monomodale / multi-capteurs	INRIA XMAS dataset (Weinland *et al.*, 2006)	5 cameras RVB synchronisées	11 actions	10 personnes, arrière-plan et luminosité fixes.
	i3DPost Multi-view Human actions dataset (Gkalelis *et al.*, 2009)	8 cameras RVB HD synchronisées	12 actions	8 personnes, caméras et luminosité fixes, fond bleu.
	MuHAVi dataset (Singh *et al.*, 2010)	8 cameras RVB	17 actions	14 personnes, caméras et luminosité fixes.
	WARD dataset (Yang *et al.*, 2009)	5 capteurs inertiels	13 actions	20 personnes.

Tableau 9. Tableau récapitulatif des bases de données multimodales

	base de données	media	gestes	observations
multimodale	HumanEva-I database (Sigal *et al.*, 2010)	4 cameras N&B, 3 cameras RVB, système de capture de mouvements synchronisés.	6 actions	4 personnes
	HumanEva-II database (Sigal *et al.*, 2010)	idem	6 actions	2 personnes
	CMU-MMAC database (De la Torre *et al.*, 2009)	6 caméras RVB, 5 microphones, 5 capteurs inertiels, système de capture de mouvements.	5 recettes	43 personnes, 5 cameras fixes et 1 embarquée.
	3DLife Actions dataset (Gowing *et al.*, 2014)	5 kinects, 5 capteurs inertiels.	22 actions	17 personnes, kinects fixes
	3DLife Dance dataset (Essid, Lin *et al.*, 2012)	2 kinects, 5 capteurs inertiels, 12 microphones, 9 cameras RVB, 4 piézoélectriques.	11 pas de salsa	5 chorégraphies de salsa, 15 danseurs, cameras et kinects fixes.
	3DLife Dance & Actions dataset	4 kinects, 4 cameras RVB HD, 4 microphones, 4 piézoélectriques, système de capture de mouvements	8 pas de lindy, 8 pas de salsa, 22 pas de classique, 9 mouvements de fitness et 15 actions.	20 danseurs, 14 chorégraphies (lindy hop, salsa, classique), 16 personnes, 8 séquences de fitness, 5 séquences d'actions, cameras et kinect fixes.