Expert Gesture Analysis through Motion Capture using Statistical Modeling and Machine Learning

Mickaël Tits
University of Mons
September 25, 2018

A dissertation submitted to the Faculty of Engineering of the University of Mons, for the degree of Doctor of Philosophy in Engineering Science

Jury:
Dr **Nicolas D'ALESSANDRO** ..... Hovertone
Prof. **Thierry DUTOIT**, supervisor ..... University of Mons
Prof. **Bernard GOSSELIN** ..... University of Mons
Dr **Alexis HELOIR** ..... University of Valenciennes
Prof. **Marc LEMAN** ..... Ghent University
Prof. **Xavier SIEBERT**, president ..... University of Mons
Dr **Joëlle TILMANNE**, co-supervisor ..... University of Mons
Prof. **Marcelo WANDERLEY** ..... McGill University

Thesis
Everyone has the same color.

Demonstration
The black sheep idiom is a well-documented research subject in the literature, as shepherds have existed from the depths of time, and have always been intrigued by that mystery (Marques et al., 1988). This demonstration will hence be based on this subject.
In a herd populated with a black and a white sheep, the white sheep is the only white sheep of the herd. It is thus an outsider in its herd and can therefore be considered as the black sheep of the herd. Hence, both sheep are black. The herd is thus composed of two black sheep. As all the sheep of the herd have the same color, there is no outsider, i.e. no black sheep, and both sheep are thus white. All the sheep thus have the same color. As it was demonstrated that humans are like sheep (Raafat et al., 2009), humans therefore have the same color. *QED*.

Mickaël Tits
September 25, 2018

Acknowledgments

Merci tout d'abord à Thierry et Nicolas pour m'avoir donné goût à la recherche dans les domaines passionnants du traitement du signal et de l'intelligence artificielle dès le début de mon Master. Merci à Julien Biral également pour avoir été un excellent compagnon de voyage, et ami, et d'avoir partagé avec moi cette passion pendant nos études de Master. C'est grâce à vous que cette passion a pu mûrir et se transformer d'abord en un TFE, puis en une thèse. Merci à Joëlle pour avoir pris le relais, et pour m'avoir suivi, soutenu et poussé à me surpasser constamment pendant ces quatre années. Nicolas, tu m'as transmis ta passion pour la folie de la recherche, nécessaire à la création et à la révolution des idées. Joëlle, tu m'as transmis ta passion pour la rigueur de la recherche, nécessaire à la réalisation des idées. J'espère que cette passion vous emmènera loin dans la sphère de l'entreprenariat, avec Hovertone. Je suis heureux d'avoir pu être votre padawan, et j'emmènerai également cette passion avec moi dans la suite de mon chemin.

Un grand merci également à Eric Caulier, Georgette Methens et Michèle Van Hemelrijk pour leurs conseils, leur disponibilité et les enseignements qu'ils m'ont donnés sur leur discipline passionnante qu'est le Taijiquan.

Merci à ma famille, et en particulier à mes parents, et à mes prédécesseurs sur le chemin de la thèse, Jacques Tits, André Tits, Pierre-Antoine Absil, Olivier Absil, Laurent Kohler, pour avoir cultivé et pour m'avoir transmis votre passion pour la science et la recherche. Je suis heureux d'avoir suivi cette voie que vous avez tracée, et heureux également d'être déjà suivi de près. Merci, Noé, d'avoir partagé cette passion également avec moi pendant un an. Je te souhaite bonne chance pour la suite. Merci également à Isabelle et Pierre-olivier pour m'avoir soutenu également tout ce temps.

Merci à tous mes amis également pour leur soutien et leurs encouragements tout au long de ce travail de longue haleine. En particulier, merci Ambroise, tu as été un excellent compagnon, et ami, de voyage et de Citizen Fox. Merci pour m'avoir fait profiter de ton karma pendant ces quatre années. Ton altruisme, ton empathie et ton équanimité ont été une inspiration pour moi. Merci pour avoir tant de fois relu mes publications, et réfléchi posément avec moi sur des problèmes scientifiques ou existentiels. J'espère que tu trouveras le chemin de la pleine conscience, je te souhaite bonne chance pour la fin de ta thèse.

Merci à tous mes collègues du TCTS pour votre bonne humeur et votre passion également partagée au labo. Vous avez été comme une deuxième famille pour moi

(sauf Noé, qui est déjà de la première), et j'espère retrouver ça lors de ma prochaine étape, au CETIC.

Enfin, merci à ma chère et tendre, Julie, pour m'avoir soutenu et encouragé, et pour avoir supporté mes maux de tête, de ventre, de dos et mon indisponibilité chroniques pendant ces quatre années de thèse. Tu mérites un titre de docteur.

Abstract

The present thesis is a contribution to the field of human motion analysis. It studies the possibilities for a computer to interpret human gestures, and more specifically to evaluate the quality of expert gestures. These gestures are generally learned through an empirical process, limited to the subjectivity and own perception of the teacher. In order to objectify the evaluation of the quality of these gestures, researchers have proposed various measurable criteria. However, these measurements are still generally based on human observation.

Enabled by significant steps in the development of Motion Capture (MoCap) and artificial intelligence technologies, research on automatic gesture evaluation has sparked a new interest, due to its applications in education, health and entertainment. This research field is, however, recent and sparsely explored. The few studies on the subject generally focus on a small dataset, limited to a specific type of gestures, and a data representation specific to the studied discipline, hereby limiting the validity of their results. Moreover, the few proposed methods are rarely compared, due to the lack of available benchmark datasets and of reproducibility on other types of data.

The aim of this thesis is therefore to develop a generic framework for the development of an evaluation model for the expertise of a gesture. The methods proposed in this framework are designed to be reusable on various types of data and in various contexts. The framework consists of six sequential steps, for each of which an original contribution is proposed in the present thesis:

Firstly, a benchmark dataset is proposed to promote further research in the domain and allow method comparison. The dataset consists of repetitions of 13 Taijiquan techniques by 12 participants of various levels from novice to expert, resulting in a total of 2200 gestures.

Secondly, the MoCap data must be processed, in order to ensure the use of high-quality data for the design of an evaluation model. To that end, an original method is proposed for automatic and robust recovery of optical MoCap data, based on a probabilistic averaging of different individual recovery models, and the application of automatic skeleton constraints. In an experiment where missing data were simulated into a MoCap dataset, the proposed method outperforms various methods of the literature, independently of gap length, sequence duration and the number of simultaneous gaps.

Thirdly, various motion features are proposed for the representation of various aspects of motion, potentially correlated with different components of expertise. Additionally, a new set of features is proposed, inspired by Taijiquan ergonomic principles. In this respect, 36 new motion features, representing aspects of stability, joint alignments, joint optimal angles and fluidity are presented.

Fourthly, the features must be processed to provide a more relevant representation of expertise. In the present work, the morphology influence on motion is addressed. Morphology is an individual factor that has a great influence on motion, but is not related to expertise. A novel method is therefore proposed for the extraction of motion features independent of the morphology. From the linear modeling of the relation of each feature with a morphological factor, residues are extracted, providing a morphology-independent version of the motion features. As a consequence, the resulting features are (i) less correlated between each other, and (ii) enable a more relevant comparison between the gestures of various individuals, hereby allowing a more relevant modeling of expertise. Results show that the method, termed as Morphology-Independent Residual Feature Extraction (MIRFE) outperforms a baseline method (skeleton scaling) in (i) reducing the correlation with the morphological factor, and in (ii) improving the correlation with skill, for various gestures of the Taijiquan MoCap dataset, and for a large set of motion features.

Fifthly, an evaluation model must be developed from these features, allowing the prediction of the expertise level on a new gesture performed by a new user. A model based on feature statistics, dimension reduction and regression is proposed. The model is designed to be used with any motion feature, in order to be generic and relevant in different contexts, including various users and various types of gestural disciplines. Trained on the Taijiquan MoCap dataset, the model outperforms two methods of the literature for the evaluation of gestures of a new user, with a mean relative prediction error of 10% ($R = 0.909$).

Additionally, a first exploration of the use of deep learning for gesture evaluation is proposed. To that end, MoCap sequences are represented as abstract RGB images, and used for transfer learning on a pre-trained image classification convolutional neural network. Despite a lower performance ($R = 0.518$), an analysis of the results suggests that the model could achieve better performance given a larger dataset, including a larger number of novices and experts.

Sixthly, and finally, to allow a practical use of the evaluation model, a feedback system must provide an intuitive interpretation of the predicted level, allowing an effective understanding and assimilation by the user of the system. In the present work, an original and generic feedback system is proposed, based on the synthesis of an improved gesture, and its comparison to the user's original gesture. Both intuitive and precise feedback are proposed, based on (i) synchronized visualization of both gestures, and (ii) striped images highlighting the motion features that need improvement. As a validation of the proposed method, examples of feedback are proposed for various sequences of the Taijiquan MoCap dataset, showing its practical interest for objective and automated supervision.

Contents

List of Figures

List of Tables

List of acronyms

CI	Confidence Interval
CNN	Convolutional Neural Network
CoM	Center of Mass
DCT	Discrete Cosine Transform
DNN	Deep Neural Network
DOF	Degree Of Freedom
DTW	Dynamic Time Warping
EN	Elastic Net
FC	Fully-Connected
GLR	Global Linear Regression
GRNN	Generalized Regression Neural Network
G-SVR	Gaussian-Support Vector Regression
GUI	Graphical User Interface
HCI	Human Computer Interaction
HMM	Hidden Markov Model
KNN	K-Nearest Neighbors
LGRNN	Local Generalized Regression Neural Network
LI	Local Interpolation
LMA	Laban Movement Analysis
LOPO	Leave-One-Participant-Out
LPR	Local Polynomial Regression
LSB	Least-Square Boosting
L-SVR	Linear-Support Vector Regression
MAE	Mean Absolute Error
MIRFE	Morphology-Independent Residual Feature Extraction
MLP	Multi-Layer Perceptron
MoCap	motion capture
NMF	Nonnegative Matrix Factorization
NN	Neural Network
PC	Principal Component
PCA	Principal Component Analysis
PM	Principal Movement
PMA	Probabilistic Model Averaging
PSO	Particle Swarm Optimization
ReLU	Rectified Linear Unit

RMSE	Root Mean Square Error
ROM	Range Of Motion
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
SVT	Singular Value Thresholding

Introduction

Context

The present thesis is a contribution to the field of human motion analysis. It studies the possibilities for a computer to interpret human gestures, and more specifically to evaluate the quality of an 'expert gesture'. An expert gesture can be defined as any complex and precise gesture requiring a high level of motor control acquired through experience, i.e. a long training process. This type of gesture is encountered in various disciplines in sports, music, dance, or in manual works like surgery, pottery and calligraphy. All these gestures are usually learned through a long training process, either self-learned or guided by a teacher. This process is usually empirical, and limited to the subjective perception of the discipline by the learner or the teacher. The perception of a gesture and the underlying characteristics defining expertise are indeed related to one's unique body, experience and personality, and are therefore partly subjective. Moreover, this perception is generally hard to describe with words and even harder to quantify, even for an expert of the discipline.

In order to objectify the evaluation of the quality of these gestures, researchers have proposed various measurable criteria, somehow defining what expertise is, also termed motor control, dexterity, or skills according to the research domain. In the field of physiotherapy, various quantitative tests were developed to evaluate the gross and fine motor skills of patients, with various exercises for measuring motor precision, integration, manual dexterity, uni- and bilateral coordination, balance, agility, and strength (Deitz et al., 2007; Cools et al., 2009). These tests have been widely used for diagnosis with patients with cerebral palsy or after a physical injury, or with children with development disorders. In ergonomics, evaluation matrices have been developed to measure the quality of manual works in terms of ergonomy, by evaluating muscular fatigue and risks of injury (McAtamney and Corlett, 1993; Kee and Karwowski, 2001). In a more artistic context, Laban movement analysis (LMA) is a method allowing the description and interpretation of motion in terms of intention, aesthetics and efforts, and is used by dancers, actors, but also by physiotherapists (Bartenieff and Lewis, 1980; Newlove, 1993). More specifically, the particular discipline of surgery has drawn a lot of attention, due to the impact of these gestures on patient health. In this context, various evaluation matrices have also been proposed for the supervision of the surgical training (Reiley et al., 2011).

Most of these gesture evaluation methods are still based on human observation or are limited to basic measurements of the motion, such as the running speed, or the duration of a manual work. However, the use of computer technologies could provide

more consistent, automatic and objective solutions to this issue. Recent motion capture (MoCap) technologies allow the accurate and automatic recording of the motion of the entire body. On the other hand, artificial intelligence technologies, and more particularly machine learning, allow a computer to interpret various types of data, based on automatic algorithms modeling the relations between these data. These technologies could be used together, allowing an automatic modeling of expertise from MoCap data representing the analyzed gestures accurately. The models developed could therefore be more precise and objective than the human observer, and could find applications in various areas. In sports, wearable MoCap sensors are more and more used for the monitoring of athletes and the evaluation of their performance (Camomilla et al., 2018). In the field of surgery, models based on artificial intelligence have been tested for the evaluation of the surgical process (Lalys and Jannin, 2014). In a medical context, MoCap has been used to investigate the long-term effect of medication on tremor for patients with Parkinson's disease (Van Someren et al., 1998). More recently, Tahir and Manap (2012) tested machine learning algorithms for the detection of the Parkinson's disease from walking patterns. In the context of sports competition, Young and Reinkensmeyer (2014) proposed a model based on machine learning for automatic and objective jury grades in an Olympic diving competition. In various contexts, specific computational features have been developed to objectively interpret various expertise criteria in gestures, including LMA (Aristidou and Chrysanthou, 2014), ergonomics (Andreoni et al., 2009) and physiotherapy (Harrison et al., 2007).

In general, automatic gesture evaluation models could be used either by a learner for automated supervision during the learning of an expert gesture, or as a tool by a teacher, a sports coach, a choreographer or a medical doctor for a more consistent and objective monitoring. Finally, the video-game industry could use these models for the development of new video games, allowing the learning of sports and musical gestures in an entertaining context.

Motivations and original contributions

Due to the recent spreading of MoCap and machine learning technologies, and due to the range of potential applications, the analysis of expert gestures has recently sparked a particular interest in research. This research field is, however, recent and sparsely explored. The few studies on the subject generally focus on a small dataset, limited to a specific type of gesture, and a data representation specific to the studied discipline, hereby limiting the validity of their results. Moreover, the few proposed methods are rarely compared, due to the lack of available benchmark datasets and of reproducibility on other types of data.

The aim of this thesis is therefore to develop a generic framework for the development of an evaluation model for the expertise of a gesture. The methods proposed in

this framework are designed to be reusable on various types of data and in various contexts. Moreover, a benchmark dataset is proposed to promote further research in the domain and allow method comparison. The proposed models must be designed to take into consideration various aspects of motion, in order to be generic and relevant in different contexts, including various users and various types of gestural disciplines. The proposed models should also allow for a practical use, either for automated supervision, or as a support for teachers, by providing a quantified and objective feedback to the user.

Fig 1 illustrates the proposed framework. The workflow of this framework includes six sequential steps. For each one of these steps, an original contribution is presented in this thesis:

1. First, a large dataset must be collected. To be relevant for gesture evaluation, the dataset must contain a large number of participants to encode a large variability of gestures, and a large number of expertise levels. In the present thesis, a dataset of Taijiquan gestures has been recorded, using 3D and accurate full-body MoCap. Most gestural disciplines are focused on the motion of a specific body part, or on a specific purpose such as gesture aesthetics, musical sound or force production. On the contrary, Taijiquan focuses on the movement itself, allowing for the development of general physical abilities such as balance, coordination, etc., as well as mental skills such as concentration. This general expertise learned during Taijiquan practice can often be transferred to various other sports disciplines (Caulier, 2010). These characteristics make Taijiquan a well-suited discipline to study gesture expertise. The dataset recorded contains 13 classes (Taijiquan gestures), performed by 12 participants of different levels of expertise from novice to expert. All the recordings have been manually corrected and segmented, resulting in 2200 gestures ($\approx$ 170 / class). The 12 participants have been ranked by three highly experienced Taijiquan teachers, providing an index of their global level of expertise. This dataset has been published and is available for free download for research purposes (Tits et al., 2018a).[1] To the author's knowledge, this is the first published dataset of sports gestures comprising simultaneously a large number of participants (12), a large number of different classes (13), and a variety of levels of expertise.

2. Secondly, the data must be processed, in order to ensure the use of high-quality data for the design of an evaluation model. To that end, an original method is proposed for automatic and robust recovery of optical MoCap data, based on a probabilistic averaging of different individual MoCap data recovery models.

3. Thirdly, relevant motion features must be extracted from the data. Motion features allow the representation of various aspects of motion, such as dynamics, semantics, ergonomics or expressivity. In the present context, relevant motion features are those that are related to expertise. In the present thesis, an

[1] Taijiquan MoCap dataset: https://github.com/numediart/UMONS-TAICHI

original high-level representation of motion data is proposed, inspired by the ergonomic principles of Taijiquan. The ergonomy of a gesture is closely related to expertise, but this aspect has never been explored in the literature known to the author. This new type of motion features allows a relevant interpretation of the motion with a high-level of abstraction, in terms of ergonomics.

4. Fourthly, the features must be processed to provide a more relevant representation of expertise. In the present work, the morphology influence on motion is addressed. Morphology is an individual factor that has a great influence on the motion, but is not related to expertise. For instance, during a kick gesture the foot of a tall person will generally move higher than the foot of a short person. On the contrary, if a particular height of the kick is aimed, then the hip angle of the taller person will be smaller. In both cases, some features of both individuals will be very different (either foot height, or hip angle), without any indication about the quality of the performance. In this sense, the information contained in any feature about morphology can be considered as noise and should therefore be reduced as much as possible. Moreover, this information is generally redundant as it is contained in many features. In this respect, a novel method is proposed for the extraction of motion features independent of the morphology. The proposed method is based on the modeling of the relation of each feature with a morphological factor. From this model, residues are extracted, providing a morphology-independent version of the motion features. As a consequence, the resulting features are (i) less correlated between each other, and (ii) enable a more relevant comparison between the gestures of various individuals, hereby allowing a more relevant modeling of expertise.

5. Fifthly, an evaluation model must be developed from these features, allowing the prediction of the expertise level on a new gesture performed by a new user. This step has been widely explored in the present work, and two key contributions are proposed:

    - A simple, efficient and generic evaluation model is presented. It is based on the computation of basic statistics on motion features (means and standard deviations), Principal Component Analysis (PCA) and regression. It is tested with various configurations, including various feature types, different regression models and different gesture classes. Tested on the Taijiquan dataset, the proposed model outperforms two methods of the recent literature. On this dataset, the best prediction accuracy ($R = 0.909$) was obtained using a combination of global joint positions and relational features (Müller and Röder, 2006), 60 Principal Components (PCs) extracted on their statistics and L2-regularized linear regression.
    - Additionally, a first exploration of the use of deep learning for the evaluation of the expertise is proposed. The method is inspired by Laraba et al. (2017), representing MoCap data as abstract images, allowing their use with pre-trained deep-learning models for image classification. These

models are adapted using transfer learning for the regression of the level of expertise. Though the prediction accuracy is lower than with the previously presented method ($R = 0.518$), an analysis of the results suggests that the model could achieve better performance given a larger dataset, including a larger number of novices and experts.

6. Sixthly and finally, to allow a practical use of the evaluation model for learning, a feedback system must provide an intuitive interpretation of the predicted level, allowing an effective understanding and assimilation by the user of the system. In the present work, an original and generic feedback system is proposed. The method is based on the synthesis of gestures corresponding to a given level of expertise, higher than the user's level. These synthesized gestures are compared with the user's performance, allowing various types of visual feedback to the user: (i) a synchronized visualization of both gestures, (ii) a striped image representing the motion features that need improvement, and (iii) a striped image displaying the wrongly placed body joints. This feedback system can be used with any expertise evaluation model as long as it provides a continuous score. The resulting feedback is intuitive, and can be used either by a learner for automated supervision, or by a teacher as a tool for objective supervision.

Thesis overview

Fig 1 illustrates the structure of the present thesis, showing the correspondence with the proposed framework. This structure is divided into three main parts:

- Part I presents the global background for this research, and is distributed into three chapters:
  - Chapter 1 briefly explores and discusses different definitions of the concept of expertise.
  - Chapter 2 presents different types of motion representation (i.e. motion features). These representations can be divided into low-level features (see Section 2.2), directly representing motion in terms of positions and orientations, and high-level features (see Section 2.3), allowing the abstract representation of various aspects of motion, such as semantics, expressivity or ergonomics.
  - Chapter 3 then outlines the previous works concerning the evaluation of expert gestures. The various proposed methods are classified into three categories according to the type of score provided to represent the expertise: a score can be directly derived from a specific feature designed

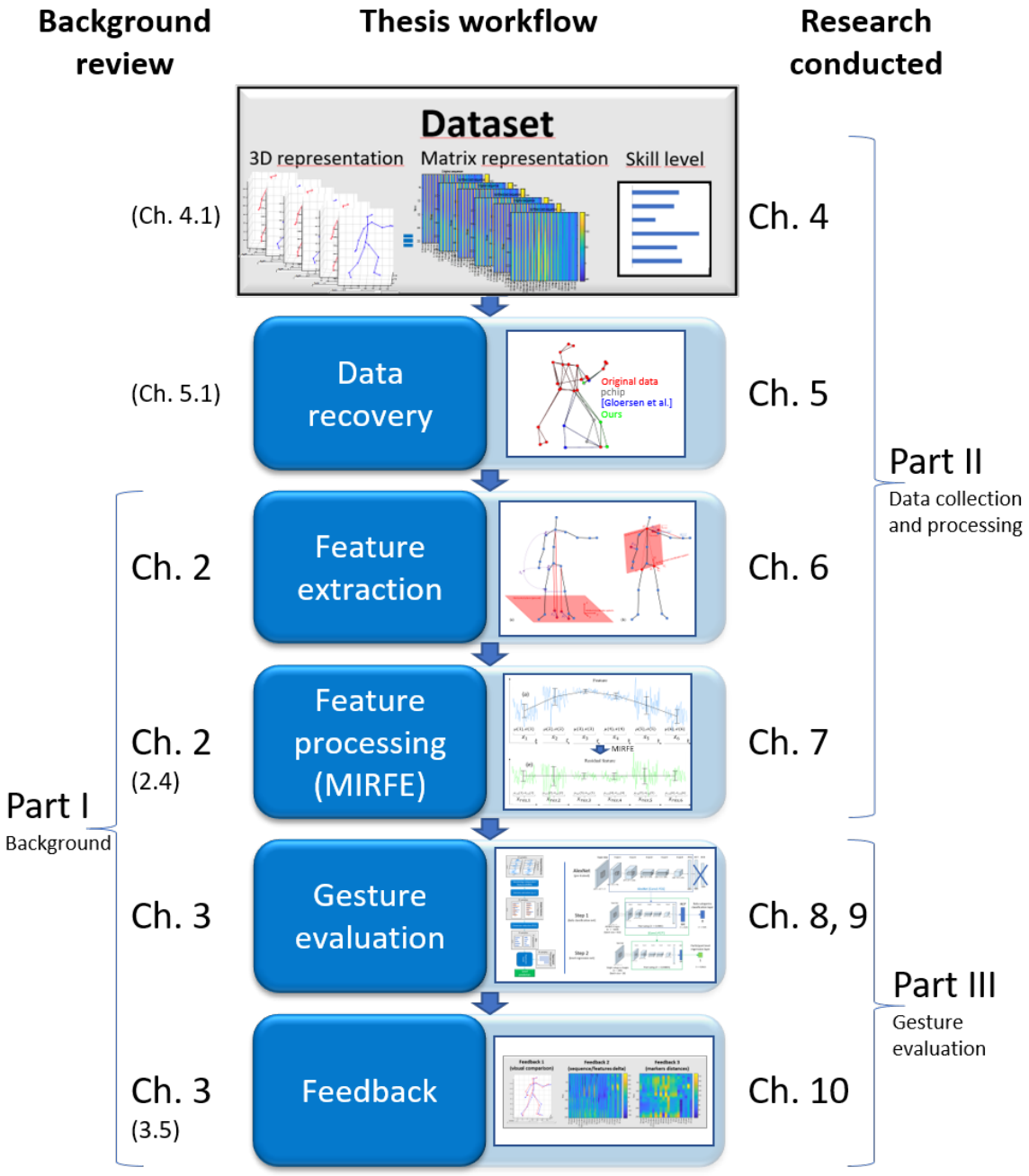

Figure 1: Workflow of the proposed framework, and correspondence with the manuscript structure.

to represent a component of expertise (such as coordination, stability, or complexity indices, see Section 3.4.1) ; a similarity measure can be computed between a learner's gesture and a model of the ideal gesture (see Section 3.4.2) ; or a score can be predicted using a classification or a regression model (see Section 3.4.3). Additionally, the few works proposing methods for feedback are presented in Section 3.4.4.

- Part II presents the proposed contributions of the present thesis concerning the collection and processing of MoCap data. These processing steps aim at a more relevant representation of the gesture, containing maximum information about its expertise level:
  - Chapter 4 presents a new dataset of Taijiquan gestures, including 12 participants of different levels of expertise, from novice to expert, and 13 classes of Taijiquan techniques. This dataset is used as a benchmark for the methods proposed in this thesis.[2]
  - Chapter 5 presents and discusses a method for robust and automatic recovery of MoCap data, based on soft skeleton constraints and model averaging.[3]
  - Chapter 6 presents a new set of motion features inspired by Taijiquan ergonomic principles. These features can be divided into stability features (see Section 6.2), joint alignments (see Section 6.3), favorable angles (see Section 6.4) and fluidity features (see Section 6.5).
  - Chapter 7 presents and discusses an original method for the extraction of morphology-independent motion features, based on the extraction of residuals of a regression model predicting a morphological factor from the original features (Morphology-Independent Residual Feature Extraction, MIRFE). The method is validated on the Taijiquan MoCap dataset.[4]
- In Part III, different methods are proposed for the evaluation of expertise, and for feedback on the learning of an expert gesture:
  - Chapter 8 presents a generic evaluation model, based on feature statistics and classical machine learning. The model is based on PCA and regression of the expertise level from statistics computed on motion features. The proposed approach is tested on the Taijiquan MoCap dataset with various types of motion features presented in the previous chapters (see Section 8.3.1). The use of MIRFE is validated with the proposed method (see Section 8.3.2). Various regression models are then tested (see Section 8.3.3), and are compared with methods of the recent literature (see Section 8.3.4).
  - In Chapter 9, an exploration of the use of deep learning for gesture evaluation is proposed. To that end, a method proposed by Laraba et al. (2017),

[2]This chapter is partly reproduced from Tits et al. (2018a).
[3]This chapter is partly reproduced from Tits et al. (2018b).
[4]This chapter is partly reproduced from Tits et al. (2017).

allowing representation of MoCap data as abstract images, is adapted for the regression of the level of expertise. The method is based on a double transfer-learning step: a classification model is first trained for different Taijiquan classes (see Section 9.3.1), and a regression model is then trained for the prediction of the level of expertise (see Section 9.3.2).

– Finally, Chapter 10 presents an original and generic feedback system based on the synthesis of a feedback gesture corresponding to a particular level of expertise. A gesture performed by a user of the system is first evaluated through an evaluation model. It is then compared with a feedback gesture corresponding to an improved level of expertise, allowing a highlighting of the motion features that need improvement for the user to reach a better level of expertise. The synthesis system is quantitatively validated through a re-evaluation with the evaluation model (see Section 10.3.1). A qualitative validation is then proposed through various examples of the use of the system (see Section 10.3.2).

Conclusions are then provided at the end of the manuscript.

Part I

Background

Chapter 1

What is expertise ?

Expertise: "*Expertise is special skill or knowledge that is acquired by training, study, or practice.*" (Collins)[1]

Expert: "*having, involving, or displaying special skill or knowledge derived from training or experience.*" (Merriam-Webster)[2]

Skill: "*The ability to do something well; expertise.*" (Oxford)[3]

Though it is not the goal of this thesis to dissert on the definition of expertise, it is relevant to provide a brief picture of the research conducted on the subject. 'Gestural expertise' is a complex concept, involving different physiological and psychological components. Studies on this subject found in the literature explore very different aspects of the question, showing the multidisciplinary nature of this research. According to the context or field of research, gestural expertise overlaps with different keywords: "skilled performance", "motor skills", "motor control", "efficiency", "dexterity". An important component of the definition of expertise is that it can only be acquired through the training of a skill, i.e. through experience.

Expertise is a concept that has already been explored in ancient Chinese philosophy. A famous excerpt from a traditional text called the *Zhuangzi*, pillar of the Taoist philosophy and written more than two thousand years ago, draws the encounter of a prince with a dexterous butcher (Tzu, 1964):

Cook Ting was cutting up an ox for Lord Wen-hui. As every touch of his hand, every heave of his shoulder, every move of his feet, every thrust of his knee — zip! zoop! He slithered

[1]Expertise (Collins), retrieved on 20/07/2018: https://www.collinsdictionary.com/dictionary/english/expertise

[2]Expert (Merriam-Webster), retrieved on 20/07/2018: https://www.merriam-webster.com/dictionary/expert

[3]Skill (Oxford), retrieved on 20/07/2018: https://en.oxforddictionaries.com/definition/skill

the knife along with a zing, and all was in perfect rhythm, as though he were performing the dance of the Mulberry Grove or keeping time to the Ching-shou music.

"Ah, this is marvelous!" said Lord Wen-hui. "Imagine skill reaching such heights!"

Cook Ting laid down his knife and replied, "What I care about is the Way, which goes beyond skill. When I first began cutting up oxen, all I could see was the ox itself. After three years I no longer saw the whole ox. And now — now I go at it by spirit and don't look with my eyes. Perception and understanding have come to a stop and spirit moves where it wants. I go along with the natural makeup, strike in the big hollows, guide the knife through the big openings, and following things as they are. So I never touch the smallest ligament or tendon, much less a main joint.

"A good cook changes his knife once a year — because he cuts. A mediocre cook changes his knife once a month — because he hacks. I've had this knife of mine for nineteen years and I've cut up thousands of oxen with it, and yet the blade is as good as though it had just come from the grindstone. There are spaces between the joints, and the blade of the knife has really no thickness. If you insert what has no thickness into such spaces, then there's plenty of room — more than enough for the blade to play about it. That's why after nineteen years the blade of my knife is still as good as when it first came from the grindstone.

"However, whenever I come to a complicated place, I size up the difficulties, tell myself to watch out and be careful, keep my eyes on what I'm doing, work very slowly, and move the knife with the greatest subtlety, until — flop! the whole thing comes apart like a clod of earth crumbling to the ground. I stand there holding the knife and look all around me, completely satisfied and reluctant to move on, and then I wipe off the knife and put it away."

"Excellent!" said Lord Wen-hui. "I have heard the words of Cook Ting and learned how to care for life!"

This excerpt enlightens the idea of a particular knowledge acquired through an extensive learning process, and through different steps. A common butcher hacks, a good butcher cuts, and the expert has integrated the whole task and can carve an ox with eyes closed, and with a specific state of mind acquired through a systematic training.

From this Chinese philosophy, Taijiquan was developed. Taijiquan is a Chinese martial art, but it can be considered more broadly as an art of body awareness. It conciliates three components to define gestural expertise: the body external mechanics, the internal feeling or the mental image, and a spiritual aspect related to concepts of flow, trance and meditation. These three aspects may be seen as three major steps during the training for mastering any gestural discipline (Caulier, 2010, 2014, 2015).

In the field of cognitive science, the process of learning an activity and expertise have been studied (Ericsson and Lehmann, 1996; HAUW, 2009; FLEURANCE, 2009). According to some authors, gestural expertise refers to a mental image of the gesture,

which then becomes finer and more stable through training (Cadopi, 2005; HAUW, 2009). Others propose that expertise depends mainly on the ability to adapt to circumstances when performing a gesture. These faculties of adaptation themselves depend on the training of the gesture in different contexts (King and Yeadon, 2003; HAUW, 2009).

In more practical terms, psychologists and physiotherapists defined the motor skill, and divided it down into several concepts. First, they divided it into two main categories: the gross motor skills and the fine motor skills, according to whether gestures are global movements of the body, or precise movements involving a specific group of body muscles. To evaluate the motor skill, practical tests for fine and gross motor skills were developed. These tests were then divided down into different subtests to evaluate the various components of the motor skills: precision, integration, manual dexterity, uni- and bilateral coordination, balance, speed, agility, strength (Deitz et al., 2007; Cools et al., 2009).

In the field of neuroscience, the process of learning has been studied at the brain level, and allowed the discovery of brain plasticity. This mechanism allows a restructuring of synapses in the brain through motor training, to optimize the motor control, and thus to produce more efficient and economic gestures (Kami et al., 1995).

In the context of ergonomics, the quality of a gesture is assessed over the optimization of coordinated movements of all the parts of the body in the production of the gesture, to minimize body stress. This biomechanical optimization leads to energy savings, but also reduces the risk of injury (Andreoni et al., 2009; Multon and Olivier, 2013; Multon, 2013).

In the context of the arts, and more precisely in dance, Laban Movement Analysis (LMA), from the name of the choreographer Rudolf Laban, allows the analysis of the quality of dance motion from an intentional and aesthetic point of view. Laban defined four main descriptions of motion (Aristidou and Chrysanthou, 2014):

- Body: description of the physical and structural characteristics of the body (positions and orientations);
- Effort or dynamics: description of the intention and dynamic characteristics of the motion;
- Form: description of the overall shape of the body and its aesthetic appearance (volume, height, etc.);
- Space: description of the relationship between motion and the environment.

Regardless of the research domain, whether in cognitive science, physiology or arts, the description of the quality of a gesture is divided into multiple components, which are more relevant if used in a complimentary manner.

For an expert, it is difficult to define what expertise is with words. The expert simply knows and feels it from experience, just as the Taoist butcher. In this research, the goal is not a precise definition of expertise. The aim is instead to model from motion data an expert's perception of expertise. The concept of expertise is then not described with words, but with an algorithm. In this research, we will computationally represent expertise. To that end, we endorse this aspect of multiplicity and complexity of its definition. In Chapter 2, we present techniques that allow a quantitative representation of expertise from different aspects (termed as *features* below), including coordination, stability, energy, accuracy, etc., and thus leading to a more robust description of expertise. In Chapter 3, we then present algorithms used to model the expert's perception of expertise from these various features.

Chapter 2

Motion Capture and Representations

Contents

2.1 Introduction

MoCap appeared together with the development of instantaneous photography and cinematography at the end of the 19th century, making movement recording possible. Before the rise of this technology, it was difficult to perceive and measure complex movements. At that time, Etienne-Jules Marey invented *cyclography*, the ancestor of recent optical motion capture methods. According to that technique, a patient making a periodic movement, and wearing a black suit with narrow white tapes placed along each limb of her/his body, is captured with several photographic exposures on

a single plate. This produces overlapping pictures allowing direct representation of motion on a single image, and accurate analysis thanks to the white tape. Cyclography was further developed into *kymocyclography* by Nikolai Bernstein in 1927, using electric bulbs as markers instead of tape, and a slowly and evenly moving photographic film instead of a single plate. The bulbs captured on the photographic film drew wavelike curves, easy to decipher (Whiting, 1983).

The next step in the development of optical motion capture technology was stereoscopic recording of movements, allowing recording of an object with three spatial coordinates. This was achieved by recording the same scene from different points of observation. Details on the history of MoCap systems can be found in Whiting (1983), Kay et al. (2003) and Metcalf et al. (2014).

During the last three decades, different MoCap systems have been developed to allow accurate 3D measures. These systems can be divided into two main categories: intrusive and non-intrusive systems. Intrusive systems use elements fixed on the object to be captured, like an exoskeleton, inertial measurement units, magnetic systems, or optical markers. Non-intrusive systems do not need the placement of intrusive elements on the target. These systems, based on cameras, have a significant advantage as they are not intrusive, and hence allow freer target movements. For instance, the Microsoft Kinect[1] is a single camera allowing the extraction of a 3D map from a single 3D depth infrared sensor. OpenStage[2] is a multi-camera system, using shape-from-silhouette construction to extract a visual hull of a body. However, the accuracy of these markerless systems is still below that of intrusive systems (Brooks and Czarowicz, 2012; Mundermann et al., 2005). These systems are therefore more adapted to less demanding applications.

In the present work, we used a state-of-the-art MoCap system using passive optical markers manufactured by Qualisys [3]. This system was chosen for its accuracy (< 1 mm) and capability for recording motion at a fixed frame rate up to 400 fps.

This system allows measuring 3D positions of each marker placed on the body at a fixed frame rate. In the remaining of this thesis, a MoCap recording will be referred to as a *motion sequence*. A motion sequence can be considered as a matrix representing all the trajectories of all the recorded markers during the entire sequence. This matrix has the dimension $N \times (3 \cdot M)$, where N is the number of frames of the sequence, and M is the number of recorded markers. A marker trajectory p_j $(j \in 1, ..., M)$ will be represented as an $N \times 3$ matrix.

MoCap data can be represented in various manners, with different advantages and drawbacks. Basic representations (referred to as *low-level features* below) are directly derived from recording systems and include various types of positions and orientations. From these low-level features, higher-level representations may be extracted

[1]Kinect: http://www.microsoft.com/en-us/kinectforwindows/
[2]OpenStage: http://www.organicmotion.com/open-stage-2dot4-release/
[3]Qualisys: www.qualisys.com

(referred to as *high-level features* below). These features may represent various specific aspects of motion such as kinetics, ergonomics, expressiveness and intentions. This Chapter is not intended to present the vast ensemble of motion features. It is rather focused on some major types of motion features that were used for the modeling of expertise in previous works and in the present thesis. Section 2.2 presents different types of low-level features, including positions and orientations (2.2.1), their representation in global or local coordinate systems (2.2.2), as well as some basic preprocessing steps for their effective use (2.2.3). Different types of high-level features are then presented in Section 2.3, including kinematics and kinetics (2.3.2), relational features (2.3.3), expressive features (2.3.4), mathematical decomposition (2.3.5) and ergonomic features (2.3.6). A few studies on the influence of individual factors on these features are then briefly presented in Section 2.4. More specifically, morphology is an individual factor having a direct influence on motion features, making difficult the analysis of gestures performed by several individuals. The few works focused on morphology-independent features will be briefly presented. Finally, a summary of the presented features is provided in Section 2.5, as well as a discussion on their advantages and drawbacks.

2.2 Motion low-level representations

2.2.1 Positions and orientations

A motion sequence can be represented using only **3D coordinates** (positions). These positions can be related to markers placed on the body surface. However, it is more convenient to represent motion using landmarks corresponding to the centers of the body joints (e.g., shoulders, elbows, wrist, etc.). This representation allows to simply describe the movement of all body kinematic chains. Fig 2.1 shows an example of joint representation of the body. From positions of surface markers, a skeleton is reconstructed (using a biomechanical software such as Visual3D^{TM}[4]. A few centers can then be used as landmarks, describing the trajectories of the main kinematic chains of the body (both legs, both arms, and spine). Nonetheless, this representation is incomplete, as a limb can also rotate without changing joint positions (e.g., pronation/supination of the forearm). It is hence convenient to add orientation information to the motion sequence representation. This orientation can be expressed using for instance Euler angles, a rotation matrix or a quaternion, among other representations.

Euler angles represent a 3D rotation by three successive rotations around an axis (see Fig 2.2). It hence requires three parameters (angles ψ, ϑ and φ). A **rotation matrix** is a 3×3 matrix where each column represents each new axis coordinates (Ox', Oy' and Oz' see Fig 2.2) in the original system Oxyz. This representation hence requires

[4]Visual3D^{TM}: http://www2.c-motion.com/products/visual3d/

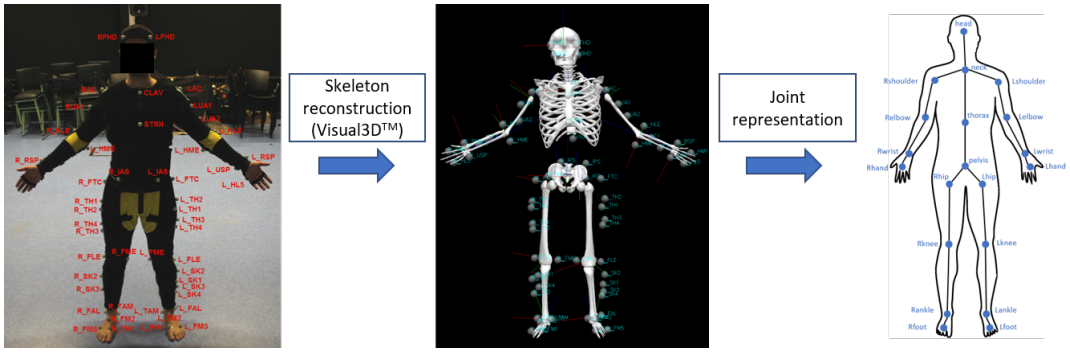

Figure 2.1: Joint representation of the body.

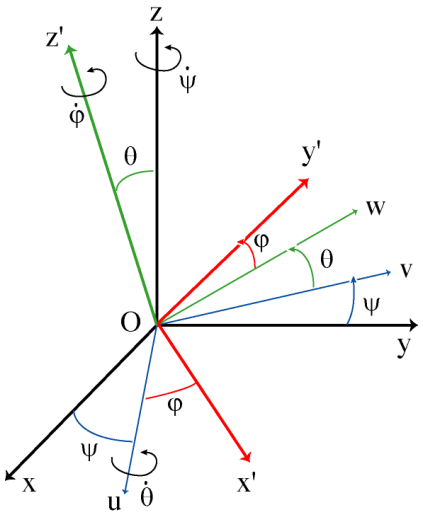

Figure 2.2: Rotation using Euler angles (ψ, ϑ and φ). The original system is in black (Oxyz), the first rotation in blue (ψ around z), the second rotation in green (ϑ around u), the third rotation in red (φ around z'). The rotated system is Ox'y'z'. (Source: Wikipedia)

nine parameters. This representation is useful for linear algebra as a rotation can be performed by a matrix product. It can easily be shown that the parameters of the rotation matrix can be obtained with the Euler angles.

An advantage of the Euler representation its compactness. However, it suffers from a major drawback, as the same rotation can be represented with several angles combinations, leading to discontinuities in the representation of a motion sequence.

Quaternions, introduced by Hamilton (1866), are a generalization of complex numbers, and allow another compact representation of a 3D rotation. A quaternion is composed of a scalar value, and a 3D imaginary part (i.e. 3 hypercomplex values). An advantage of quaternions is that they uniquely represent any 3D rotation, without discontinuities. More information on 3D rotation representations can be found in Tilmanne (2013).

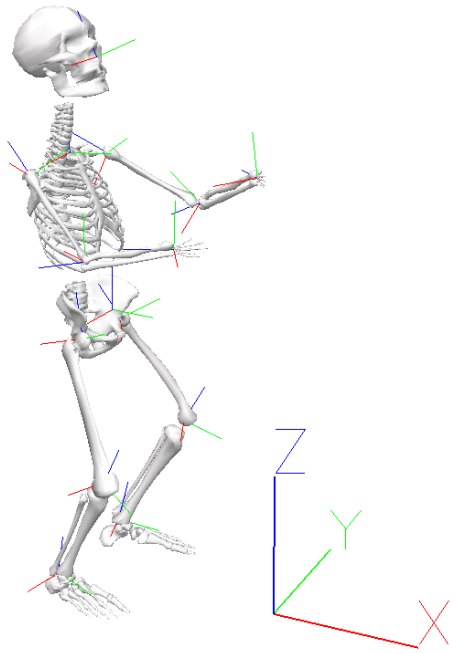

Figure 2.3: Local coordinate systems in Visual3D^{TM}.

2.2.2 Global and local coordinate system

Positions and orientations are always expressed according to a coordinate system. By default, all Qualisys data are represented according to a global coordinate system generally placed on the ground (defined as the origin) with a vertical z-axis. The position of each marker is thus expressed according to the same coordinate system, and is called global position. When representing a motion sequence with joints center positions, it can be relevant to consider a coordinate system located on a parent joint. An example of joints local coordinate systems is shown in Fig 2.3. For instance, the position and orientation of the elbow can be defined according to a coordinate system placed on the shoulder and oriented according to the upper arm. A property of this representation is that a movement of a joint according to any of its DOF will not modify the local position of all the children joints. This allows a reduction of the redundancy of the representation of joint positions.

2.2.3 Modified low-level representation

A major difficulty with dealing with this type of representation, referred to as *low-level features* below, is that two similar motions may be represented with highly different features. For instance, two people facing each other and imitating each other's motion will have very different global positions. Their local positions will also differ, at least because of their different sizes. To deal with these issues, it is common

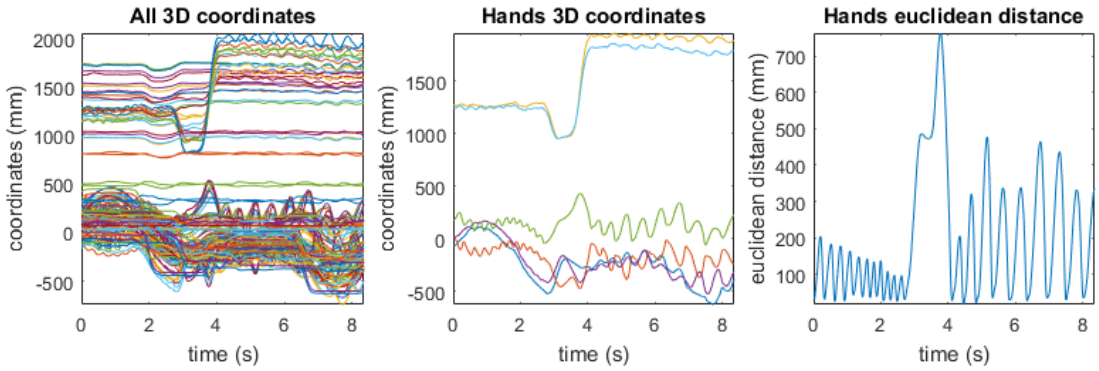

Figure 2.4: Applauding performance analysis example. Left: motion sequence raw 3D coordinates. Center: both hands raw 3D coordinates. Right: hands Euclidean distance.

practice to use a **rigid transformation** (i.e. a **translation** and a **rotation**) of the global coordinate system to align it with specific joints. For instance, the origin $(0, 0, 0)$ could be placed between both feet, or at the pelvis location. Another common practice is the **scaling** of the positions according to people size, in order to reduce the influence of morphology on the features. The effect of individual factors such as morphology is further discussed in Section 2.4.

2.3 Motion high-level representations

2.3.1 Introduction

To the human eye, the signal derived from these low-level representations is complicated to interpret. If we imagine a task where a machine should predict if a person is clapping and at which frequency from this complex signal, it would be challenging. Instead, we can use prior knowledge and design a higher-level representation of the movement from this signal. We can extract 3D coordinates of both hands centers, and compute their Euclidean distance. Fig 2.4 displays this computation on a clapping motion sequence. Using this higher-level representation, referred to below a *high-level feature*, it is easier, even to the human eye, to determine if the person is clapping, and it is also easier to determine the clapping frequency.

As explained in Chapter 1, motion has been studied in many different research areas. According to the context, many different aspects of motion may be studied. Different types of higher-level representations of motion may hence be extracted. For instance, in dance, LMA is often used to analyze movement. In this frame, motion may be represented in terms of general body shape or interaction with the scene. For instance, the bounding box of the body may represent the shape, and the covered area during

a performance may represent the interaction with the scene. These motion features are based on specific prior knowledge of the domain. They are used to indirectly represent the intentions or functions of motion. These higher-level representations are easier to interpret and make motion analysis easier.

In the following sections, different types of high-level features commonly used in the literature will be presented, including kinematic and kinetic features (2.3.2), relational features (2.3.3), LMA-based features (2.3.4), mathematic decomposition of motion (2.3.5) and ergonomic features (2.3.6).

2.3.2 Kinematic and kinetic features

Kinematics and kinetics are two branches of classical mechanics. Kinematics describes the geometrical aspects of motion, and kinetic is the study of the relations between the motion and its causes, i.e. mass and the forces applied to it.

Kinematic features such as limbs and joints velocities, accelerations and jerks can easily be derived from positions and orientations. Kinetic features require an estimation of the body mass distribution. Most of the methods found in the literature to extract kinetic cues from MoCap are based on inertia parameter tables. Zatsiorsky (1990) measured relative body segments masses, and center of mass positions on a sample of college-aged Caucasian males and females, using a gamma-ray scanning technique. These measurements were then adjusted by De Leva (1996). These measurements allow an estimation of the mass distribution of the body from its size and total mass.

To compute the body kinetic energy, the body center of mass must be extracted. The center of mass (CoM) of the body corresponds to the weighted average position of all the points of this body. A single point at this position with a mass corresponding to the total mass of the body has the same inertial properties as the entire body.

The **CoM of the body** can be computed as the weighted sum of the CoM positions of each body segment:

$$\overrightarrow{CoM_{body}} = \frac{\sum_{j=1}^{N} W_j \cdot \overrightarrow{CoM_j}}{\sum_{j=1}^{N} W_j}, \tag{2.1}$$

where N is the number of body segments considered in the sum, W_j is the mass of the body segment j and $\overrightarrow{CoM_j}$ is its center of mass, estimated from inertia parameter tables (De Leva, 1996).

Finally, the **kinetic energy of the body** can be expressed as:

$$E = \frac{1}{2} m \cdot \|\overrightarrow{v_{CoM}}\|^2, \tag{2.2}$$

where m is the total mass of the body, and $\overrightarrow{v_{CoM}}$ is the velocity of the body CoM.

2.3.3 Relational features - Müller

Motion can be described in terms of kinematic relations between different body joints, segments of limbs. These features are generally designed based on prior knowledge of human motion, such as limbs functionality and motion semantic. An advantage of these features is that they are generally independent of the coordinate system. The Euclidean distance of both hands is an example of relational feature. It is not the aim of this Section to present the vast (if not infinite) ensemble of relational features, which can be intuitively designed by the researcher for any specific case, such as the stride length or width for walking pattern analysis, or the distance between hands for "clapping analysis". However, a general framework based on relational features to semantically describe the movement of the full body in any context can be relevant.

In this respect, Müller et al. (2005) proposed a set of 39 binary relational features describing the relations between different body joints. Originally, this set of features was used for motion retrieval, and used with Dynamic Time Warping (DTW, Vintsyuk, 1968). However, these features were also used for various tasks, such as classification (Müller and Röder, 2006), segmentation (Müller and Röder, 2008), annotation (Müller et al., 2009), and gesture evaluation (Laraba and Tilmanne, 2016). Fig 2.5 summarizes the features proposed by Müller et al. (2005).

As indicated in this table, these features are based on six different relation types:

- F_{angle}: the angle between two segments defined by their joints (j_1 to j_2 and j_3 to j_4);
- F_{fast}: the normal speed of a joint (j_1). Both hands, both feet and the root speeds are computed;
- F_{plane}: the distance a joint (j_4 in the table) to a plane passing through three joints (j_1, j_2 and j_3);
- F_{nplane}: the distance between a joint (j_4) and a plane passing through the joint $j_3$and normal to the segment from $j_1$to j_2;
- F_{move}: the speed of a joint (j_4) in reference to another joint (j_3) in the direction of the segment $j_1 \rightarrow j_2$;
- F_{nmove}: the speed of a joint (j_4) in reference to another joint (j_1) in a given direction (perpendicular to a plane passing through three joints j_1, j_2 and j_3).

For instance, the first feature (of the type F_{nmove}) computes the forward speed of the right wrist, i.e. in a direction perpendicular to the torso, defined as a plane passing through both hips and the neck; feature 7 (F_{angle}) computes the elbow flexion angle, and feature 39 (F_{fast}) computes the normal speed of the root. From these

relations, binary features are extracted to semantically describe the movement of the body. For instance, the first feature indicates if the right wrist is moving forward, feature 7 indicates if the elbow is bent, and feature 39 indicates if the root is moving fast. Thresholds for these binary decisions are defined by the parameters θ_1 and θ_2 in the table, and used with a Schimtt trigger (Schmitt, 1938). These thresholds were defined relatively to different parameters of the body morphology (humerus length, hip width and shoulder width, see Fig 2.5), in order to make them invariant to morphology.

Baak (2013) proposed a fortieth feature (F_{40}) for computing the angular velocity of the root orientation. This feature was originally designed to discriminate between turning and non-turning full-body motions.

2.3.4 Expressive features - Laban

Rudolf Laban is a Hungarian choreographer (15 December 1879 – 1 July 1958). He is famous for his proposition of a dance notation system known today as Labanotation. He is also famous for the development of a method for the analysis of movement qualities, divided in four main categories:

- **Body** (body): physical and structural body characteristics (movements and orientations)
- **Effort** or dynamics: description of the intention and dynamic characteristics of the movement
- **Form**: description of the overall body shape and its aesthetic appearance (volume, height, etc.)
- **Space**: description of the relationship between movement and the environment

From his theory, many researchers developed algorithms extracting these qualities. Aristidou et al. (2015) proposed a framework based on 27 hand-crafted features for evaluation of folk dance based on Laban Movement Analysis (LMA). The body component was described by eight cues similar to Müller's relational features described in Section 2.3.3, such as distances between hands, feet, hips and the head, and the height of the pelvis. The effort component is mainly represented by kinematic features including velocities, accelerations and jerks of the root and the body end-effectors (hands and feet). To describe the intentional effort, they compute the head orientation in reference to the body movement direction. The shape component is represented by bounding volumes, as well as the torso height and hands level. These cues are also similar to Müller's relational features. Finally, the space component is described by the distance and the area covered, over a period, by the projection of

ID	Set	Type	j_1	j_2	j_3	j_4	θ_1	θ_2	Description
F_1 / F_2	u	F_{nmove}	neck	rhip	lhip	rwrist	$1.8\,\text{hl}\,\text{s}^{-1}$	$1.3\,\text{hl}\,\text{s}^{-1}$	rhand moving forwards
F_3 / F_4	u	F_{nplane}	chest	neck	neck	rwrist	0.2 hl	0 hl	rhand above neck
F_5 / F_6	u	F_{move}	belly	chest	chest	rwrist	$1.8\,\text{hl}\,\text{s}^{-1}$	$1.3\,\text{hl}\,\text{s}^{-1}$	rhand moving upwards
F_7 / F_8	u	F_{angle}	relbow	rshoulder	relbow	rwrist	$[0°, 110°]$	$[0°, 120°]$	relbow bent
F_9	u	F_{nplane}	lshoulder	rshoulder	lwrist	rwrist	2.5 sw	2 sw	hands far apart, sideways
F_{10}	u	F_{move}	lwrist	rwrist	rwrist	lwrist	$1.4\,\text{hl}\,\text{s}^{-1}$	$1.2\,\text{hl}\,\text{s}^{-1}$	hands approaching each other
F_{11} / F_{12}	u	F_{move}	rwrist	root	lwrist	root	$1.4\,\text{hl}\,\text{s}^{-1}$	$1.2\,\text{hl}\,\text{s}^{-1}$	rhand moving away from root
F_{13} / F_{14}	u	F_{fast}	rwrist				$2.5\,\text{hl}\,\text{s}^{-1}$	$2\ \text{hl}\,\text{s}^{-1}$	rhand fast
F_{15} / F_{16}	ℓ	F_{plane}	root	lhip	ltoes	rankle	0.38 hl	0 hl	rfoot behind lleg
F_{17} / F_{18}	ℓ	F_{nplane}	$(0,0,0)^\top$	$(0,1,0)^\top$	$(0, Y_{\text{min}}, 0)^\top$	rankle	1.2 hl	1 hl	rfoot raised
F_{19}	ℓ	F_{nplane}	lhip	rhip	lankle	rankle	2.1 hw	1.8 hw	feet far apart, sideways
F_{20} / F_{21}	ℓ	F_{angle}	rknee	rhip	rknee	rankle	$[0°, 110°]$	$[0°, 120°]$	rknee bent
F_{22}	ℓ		Plane Π fixed at lhip, normal rhip→lhip. Test: rankle closer to Π than lankle?						feet crossed over
F_{23}	ℓ		Consider velocity v of rankle relative to lankle in rankle→lankle direction. Test: projection of v onto rhip→lhip line large?						feet moving towards each other, sideways
F_{24}	ℓ		Same as above, but use lankle→rankle instead of rankle→lankle direction.						feet moving apart, sideways
F_{25} / F_{26}	ℓ	F_{fast}	rankle				$2.5\,\text{hl}\,\text{s}^{-1}$	$2\,\text{hl}\,\text{s}^{-1}$	rfoot fast
F_{27} / F_{28}	m	F_{angle}	neck	root	rshoulder	relbow	$[25°, 180°]$	$[20°, 180°]$	rhumerus abducted
F_{29} / F_{30}	m	F_{angle}	neck	root	rhip	rknee	$[50°, 180°]$	$[45°, 180°]$	rfemur abducted
F_{31}	m	F_{plane}	rankle	neck	lankle	root	0.5 hl	0.35 hl	root behind frontal plane
F_{32}	m	F_{angle}	neck	root	$(0,0,0)^\top$	$(0,1,0)^\top$	$[70°, 110°]$	$[60°, 120°]$	spine horizontal
F_{33} / F_{34}	m	F_{nplane}	$(0,0,0)^\top$	$(0,-1,0)^\top$	$(0, Y_{\text{min}}, 0)^\top$	rwrist	-1.2 hl	-1.4 hl	rhand lowered
F_{35} / F_{36}	m		Plane Π through rhip, lhip, neck. Test: rshoulder closer to Π than lshoulder?						shoulders rotated right
F_{37}	m		Test: Y_{min} and Y_{max} close together?						Y-extents of body small
F_{38}	m		Project all joints onto XZ-plane. Test: diameter of projected point set large?						XZ-extents of body large
F_{39}	m	F_{fast}	root				$2.3\,\text{hl}\,\text{s}^{-1}$	$2\,\text{hl}\,\text{s}^{-1}$	root fast

Figure 2.5: Müller's relational features. "hl" = humerus length, "hw" = hip width, "sw" = shoulder width. Reproduced from Müller and Röder (2006).

the root joint on the ground. Though the proposed features are not original, the interest of this framework arises from the variety of types of features that are proposed as an ensemble, to describe all aspects of dance motion as described by LMA. This explains why feature sets derived from LMA have been used for various tasks, including computer animation (Chi et al., 2000; Torresani et al., 2007; Zhao and Badler, 2005), motion segmentation (Bouchard and Badler, 2007), retrieval (Kapadia et al., 2013), indexing (Aristidou and Chrysanthou, 2014), and evaluation (Aristidou et al., 2015).

2.3.5 Mathematical decomposition of motion

2.3.5.1 Introduction

Instead of using prior knowledge on motion to decompose it into various handcrafted features, another approach is the use of mathematical tools allowing extraction of new representations of data. These tools, widely used in signal processing or in machine learning, can be applied onto motion data. Though it is not the aim of this section to review all signal processing and machine learning techniques adapted to motion data, a few well-known algorithms will be briefly presented, including frequency decomposition (2.3.5.2) and eigenmovement decomposition (2.3.5.3). Both of these techniques were used for expertise modeling (Federolf et al., 2012; Zhang et al., 2010).

More details on the use of other mathematical tools on motion data can be found in Samadani et al. (2013), where different statistical dimensionality reduction techniques such as supervised Principal Component Analysis, Isomap and functional Fisher Discriminant Analysis have been tested for motion representation. For the interested reader, the development of motion manifolds based on machine learning, and especially deep learning, is also explored in Holden et al. (2015, 2016).

2.3.5.2 Fourier analysis - Frequency decomposition

MoCap data can be considered as a discrete multidimensional signal. During the last three decades, as research on MoCap was expanding, many signal processing techniques already extensively used in image and speech processing were applied to motion. Bruderlin and Williams (1995), in their article simply entitled "Motion Signal Processing", showed the interest of multiresolution filtering, a technique initially used on images, to decompose motion. The intuition was that low frequencies contained general, gross motion patterns, while high frequencies contained motion details and subtleties, as well as most of the noise.

Unuma et al. (1995) used Fourier analysis to generate human walking animations with emotion. Interpolation in the frequency domain allowed them to synthesize

a smooth transition from walking to running motion, as well as normal walking to brisk walking motion. These works showed that frequencies efficiently encode information about motion style and emotion. However, their interest is limited to periodic motions like walking.

2.3.5.3 Principal Component Analysis - Eigenmovement decomposition

Principal Component Analysis (PCA, Pearson, 1901) is a widespread mathematical tool, allowing to extract, from correlated observational variables, a new set of orthogonal (linearly independent) variables, called principal components (PCs). These PCs are linear combinations of the original variables, obtained in such a way that the first PC has the largest possible variance, i.e. following the axis with the largest variability in the observed data. Each one of the following PCs follows the orthogonal axis with the largest remaining variance. This technique allows to efficiently reduce the size of a set of variables, by keeping only the PCs with the largest variance (i.e. the first PCs).

MoCap data consist of complex and highly multidimensional signals, including 3D coordinates or orientations of various joints. However, due to the skeleton structure, all these variables are highly correlated. PCA is thus an efficient technique to reduce the signal complexity by removing the redundancy of all the variables, and extracting a few PCs encoding most of the information contained in the signal.

PCA has been first used to decompose gait patterns. Troje (2002) used PCA to decompose gait patterns in PCs, following axes called "eigenpostures". As these PCs were similar to sinusoids due to the periodic pattern of walking, he extracted their frequencies and phases, and represented a whole gait pattern using a set of parameters including first PCs' frequencies, phases and axes. From a dataset of these parameters extracted from different participants, he performed a second PCA to extract new PCs called "eigenwalkers". He then used this representation for gender classification, and for walking synthesis.

This technique was then used for various applications, such as person identification (Troje et al., 2005) or synthesis of expressive gait (Tilmanne and Dutoit, 2010). In the context of gesture evaluation, Federolf et al. (2012) proposed the use of eigenmovements to quantify skill in sport. They analyzed alpine skiing as an example. They show that PCs can be visualized as "principal movements" using a back-projection on the original variables. They show that the visualization of these principal movements allows a semantic interpretation. For instance, in their case, the first principal movement represented lateral body inclination, the second one represented flexion-extension of the legs, and the third one represented a rotation of the skis and the upper torso.

PCA is more general than frequency domain analysis, as it is not limited to the motion decomposition into periodic patterns. As this technique was more efficient and

general than frequency analysis, the same idea has then been successfully applied in many disciplines such as competitive diving (Young and Reinkensmeyer, 2014), karate (Zago et al., 2016), handball (Helm et al., 2017), soccer (Abdullah et al., 2017), cross-country skiing (Gløersen et al., 2017), and juggling (Zago et al., 2017).

2.3.6 Ergonomics

2.3.6.1 Introduction

Ergonomic features are related to biomechanical aspects of the movement. They are based on musculoskeletal modeling of the body, aiming to describe the movement quality in terms of comfort, robustness, or load. Ergonomics is closely related to skill, especially in sports disciplines, as it is the study of motor control effectiveness while minimizing energy expenditure and risks of injury. However, few previous work known to the author used ergonomic features for evaluation of expertise. Andreoni et al. (2009) proposed a method based on perceived discomfort (see Section 2.3.6.4) to automatically assess the ergonomy of a posture from MoCap data, showing a potential use of ergonomic features in motion quantitative assessment. More recently, coordination indices were also developed for the evaluation of sports gestures (see Section 2.3.6.6) (Kim et al., 2011; Alborno et al., 2017; Morel et al., 2016).

In this respect, different types of ergonomic features used in various fields of motion analysis, such as medical research and animation, are presented in this Section. Some of these features are used in the present thesis for the development of gesture evaluation models. More particularly, a new set of features has been developed, inspired by ergonomic features, and especially Taijiquan ergonomic principles (see Chapter 6).

The following Sections will present different ergonomic characteristics of motion, including balance (2.3.6.2), the degrees of freedom and ranges of motion of the body (2.3.6.4), postural load or discomfort (2.3.6.4), torques (2.3.6.5) and coordination (2.3.6.6).

2.3.6.2 Balance

Balance is "the stability produced by even distribution of weight on each side of the vertical axis" (Merriam-Webster[5]). Balance is an essential component of motor control, as it allows to move without losing one's stability, and in extreme situations, falling. Balance has been vastly explored, mostly in the area of gait analysis (Bruijn

[5]Balance definition (Merriam-Webster): https://www.merriam-webster.com/dictionary/balance

et al., 2013). Nonetheless, balance has been presented in the literature as an expressive feature as well, or for motion indexing (Tilmanne et al., 2015; Kapadia et al., 2013; Larboulette and Gibet, 2015).

A basic measure of balance is based on the computation of the body CoM (see eq. 2.1), and the support base. The support base is defined as the area between the points of the body that are in contact with the ground. For instance, the support base of a standing person is defined by the area between her/his two feet extremities. Balance can then be derived from the distance between the support base center and the projection of the CoM on the ground (Tilmanne et al., 2015). A larger distance means a lower balance.

However, this basic measure does not take into account the velocity and inertia of the CoM, and the forces implied to keep the CoM above the support base. To assess balance in dynamical situations, Hof et al. (2005) proposed the extrapolated center of mass (XCoM), based on an extension of this model. This extension is inspired by an inverted pendulum system, adding a linear function of the CoM velocity to its position in evaluating its distance with the support base center:

$$XCoM = CoM + \frac{V_{CoM}}{\omega_0}, \tag{2.3}$$

where V_{CoM} is the velocity of the CoM and ω_0 is the eigenfrequency of the inverted pendulum:

$$\omega_0 = \sqrt{\frac{g}{l}}, \tag{2.4}$$

where g is the grativational acceleration and l is the length of the inverted pendulum equivalent to the body. However this model considers the body as a simple pendulum (i.e. a rope and a mass), and not as a kinematic chain.

To take into account the forces implied in keeping the CoM above the support base, Duclos et al. (2009) proposed a method for quantifying the force needed to keep the CoM above the support base (called the stabilizing force), and the destabilizing force moving the CoM outside the support base. However, this method requires a measurement of ground reactions forces (using for instance a force plate).

2.3.6.3 Degrees of freedom and ranges of motion

A **degree of freedom (DOF)** is any independent part of the body motion, i.e. the rotation of a joint. The **range of motion (ROM)** defines the maximal rotation that can be applied along each DOF. For example, the ROM of the DOF corresponding

	Neck	**Shoulder**	**Elbow**	**Wrist**
X+	Flexion	Flexion	Flexion	Flexion
X-	Extension	Extension	-	Extension
Y+	Left bending	Adduction	-	Radial deviation
Y-	Right bending	Abduction	-	Ulnar deviation
Z+	Left rotation	Medial rotation	Supination	-
Z-	Right rotation	Lateral rotation	Pronation	-
	Spine	**Hip**	**Knee**	**Ankle**
X+	Flexion	Flexion	Flexion	Flexion
X-	Extension	Extension	-	Extension
Y+	Left bending	Adduction	-	Adduction
Y-	Right bending	Abduction	-	Abduction
Z+	Left rotation	Internal rotation	-	-
Z-	Right rotation	External rotation	-	-

Table 2.1: List of typical measured ROM.

to the flexion of an elbow usually varies from 140 to 159° according to individuals (NASA, 1995).

ROM are generally used in ergonomics to evaluate suppleness or disability (Boone and Azen, 1979). In motion processing, ROM can also be used as reference to normalize joint motions, to represent movement as a portion of the total possible motion of a joint.

ROM are usually expressed as Euler angle deviations from a reference position, corresponding to a rest posture (standing, lying or sitting posture). In this manner, at most six ROM are given for each joint, i.e. one for each rotation axis (x, y, z) on each way (positive, negative). For instance, the shoulder is the most mobile body joint and is defined by six ROM: flexion and extension, adduction and abduction, medial rotation and lateral rotation. On the opposite, the knee has only one degree of freedom, on one way only, the flexion, and the other five ROM are considered as zero. Table 2.1 lists different measurable ROM of each joint, as proposed in Kee and Karwowski (2003). The axes and signs are arbitrarily chosen in this example.

Normal values for these ROM can be found in Boone and Azen (1979), or in the NASA man-systems integration standards (NASA, 1995)[6], among other sources.

Movement representation through DOF and ROM has the advantage to be anatomically meaningful, as each component corresponds to a degree of freedom of the body. Motion normalization may have an interest when computing relationships between joints in a movement, or comparing motions of various individuals. However, a limitation is due to the fact that joint ROM are not strictly independent variables. The maximal orientation (the ROM) on a DOF can indeed depend on the orientation on

[6]NASA standards : http://msis.jsc.nasa.gov/sections/section03.htm

another DOF. Haering et al. (2014) showed these interactions between DOF for the particular case of the shoulder, and proposed a representation method based on a 3D hull to account for these interactions.

2.3.6.4 Postural load (perceived discomfort)

The **postural load**, or perceived discomfort, or stress, of a joint is an indication of the perceived stressfulness of a joint, according to its orientation, taking into account each degree of freedom of the joint. The postural load of the whole body provides information about the comfort of a posture, and on the risk of injury.

The evaluation of the postural load has mainly applications in industrial ergonomics, to correct workers postures, or to design ergonomic workplaces. However, it can also find an interest in movement performance assessment, by evaluating if the movement is optimized in terms of stress, i.e. minimizing the pain, and risks of injury.

The **joint stress** can be defined as the sum of the perceived discomforts associated to the independent movement of a joint on each degree of freedom. The perceived discomfort on a degree of freedom can be interpolated from a table of perceived discomforts indices in relation with joint angles. Such tables can be found in ergonomics literature (Kee and Karwowski, 2001, 2003; Andreoni et al., 2009). The overall postural load can then be deduced from the joint stresses. Kee and Karwowski (2001) defined a postural load index as the sum of all perceived discomforts for each joint:

$$\begin{aligned} S_j &= \sum_{i=1}^{m_j} S_{ij} \\ Postural\ load &= \sum_{j=1}^{n} S_j \end{aligned} \tag{2.5}$$

where S_j is the stress of the joint j, S_{ij} the part of the stress independently associated to the degree of freedom i for the joint j, and m_j the number of degrees of freedom of the joint j.

Andreoni et al. (2009) proposed a method based on a weighted sum of each joint stress, to evaluate the overall postural load. The weights (W_j) of each joint are proportional to the mass of the distal body district corresponding to each joint (as measured by Zatsiorsky, 1990):

$$Postural\ load = \sum_{j=1}^{n} W_j S_j \tag{2.6}$$

They tested experimentally their method on a reaching task. During the experiment, participants were recorded with a MoCap system while reaching, with one hand, different points regularly placed from the ground to the top of an experimental structure (see Figure 2.6). From the MoCap data, they calculated the postural load for the reaching movement of each point on the structure, and compared the results of their method to the previous method from Kee and Karwowski (2001).

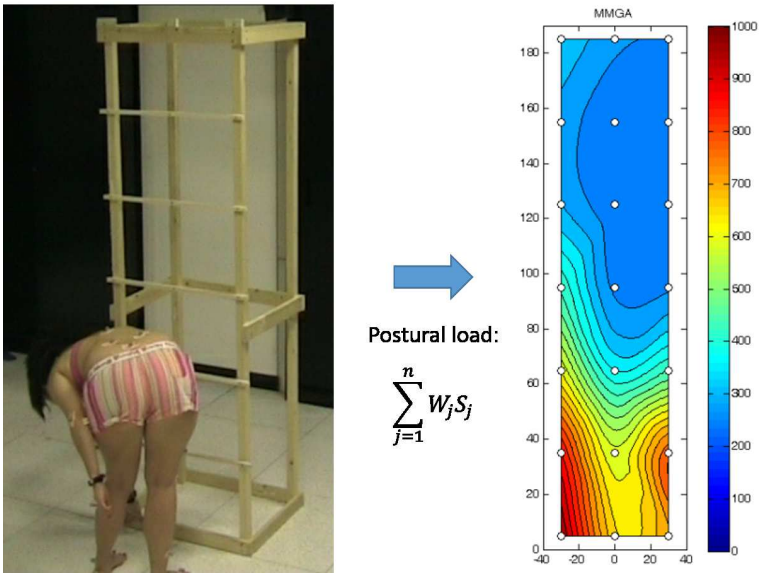

Figure 2.6: Postural load on a reaching task. Reproduced from Andreoni et al. (2009).

The postural load is a feature specifically developed for ergonomics research. Nonetheless, an interest can be found in the assessment of sport or dance performance. In sport, an efficient gesture will minimize the global stress for the same outcome, while distributing the load on each joint according to their robustness. In dance, on the other hand, the global stress may be high, displaying the emotion and intention of the dancer through her/his movements.

The postural load presents several limitations. First, as its name suggests, the feature is based on the posture only, and does not account for the movement dynamic. Moreover, the perceived discomfort tables found in the literature rarely take the gravity into consideration, nor the coupling between ROM and joints.

2.3.6.5 Torques

Torques represent the forces that are responsible for joint rotations. They allow the definition of the dynamic state of each body joint.

Though they could be classified as kinetic features, torques have direct applications in ergonomics, to evaluate muscles exertions and articular loads. They are also used in computer animation, to evaluate the naturalness of a motion (Multon, 2013).

Joint torques cannot be directly measured, but they can be calculated through different "inverse dynamics" methods. One approach is to consider the body as a mechanical system limited to rigid isolated segments (S_i). For the center of mass of each segment, the resultant force and variation of angular momentum can be expressed as follows, considering external forces and torques (F_e^i, T_e^i) such as gravity and ground reaction, and internal forces and torques (F_m^i, T_m^i), due to muscle activity :

$$\begin{cases} m_i \gamma_i = F_e^i + F_m^i \\ \dfrac{dL_i}{dt} = M_{F_e^i} + M_{F_m^i} + T_e^i + T_m^i \end{cases} \tag{2.7}$$

where m_i, γ_i and L_i are the mass, acceleration, and angular momentum of the center of mass of the segment S_i , respectively. As connected segments apply forces and torques on each other, the problem is resolved iteratively, from an extremity of the body to another. Details can be found in Multon (2013).

Joint torques have the advantage of being more objective features than the perceived postural load, as they only depend on physical equations, and not on the subject's feelings. They allow for a precise ergonomic assessment of gestures and postures, and are therefore widely used in the evaluation of the ergonomics of user interfaces Bachynskyi et al. (2015).

However, dynamic motion models generally depend on weight approximations for each body segment, which can vary a lot from a person to another. Moreover, noisy motion data can result in a very inaccurate computation of accelerations. Finally, a specific instrument such as a force plate is needed to efficiently estimate ground reaction forces.

2.3.6.6 Coordination

Coordination is "the harmonious functioning of parts for effective results" (Merriam-Webster[7]). Coordination is the process of synchronizing joints and limbs movements to efficiently perform a gesture. Coordination has been studied in various fields, such as gait analysis (Dejnabadi et al., 2008), music (Furuya et al., 2011, 2015), and sports (Kim et al., 2011; Alborno et al., 2017; Morel et al., 2016).

To assess the **inter-joint coordination** of a leg in a Taekwondo kick, Kim et al. (2011) proposed an index based on local joints angular velocities:

$$IIC = (\overrightarrow{\omega_H} / \left\|\overrightarrow{\omega_H}\right\|) \cdot (\overrightarrow{\omega_K} / \left\|\overrightarrow{\omega_K}\right\|), \tag{2.8}$$

[7]Coordination definition (Merriam-Webster): https://www.merriam-webster.com/dictionary/coordination

where $\overrightarrow{\omega_H}$ and $\overrightarrow{\omega_K}$ are respectively the hip and knee 3D angular velocities. If these vectors are in the same direction, the inter-joint coordination (IIC) is 1. $IIC = -1$ if they are in an opposite direction.

To evaluate **limb synchronization** in karate, Alborno et al. (2017) extracted limbs acceleration peaks using progressive filtering. They then analyzed time delay relationships between the peaks of each limb.

Dejnabadi et al. (2008) proposed a method based on a neural network to model walking coordination at different stride lengths and walking speeds. The neural network was designed to reconstruct a normal walking pattern from two input parameters: stride length and cadence. To evaluate the coordination of a walking motion, they compared it to a reconstructed walking pattern at the same stride length and speed, and they computed a coordination index from joint angles differences.

Morel et al. (2016) proposed a method based on Dynamic Time Warping (DTW) for evaluation of limb synchronization during a sports gesture. From a dataset of gestures performed by experts and aligned with DTW, they compute an average gesture, supposed as a perfectly synchronized gesture. For a new gesture, they perform a temporal alignment of each limb's data separately (called local alignment). They then evaluate the gesture synchronization from the delay between two limbs alignments.

PCA (see Section 2.3.5.3) can be seen as a tool for coordination analysis, where each eigenmovement can be seen as a coordinated part of movement (Daffertshofer et al., 2004).

2.4 Multifactor influence

In a dataset of any type of motion performance, a large number of factors can influence the way a performer moves. Factors are the variables, intra- or inter-individual, which can have an effect on motion. Inter-individual factors may be of various types, including social factors (e.g., culture, education), psychological factors (e.g., personality, emotions, state of concentration), physiological factors (e.g., gender, age, morphology, force, suppleness), or psycho-physiological factors (e.g., handedness, motor skills). On the other hand, intra-individual factors are related to the performance, independently of the performer, such as the type of dance, or the purpose of a particular exercise. Many specific research on these factors has been conducted in different contexts, such as clinical factor analysis on gait patterns (Lord et al., 2012; Verlinden et al., 2013), or investigation of emotion factors effects on body action and posture (Dael et al., 2012), and on kinematics of locomotion (Barliya et al., 2013), to mention a few. All these factors may have different effects on motion features, effects which are generally difficult to analyze separately.

Morphology is a particular factor that has a direct influence on motion, making comparisons between gestures of different individuals difficult. To alleviate this issue, different motion data representations have been proposed. Sie et al. (2014) proposed a simple skeleton scaling method, by placing the coordinate system on a reference node of the body (i.e. on the pelvis), and dividing all nodes coordinates by the torso height. Features can then be extracted on these scaled data. This method was later used by Morel et al. (2016) for evaluation of tennis serve. It has the advantage to be very simple, but has many limitations. It is based on the simplistic hypothesis that the motion of a short individual should be an homothety of those of a tall individual. However, weight, height of the center of mass, shoulder width, and hip width, among others, may also influence motion in different ways, including inertia, balance, speed and power. These characteristics will be altered by this basic scaling.

Kulpa et al. (2005) developed a morphology-invariant representation of motion, originally developed for animation, where they defined limbs with variable lengths. Each limb (legs and arms) is defined by the position of its end-effector, and by a plane where the middle joint (knee and elbow) is located. The spine is represented as a spline. This representation allows reconstruction of the motion to fit specific constraints. However, it does not fully store the actual motion, and it modifies it to fit these constraints. It is relevant for animation and motion retrieval, but is not suited for motion analysis, which can require motion details that are lost in this representation.

Müller et al. (2005) proposed a specific feature set based on 40 logical relational features, as described in Section 2.3.3. The boundaries for the logical decision were defined by different body segment lengths such as the humerus length or the shoulder width, so that each feature is scaled by a custom pre-defined body characteristic. However, the method is limited to the specific proposed feature set, and does not allow extraction of other high-level features.

In Chapter 7, we propose a novel method for removal of the influence of any quantifiable factor on any motion feature. This method has been published in Tits et al. (2017).

2.5 Discussion and conclusion

In this Chapter, we presented various motion representations, classified into two main categories:

- low-level features, including positions and orientations, both in global or local coordinate systems;
- and high-level features, including kinematic, kinetic, relational, expressive, ergonomic and decomposition features, among others.

Among these categories, a continuous scale could have been used. First, kinematic and kinetic features are close to low-level features, as they are directly derived from basic mechanics, without using any a priori knowledge on human motion. Secondly, relational features and eigenmovements aim at a functional description of human motion. Nonetheless, this representation is closer to a direct description of motion than higher-level representations like Laban or coordination features. The latter are indeed based on prior knowledge of human motion, and respectively intend for a description of the motion quality in terms of expressiveness or ergonomics. On this continuous scale, relational features and eigenmovements could be seen as mid-level features, and expressive or some ergonomic features as higher-level features.

High-level representations have the assumed advantage to be easier to interpret than low-level ones, and may be more representative of motion for the analysis of the intention, expressiveness, or expertise in the particular case of ergonomic features.

The drawback of this type of representation is that it does not ensure a full description of the motion. On the opposite, a low-level feature set including all joint positions and orientations intrinsically ensures a full representation of motion (limited to the data recorded by the MoCap system). High-level features are inevitably extracted from low-level features. Therefore, they do not add new information about motion, but rather make it easier to interpret. Consequently, representing motion using a high-level feature set may lead to a loss of information if all low-level features cannot be retrieved from them. For instance, Müller features indicate if hands are far apart sideways (F_9 in Fig 2.5), but the information of the absolute positions of both hands is lost in this representation. This representation is sufficient in the case of clapping detection, and is more relevant than hand positions as a single feature is needed, compared to the six 3D coordinates of both hands. However hand positions would be needed to analyze specific characteristics of the clapping gesture such as its ergonomy, or to discriminate between a flamenco clapping (with both hands on the same side of the body), or basic clapping with hands in front of the body. On the other side, higher-level features like postural loads and torques would ease the analysis of the ergonomy of the gesture, but do not allow the detection of a clapping motion, nor the extraction of other specific characteristics such as the clapping frequency. Low-level representations such as joint positions allow analysis of all these characteristics, although it would require a more complex modeling of the relations between the joint positions and the analyzed characteristics.

In the present thesis, different types of features, including both low-level and high-level features, are tested for the development of gesture evaluation and feedback models (see Part III). In particular, a new feature set inspired by Taijiquan ergonomic principles is presented in Chapter 6.

Finally, all these features may be influenced by various factors, and especially morphology, making difficult comparison between motions of different individuals. Various previous works proposed an adapted data representation to reduce morphology influence (see Section 2.4), and a novel and more general method is proposed in Chapter 7.

Chapter 3

Expert gesture evaluation: a state of the art

Contents

3.1 Introduction

The ability of a computer to interpret human gestures, and particularly to evaluate the quality of an expert gesture, mainly depends on two major technologies that are MoCap and machine learning. MoCap technologies allow the automatic and accurate recording of human motion in 3D. On the other hand, machine learning is a branch of artificial intelligence concerned about the development of mathematical models for the automatic prediction of a variable (such as expertise) from other related variables (such as gestures). In order to understand better the context of the present research, a short history about the developments in both domains is provided. Section 3.2 presents a history of the use of machine learning algorithms for the analysis of human activity. Section 3.3 then provides a history of the development of 3D full-body MoCap technologies, and their use for human motion analysis, with or without machine learning. Section 3.4 then presents the state of the art concerning expert gesture evaluation.

3.2 Machine learning for human activity analysis

During the last decades, a vast number of new technologies has emerged, driven by electronics, mathematics and computer science. One of the most important technologies developed in contemporary computer science is machine learning. More specifically, applied to human activity analysis, machine learning allows a computer to decipher and interpret a human's state or actions.

Machine learning techniques are algorithms that can be applied on any type of data, allowing a computer to automatically learn a mathematical model from these data. These algorithms are usually used to automatically predict a dependent variable (the *outcome*) from various input variables (the *predictors*). The outcome can be a category, such as a word or a person's gender, or a continuous variable, such as age, or the scale of an emotion. The predictors are for instance pixels of an image, an audio signal, bio-signals, or various types of *features* extracted from them using signal processing techniques. The modeling procedure is generally based on the probability theory, or on the optimization of a mathematical function representing the relation between the predictors and the outcome (Dietterich, 2002; Kotsiantis et al., 2007; LeCun et al., 2015).

In the particular field of human activity analysis, these techniques have been vastly explored allowing machines to recognize, synthesize and characterize various types of human actions measured with different sensing and computer vision systems. From audio data, speech recognition (Rabiner and Juang, 1993) and synthesis (Dutoit, 1997) have been extensively studied during the late 20^{th} century, and used in many applications including GPS guiding, text-to-speech, speech-to-text, and more recently virtual assistants. From face images and videos, machine learning techniques allow face recognition (Zhao et al., 2003), synthesis (Blanz and Vetter, 1999), and facial expression recognition (Cohen et al., 2003). More broadly, images have also been used for many tasks such as pose recognition (Bradski and Davis, 2002), sign-language recognition (Starner et al., 1998) and handwriting recognition (Plamondon and Srihari, 2000). Moreover, text data allow semantic analysis and text generation (Sebastiani, 2002). Most of these types of signals are now being used with the latest state-of-the-art machine learning algorithms, i.e. deep learning (LeCun et al., 2015). A fundamental aspect of machine learning, and especially deep learning, is that models can learn extremely complex relations between variables, if a large dataset and enough computational resources are provided.

Finally, another major but still emerging branch of human activity analysis is the analysis of human motion. Unlike speech which is recorded from a single microphone, or faces which are recorded from a single 2D camera, full-body articulated motion is a complex signal, difficult to record with sensors. In the early 2000s, most of the studies on human motion analysis are therefore limited by the type of motion representation, generally based on 2D silhouettes extracted from a single camera

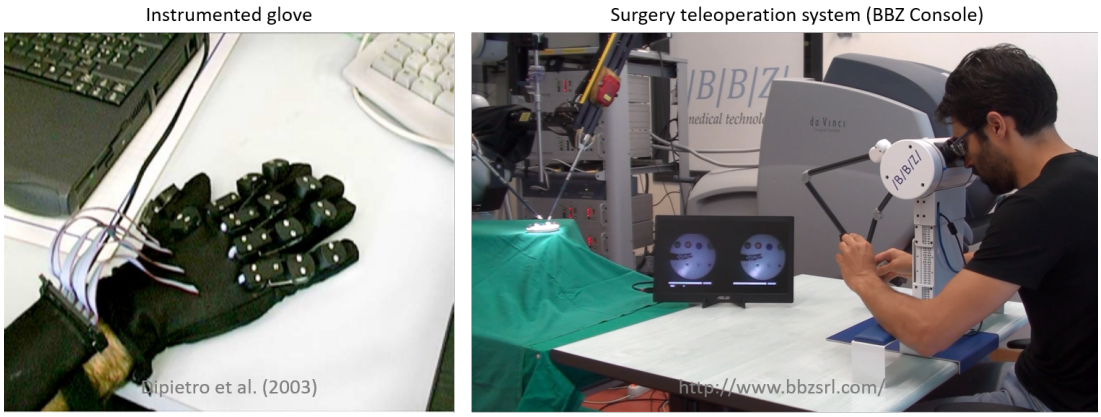

Figure 3.1: The instrumented glove (Dipietro et al., 2003) and a surgery teleoperation system (BBZ Console) are both electronics devices allowing the recording of specific 3D motions.

(Poppe, 2010). Only hand gesture recognition was already explored using specific MoCap sensors such as an instrumented glove (see Fig 3.1, left image) (LaViola, 1999). Other very specific domains have also been explored, such as evaluation of teleoperation and minimally-invasive-surgery gestures, thanks to the specific electronic devices required in these disciplines (see Fig 3.1, right image) (Hannaford and Lee, 1991; Rosen et al., 2001). However, in this case, the study is limited to the motion of the manipulated tools rather than body motion itself.

3.3 3D full-body motion capture

MoCap is the process of recording the positions and/or orientations of any object or body part, through any type of sensor (see examples Fig 3.2). Though the concept appeared by the end of the 19th century through cyclography (Marey, 1873), the first commercial manufacturers of full-body MoCap systems appeared about a century later, such as Polhemus (1969), Motion-Analysis (1982), and VICON (1984). At the same period, their systems began to be used for clinical research (Taylor et al., 1982) and for character animation in video games and movies (Sturman, 1994). However, before the end of the 1990s, the use of MoCap systems was generally limited to 2D unstable recordings, requiring a time-consuming data post-treatment to revise and label all the recorded positions. The development of accurate 3D and automatic MoCap systems (Herda et al., 2000; Kakadiaris and Metaxas, 2000) allowed their effective use for wider applications. As commercial MoCap systems integrated these developments, the first 3D full-body MoCap datasets were recorded and published on the web to promote research development in the domain (CMU, 2003; Ma et al., 2006; Müller et al., 2007). The high cost of these measurement systems[1] may also

[1]The cost of a state-of-the-art 3D MoCap system (accuracy $<$ 25 mm, frame rate $>$ 100 fps) is generally higher than 10.000$ still today (Romero et al., 2017).

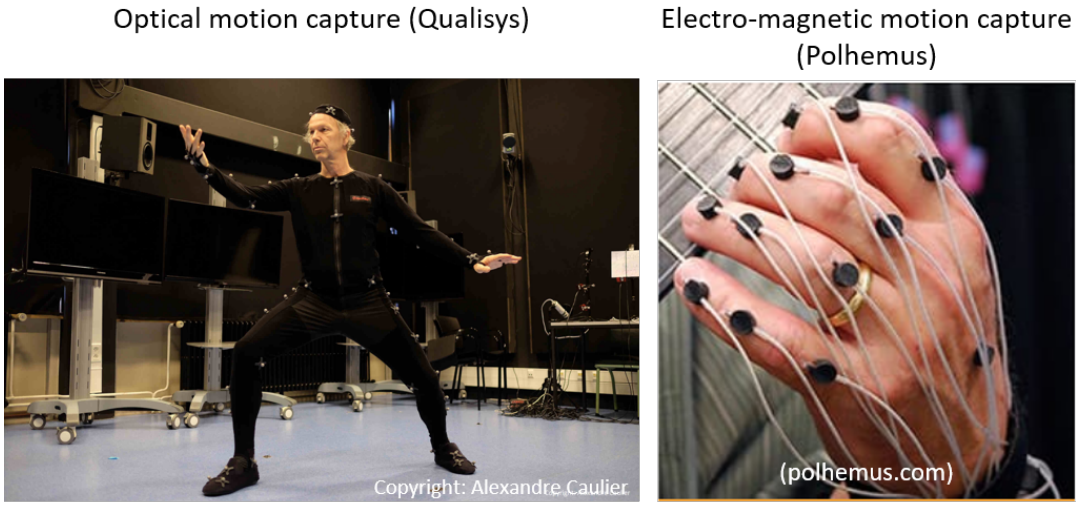

Figure 3.2: State-of-the-art magnetic and optical MoCap systems. Left: Qualisys (optical). Right: Polhemus (electromagnetic).

partly explain the delay of the research on human motion analysis compared with other domains, where a low-cost microphone or camera is sufficient to record a large dataset or real-time data. Nevertheless, although such systems are still rarely used in research because of their price, low-cost devices such as inertial sensors and depth cameras allowed a first insight of the potential of motion analysis.

The recent spreading of MoCap systems is illustrated in two specific and recently very active research areas (see Fig 3.3): Lalys and Jannin (2014) proposed a systematic review of surgical process modeling, showing the growing interest on the subject from 2007. Note that most of the studies were still based on observation, surgical robot data measurements and video recordings, and that only 15% of the studies actually used MoCap devices. In a different context, Camomilla et al. (2018) presented a thorough systematic review of the use of inexpensive wearable MoCap devices (inertial sensors) for sports performance analysis, expanding from 2009 as illustrated in Fig 3.3. Most of these studies are still based on the use of a few sensors, placed either on the body or on an object (e.g. a racket or a golf club). Among the 286 publications selected in this review, 19 were based on machine learning, generally for activity detection, recognition and evaluation.

During the last few years, the use of high-quality 3D full-body MoCap systems has been expanding, especially with multi-camera optical MoCap systems (e.g., Qualisys^{TM}, see Fig 3.2, left image) and electromagnetic systems (e.g., Polhemus^{TM}, see Fig 3.2, right image).

These technologies have then been used in biomechanics (Olesh et al., 2014), in animation (such as *Happy Feet* and *Avatar* movies (Fischer, 2018)), as well as in music (Jensenius and Wanderley, 2010), dance (Aristidou et al., 2015) and sports analysis

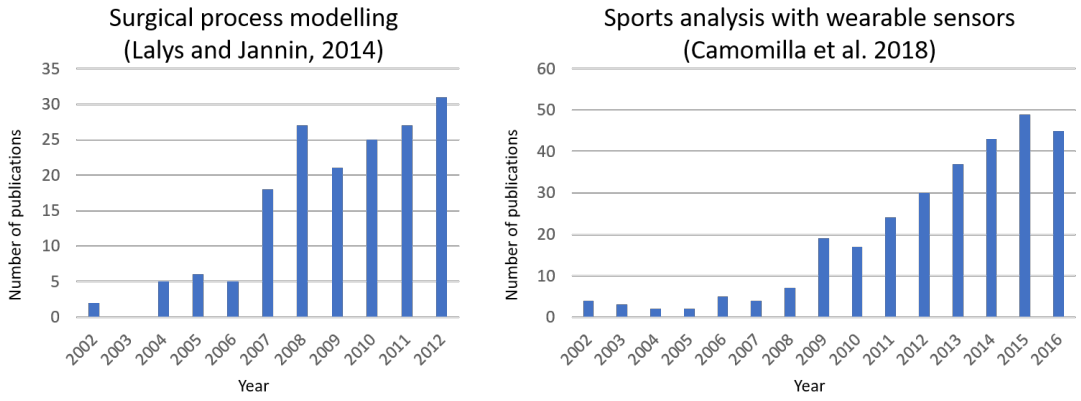

Figure 3.3: Evolution of the research on (left) surgical process modeling and (right) sports analysis with wearable sensors. The results are respectively reproduced from Lalys and Jannin (2014) and Camomilla et al. (2018).

(Federolf et al., 2012). More particularly, machine learning algorithms have been progressively tested on motion data, allowing computational interpretation of human motion, such as expressivity (Tilmanne and Dutoit, 2010; Dael et al., 2012), coordination (Daffertshofer et al., 2004; Dejnabadi et al., 2008), style (Aristidou and Chrysanthou, 2014) and skill (Federolf et al., 2012).

3.4 Expert gesture evaluation

The development of machine learning and MoCap technologies has unleashed research possibilities in many areas. One of the outstanding challenges is the analysis, and more particularly the evaluation of skill in an expert gesture. Gesture evaluation is essential in many disciplines, and has already been explored in different areas, including sports (Federolf et al., 2012), music (Tits et al., 2015), dance (Laraba and Tilmanne, 2016), rehabilitation (Pham et al., 2016), driving (Zhang et al., 2010) and even surgery (Megali et al., 2006). Regardless of the discipline, the typical approach to evaluating a gesture can be divided in four main steps:

- First, motion data must be captured. A dataset is recorded, generally including several individuals with different skill levels, for instance professionals or teachers and learners.
- Secondly, from these raw motion data, relevant features can be extracted. These features aim to represent motion in an efficient way regarding the targeted task. They can be derived for instance from body kinematics, kinetics, or motion decomposition. They can also represent relations based on prior domain-knowledge and semantic interpretation, such as expressivity, biomechanics, ergonomics or functionality (see Chapters 2 and 6).

- Thirdly, features can be analyzed, allowing selection or extraction of the best feature set for the targeted task. A common technique for multi-factor dependent data such as a MoCap data is factor analysis (Tits et al., 2016). The goal of this step is to reduce the global amount of information contained in the data to keep the most relevant and reliable information regarding the targeted task. For instance, this step enables to discard unwanted bias, such as morphology, age or expression, while keeping features more dependent on the targeted factor, i.e. skill in the present context. Features can also be post-processed, in order to extract a more relevant representation of motion. Dimensionality reduction techniques such as PCA can be used to extract a smaller set of relevant features. Morhology-independent features may also be extracted (see Section 2.4 and Chapter 7).
- Finally, the selected features, either low-level or high-level, are used to develop a model for the evaluation of the skill of a gesture. To validate an evaluation method, the results are generally compared to annotations provided by one or more expert of the discipline. These annotations may be either scores (for instance on a scale 0-10), classes (beginner, intermediate, expert), or a ranking of the participants. For specific sports, these scores can be derived from objective measures, such as a timing in a race, the speed or the accuracy of a shot, the number of balls a juggler can manage, etc. If none of these variables are available, the number of years and intensity of practice in a discipline can also be used to estimate the level of expertise.

Three major types of computational evaluation models can be found in the literature, and are presented in the following sections:

- Extract an unsupervised score directly from the features, i.e. without any use of annotations in the design of the model (see Section 3.4.1)
- Compute a similarity measure from a model of the ideal gesture (see Section 3.4.2)
- Predict a score using a regression or classification model (see Section 3.4.3)

A main interest of gesture evaluation is the ability to provide feedback on a performance. However, most of the previous works focus on the evaluation process,and do not explicitly show how to use evaluation results to provide feedback to a user. The few works proposing solutions for feedback on a gestural performance are presented in Section 3.4.4.

3.4.1 Direct (unsupervised) score

A score can be directly derived from the features and provide a quality index for the gesture. This score can actually be considered as an even higher-level feature. It

can then be compared to the annotations, taken as reference, using statistical tests, or a correlation index. Another validation is to evaluate a subject before and after a training period, with the hypothesis that the subject has progressed during this period.

Leroy et al. (2008) analyzed the influence of expertise on postural organization during juggling. On a dataset including two groups of juggles (five experts and five intermediate jugglers), they extracted specific high-level features, including elbow flexion/extension latency and maximal lateral oscillation of the pelvis. Using statistical tests, they showed significant differences in some of these features, allowing them to conclude that expertise influences the posture during juggling.

In a preliminary study, piano gestures were analyzed using PCA (Tits et al., 2015). PCs and eigenvalues were extracted separately for each motion sequences performed by each piano player. Then, from the eigenvalues, the cumulative variance of PCs were computed, and the number of PCs required to capture 95% of the total variance of the data were calculated. Finally, these numbers were compared to the experience of the pianists and showed that experienced pianists tend to use more PCs, reflecting a higher complexity of their movements.

Dadashi et al. (2015) analyzed front-crawl performances from nine professional and nine recreational swimmers. From accelerometers placed on both wrists and on the lower-back, they extracted specific features, including durations of arms pulling and pushing phases, time differences between both arms phases (representing their coordination), the stroke rate, length and velocity. They then computed the Cauchy Index for these variables, reflecting the differences of these features for two complete laps (50 m). This index represents the variability of the swimming patterns across different laps for a swimmer. Using an analysis of variance (ANOVA), they showed a significant difference of the Cauchy Index between the professional and recreational swimmers, indicating that professional swimmers have more stable swimming patterns.

Gløersen et al. (2017) analyzed motion of professional skiers using PCA as well as COM movement features. From the PCs extracted from a matrix containing all motion sequences (also called Principal Movements (PMs)), they computed specific features such as peak-to-peak differences, timing differences between stride cycles and stride cycle symmetry. They then compared all these features to the skiers' rankings in the International Ski Federation (FIS points), using Pearson's correlation. They showed high relations between FIS points and some of these features, including $R = 0.92$ ($CI = [0.45, 0.99]$) for COM lateral movement amplitude, and $R = -0.87$ ($CI = [-0.21, -0.99]$) for the periodicity of the fourth principal movement, accounting for an asymmetric movement of the legs in the sagittal and coronal planes.

Zago et al. (2017) also used PCA to analyze motions of street jugglers juggling 3, 4 or 5 balls. From the PMs, they computed autocorrelations and relative amplitudes.

Then, a second PCA was performed separately for each motion sequence of each participant. From these PCs, they extracted specific features, including variances of the first four PCs, and the residual variance (i.e. the sum of the variances of all the other PCs). For all these features, they used Mann-Whitney tests and showed significant differences between two groups, defined as the expert group (including participants able to juggle 6 or more balls), and the intermediate group (able to juggle 5 balls).

Alborno et al. (2017) proposed a score for evaluation of the synchronization of karate motions. To that end, they extracted limbs CoM acceleration peaks using progressive filtering, and analyzed time delay relationships between the peaks of each limb. They derived an average degree of synchronization from the mean of these delays. To validate their score, they conducted statistical tests on two populations of karateka (a high-skilled population and a low-skilled one), and showed a significant difference between their synchronization scores for both arms and legs motions.

3.4.2 Similarity measure

A model of an ideal gesture can be derived from a part of the dataset, containing only gestures performed by experts of the discipline. A similarity measure can then be obtained by comparing a gesture to this model. This similarity measure can be considered as a score or a higher-level feature and be validated in the same way.

Bianco and Tisato (2013) used Dynamic Time Warping (DTW, Vintsyuk 1968) to evaluate karate techniques. DTW is an algorithm allowing the warping of a time series to align it temporally with another one (called the reference). It finds the optimal warping of frame indices of the time series to minimize its distance with the reference. From the 3D coordinates of 15 joints recorded with a Microsoft Kinect V1, they extracted a set of 14 angles between joints triplets. They then aligned this 14-dimensional time series to a reference one (performed by an expert), using DTW. They extracted a score on a $[0, 10]$ range, using a logistic function of the distance between the aligned time series and the reference one.

Aristidou et al. (2015) developed a similarity measure based on Laban features. From a set of 27 temporal features, they extracted statistics on a sliding window, including maximum, minimum, mean and standard deviation. For each window, they simply calculated the correlation between a novice trial and a reference one performed by a teacher. They iterated the same process either separately for each category of Laban features (body, effort, shape and space, see Section 2.3.4), or for the entire set of features. As a demonstration, they integrated their method in a dance learning platform, but did not propose any quantitative validation.

Pham et al. (2016) proposed a summative scoring system based on a trajectory double integral to define the spatial distance between two trajectories. They tested their score

in a sign word recognition task, on the public dataset Australian Sign Language (Kadous et al., 1995). Then, to validate their score as an evaluation method, they computed a score for a novice signer at different periods, and verified that there was a progression after several attempts for the same sign word.

Chen et al. (2016) used a k-Nearest Neighbors classification model to recognize three types of Timpani percussive gestures (legato, accent and vertical accent) . The classification model was trained on a teacher's motion, using arms and stick kinematic features. The classification rates for each participant were then used as a similarity measure to the professional percussionist's motions. As a validation, participants ranking obtained from their respective classification rates were compared to a ranking annotated by the teacher. For the six participants, the first two were correctly ranked, as well as the last one, while the three others obtained very close scores both in annotations and in classification rates, showing the possible effectiveness of the method.

Laraba and Tilmanne (2016) used a similar approach to evaluate Walloon dance motions recorded with a Microsoft Kinect V2 and Qualisys. They trained a Hidden Markov Model (HMM, Baum and Petrie 1966) with the data of an expert recorded with a Qualisys system (high-quality data), to classify three different basic steps. The input variables of the HMM were seven relational features inspired by Müller et al. (2005), and 36 distances between pairs of leg joints. They then adapted the HMM using maximal likelihood linear regression (Laraba et al., 2015) so that it can recognize motions recorded with the Microsoft Kinect V2. From the classification rates, they extracted a similarity score based on the time-normalized log-likelihood. They tested their method on one novice for the three dance steps and showed that the scores were lower than those obtained by a teacher. As a demonstration, they also integrated their method into a Walloon dance learning platform using the Microsoft Kinect V2.

Morel et al. (2017) proposed a measure of limb spatial and temporal errors in a motion based on DTW. From a set of motion sequences performed by experts and aligned using DTW, and computed their mean and their variance, leading to an average motion, called "nominal motion", and a "spatial tolerance" based on the variance. This process was performed separately for each limb (two legs, two arms, and trunk), leading to a nominal motion and spatial tolerance for each limb. To compute the spatial error of a novice trial, they aligned each limb motion with their nominal motion using DTW, and compute a Mahalanobis distance between them, to account for the spatial tolerance. As a measure of temporal errors, they computed the timing difference between the warping of each limb motion to their respective nominal motion. They validated their spatial error measure on a dataset of tennis serves and their temporal error measure on a dataset of karate motions, both annotated by professional coaches. As a result, the best-fitting exponential curves were computed for estimating the relation between measures and annotations of one expert, leading to $R^2 = 0.57$ for the spatial error measure of tennis serves, and $R^2 = 0.4$ in average (among two expert annotators) for the temporal error measure of karate motions.

3.4.3 Score prediction

A regression or a classification model can be designed from the extracted features to predict a score. The score to predict can either be derived from an objective measure (timing, shot speed, years of practice, etc.), or be annotated by an expert.

Harrison et al. (2007) used bivariate functional PCA to analyze the development of vertical jump performance on a sample of jumps performed by 49 children of about 4 to 10 years. Each jump trial was annotated by expert observers according to criteria of developmental stage (Gallahue et al., 2006). They labeled each trial according to three developmental stages. Discriminant analysis coupled with Mahalanobis distance was then used to design a classification model. The best model yielded a mean accuracy of 61% for each development stage of the hip-ankle coordination.

Zhang et al. (2010) analyzed driving skills of different drivers. From the steering angle of the wheel, they performed a Discrete Fourier Transform (DFT, Welch 1967). They then designed a classification model to recognize the level of twelve drivers as typical or expert (two classes) from 30 normalized DFT coefficients. The model consisted of a decision fusion of three independent classifiers, including a Multi-Layer Perceptron (MLP) with a hidden layer of 40 neurons, a decision tree, and a Support Vector Machine (SVM, Boser et al. 1992) with polynomial kernels. The decision fusion was performed using a majority-voting method. They obtained a classification accuracy of 79% for a model trained on a balanced dataset of expert and typical drivers.

Harding and James (2010) analyzed snowboarding performance from half-pipe championships. From video footage captured during competitions, they extracted features related to the amount of time athletes spent in the air, and to their snowboard rotation degree. They selected the two best features according to their correlation with scores provided by the jury during the competition. They then trained a multiple linear regression model with these two features to predict these scores, resulting in $R = 0.902$ in average for the various competitions. No cross-validation procedure was used.

Young and Reinkensmeyer (2014) analyzed video-extracted 2D kinematic data of competitive diving performance. To predict a score given by a professional jury, they used PCA following Troje (2002) to extract eigenpostures and PMs. They also extracted a set of specific features commonly thought to influence dive judging, including the splash area time evolution, the board tip trajectory, and the body center coordinate trajectory. From these features, they selected different subsets, and performed a second PCA on them (following Troje (2002) again). For each subset, they designed a linear regression model with the PCs to predict the score provided by the jury. They obtained the best results ($R^2 = 0.66$) with the PMs and the three specific features, using a leave-one-participant-out cross-validation procedure.

Pirsiavash et al. (2014) also analyzed competitive diving, as well as figure skating performances. From 2D kinematic data extracted from videos, they computed a Discrete Cosine Transform (DCT, Ahmed et al. 1974) on each pose (i.e. each frame), and trained a Linear Support Vector Regression (L-SVR) to predict the score of a jury from the DCT of the pose. The developed regression models yielded $R = 0.41$ for the diving evaluation, and $R = 0.45$ for the skating evaluation.

Alexiadis and Daras (2014) used quaternionic motion representations to align MoCap data spatially and temporally, to make them more comparable. On these aligned data, they then proposed three similarity measures, using quaternionic correlations on joint positions (i) and velocities (ii). The quaternionic representation of motion data was used to be able to handle 3D coordinate variables jointly, representing them as pure quaternions (with a zero scalar value). Inspired by the 2D optical flow literature, they also proposed a similarity score based on 3D flow (iii). They proposed a weighted sum of these scores (which can be here considered as features), and optimized these weights using Particle Swarm Optimization (PSO) (Kennedy and Eberhart, 1995) to predict a ranking annotated by an expert. They validated their method on an annotated dataset of Salsa motion, by computing an adapted version of the Kendall rank correlation coefficient (Kendall, 1938). With optimized weights, the dancers were ranked with 20.5% ranking error according to the adapted Kendall's equation.

From a dataset of karate motion, Zago et al. (2016) used PCA to extract eigenvectors (called eigenpostures) and PCs weighting (PMs). They also extracted CoM kinematic features. From these features, they selected different subsets, and performed a second PCA on them. For each subset, they designed a linear regression model with the first five PCs to predict the experience of the karateka, using a leave-one-trial-out cross-validation procedure.[2] They finally obtained the best Root Mean Square Error (RMSE) for the combination of all features ($RMSE = 3.72$ years, $R^2 = 0.908$).

3.4.4 Feedback

One of the main interests of evaluating a motion is to propose feedback, allowing an improvement of the motion performed. However, most existing works limit their focus to the evaluation process, and do not explicitly show how to use evaluation results to provide feedback to the performer. Some methods are based on a rather simple direct score. The feedback procedure for these evaluation methods can be straightforward, by explaining the features involved in the score. However, the interest is limited. A system can tell the user that s/he should improve her/his synchronization, or her/his coordination, but the participant would not have a clue about

[2] A leave-one-trial-out procedure means that from the whole dataset of N trials performed by different participants, N-1 trials are used to train the model, the excluded one is then tested on that model. Note that the risk of this procedure is that the model can simply identify a participant from the training dataset, rather than really predicting its level. A better procedure would be a leave-one-participant-out procedure, excluding all trials of a participant from the training dataset.

how to do so. It would be more interesting to provide feedback about lower-level features, so that the subject knows which part of her/his motion should be corrected.

Concerning a similarity measure or a prediction of a model based on multiple features, it can be interesting to investigate which features are mostly responsible for the computed score. It would allow a feedback system to tell which feature should be modified, and how it should be modified, to improve the similarity with the model or the predicted score. An even more interesting information for the user could be a visualization of an improved version of the motion.

In that direction, Young and Reinkensmeyer (2014) proposed an original method which consists in synthesizing a new motion corresponding to a particular level. From a score, they calculated the eigenpostures and PMs which would predict this score using the linear regression coefficients (see Section 3.4.3), and using an inverse processing of the motion decomposition (following Troje 2002), they calculated a gesture corresponding to these eigenpostures and PMs. From the visualization of the synthesized gestures from different scores, they saw that better dives were performed with a straighter position of the legs, with a body path closer to the board tip, and with a narrower splash area.

Pirsiavash et al. (2014) modeled expertise from their diving and skating datasets using an L-SVR on the pose DCT (see Section 3.4.3). To provide a feedback to the user, they computed the gradient of the score obtained with L-SVR, and performed an inverse-DCT to show the modification of the joint positions leading to an improvement of the score.

Morel (2017) proposed a simple semantic feedback method, based on a previous work presented in Morel et al. (2017) (see Section 3.4.2). For each limb's spatial and temporal largest error during a trial, she indicates the time, duration and direction of the error. By comparing the trial to the nominal motion, the system indicates if a limb is too high, low, left or right for spatial error, or if it is too late or in advance.

Patrona et al. (2018) proposed a similar semantic feedback method. First, they used DTW for human action evaluation. They aligned positions as well as kinetic energies of eight joints (elbows, wrists, knees and ankles) of a test motion with a reference motion. They then proposed a semantic feedback using a fuzzy logic engine. The feedback was both visual and textual, showing key frames corresponding to the highest errors in position and velocity, and explaining which joint and at which frame was causing the highest error, both for position and velocity features, and proposing a correction to apply to the motion. This method allows an individual feedback, showing explicitly to the user how to improve her/his own technique.

3.5 Discussion and conclusion

In this chapter, a state of the art was presented on expert gesture evaluation. The various methods proposed can be divided into three main categories: evaluation

from a direct score extracted from features, a similarity measure to a model, or the prediction of a score using a regression or a classification model. Additionally, a few works also focus on the development of feedback methods. A summary of all these works can be found in Table 3.1.

These methods may have various advantages and drawbacks. Models based on direct scores do not intrinsically depend on experts annotations. They are fully based on the extraction of specific features, and do not rely on the training of an algorithm supervised with the annotations. Instead, they are usually based on prior knowledge of a specific discipline and therefore rarely allow generalization to other disciplines. Models based on similarity measures are partly supervised, since the data used for the modeling of the ideal gesture are supposedly performed by experts. However, they are not optimized to predict the level of expertise for new instances using supervised learning. For most of these methods, the validation is therefore limited to statistical significance allowing the discrimination of two groups, or to examples of the use of the method, and do not prove their accuracy in the evaluation of any performer's level. Morel et al. (2017) investigated the correlation between their proposed similarity measures and annotations. However, for the experiment, annotators were specifically asked to evaluate some characteristics of the gestures evaluated by the algorithms, such as the delay between two punches in karate, or accuracy of specific limb positions in a tennis serve. Gløersen et al. (2017) analyzed the correlations between the extracted features and the FIS points of professional skiers. However, their results were based on only six skiers, and they do not prove the generalization on other performers (no cross-validation). Moreover, a large number of features was analyzed, increasing the chance that one of them had a correlation with the FIS points due to coincidence.

On the opposite, score prediction methods are based on classification or regression models, trained to predict the level of any performer from the annotations of a dataset. The quality of these models therefore depends on the quality of the annotations. According to the discipline used as a case study, different types of annotations can be derived and can be more or less objective. They can be a ranking in a competition, or a subjective score annotated by a jury. In the last case, an average score averaged from several independent annotators is preferred to reduce its subjectivity.

Unfortunately, none of these works provide their dataset publicly as a benchmark for motion evaluation. Each one of the proposed methods is tested on a new dataset, recorded for that purpose. It is thus never possible to verify their results, and more particularly to compare different methods on the same data. Moreover, all these works are generally validated on a single category of gesture. The generalization to other gestures is left as an assumption. An exception can be raised to some degree concerning the use of eigenpostures and PMs following Troje (2002) procedure, as it has been successfully reused in a number of studies including various sports disciplines.

Reference	Type	Dataset (discipline, v, n, p)	Features	Model/Validation method
Harrison et al. (2007)	S.P.	Jump, $7 \times 2D$, 49, 49	PCA	Discriminant analysis
Leroy et al. (2008)	D.S.	Juggling, $5 \times 3D$, 100, 10	- Height of the ball - Elbow flexion/ext. - Pelvis lateral oscillations - Ball cycle duration - Elbow flex./ext. cycle - Pelvis oscillation cycle	Stat. significance tests
Zhang et al. (2010)	S.P.	Driving, 1, 1065, 12	steering angle DFT coefficients	Voting: - MLP - SVM - Decision tree
Harding and James (2010)	S.P.	Snowboarding, video, 169, 169	- 3 Air-time feat. - 4 Degree of rotation feat.	Selection with SLR Prediction with MLR
Bianco and Tisato (2013)	S.M.	Karate, $15 \times 3D$, ns, ns	DTW on 14 body angles	Example (preliminary)
Young and Reinkensmeyer (2014)	- S.P. - FB	Diving, $8 \times 2D$, 16, 16	PCA on - Eigenpostures - PMs - CoM traject. - Splash area - Board tip traject.	Linear regression
Pirsiavash et al. (2014)	S.P. FB	Diving, $25 \times 2D$, 159, ns Skating, $25 \times 2D$, 159, ns	pose-DCT	L-SVR
Alexiadis and Daras (2014)	S.M. S.P.	Salsa, ns, 9, 9	- Quaternionic R of joint pos. - Quaternionic R of joint pos. - "3D optical flow"	Score weighting using PSO
Tits et al. (2015)	FB	Piano, $27 \times 3D$, 16, 4	PCs cumulative variance, PMs	Preliminary comparison
Aristidou et al. (2015)	S.M.	Dance, $11 \times 3D$, 3, 1	Laban, sliding-window stat. R variability (Cauchy Index) of:	Example (preliminary)
Dadashi et al. (2015)	D.S.	Swimming, $2 \times 3D$, 5981, 18	- timings - stroke rate - arm coordination - stroke length and speed	Stat.significance tests
Pham et al. (2016)	S.M.	Sign language, $1 \times 3D$, 2, 1	Traject. double integral	Example (preliminary)
Chen et al. (2016)	S.M.	Timpani, $21 \times 3D$, 210, 7	Kinematics, k-NN classif. rate	Ranking comparison
Zago et al. (2016)	S.P.	Karate, $42 \times 3D$, 50, 10	PCA on - Eigenpostures - PMs - CoM kinematics	Linear regression
Laraba and Tilmanne (2016)	S.M.	Dance, $11 \times 3D$, ns, 2	HMM on - 7 Müller - 36 leg joints dist.	Example (preliminary)
- Morel et al. (2017) - Morel (2017)	- S.M. - FB	- Tennis, $25 \times 3D$, 147, 17 - Karate $25 \times 3D$, 95, 15	- DTW-Mahalanobis-distance - Alignment difference	R^2
Alborno et al. (2017)	D.S.	Karate, $25 \times 3D$, 22, 5	CoM acc. peak-delta timing	Stat.significance tests
Gløersen et al. (2017)	D.S.	Ski, $41 \times 3D$, 72, 6	- PMs peak-to-peak diff. - PMs timing diff. - PMs stride symmetry - PMs auto-correlations	R
Zago et al. (2017)	D.S.	Juggling, $23 \times 3D$, 36*, 12	- PMs relative amplitudes - PCs variances	Stat.significance tests
Patrona et al. (2018)	S.M. FB	Basic actions, $20 \times 3D$, N/A	DTW+fuzzy-logic on - pos. - KE	Visual feedback examples

Table 3.1: Motion evaluation methods. D.S.= Direct Score. S.M.= Similarity Measure. S.P. = Score Prediction. FB = feedback. v = number of captured variables. n = number of samples. p = number of participants. ns = not specified. PM = Principal movement. *: for each task. R = Pearson's correlation. R^2 = coefficient of determination. SLR: Single Linear Regression. MLR: Multiple Linear Regression..KE = Kinetic Energy.

On the other hand, while a large set of low or high-level features have been proposed in previous works (see Chapter 2), all studies on motion evaluation generally focus on a single category of low or high-level features. More particularly, few studies use ergonomic features for motion evaluation, although this category of features could be relevant for modeling expertise. A major objective of this thesis is to develop different models of expertise, using a large set of features from different categories, and confront them in the same evaluation task (see Chapters 8 and 9).

Another major limitation of the previously cited studies can be raised by comparing them to existing work in different fields of motion processing. The fields of gesture recognition, retrieval and synthesis are in high development due to their potential uses in industry, for human-machine interaction, surveillance, and animation. Recent works in this domain are based on the latest advances in artificial intelligence, i.e. deep learning and reinforcement learning (Holden et al., 2015, 2016; Neverova, 2016; Xia et al., 2017; Ding et al., 2017; Laraba et al., 2017; Devineau et al., 2018; Peng et al., 2018). The gap with research in motion evaluation may be explained by the fact that these deep-learning-based techniques require a large amount of data to be efficiently used. However, the recording of datasets for motion evaluation is a laborious task. A high-quality of the data is necessary in order to extract subtleties of the motion characterizing expertise; and a large number of participants with different levels of expertise is needed for the design of a complex model, able to discriminate a skilled or a low-skilled performance.

In this context, we recorded a new high-quality MoCap dataset of Taijiquan gestures, including 12 participants with different levels of expertise from novice to expert (see Chapter 4). We also developed a new algorithm for automatic MoCap data recovery, to ease the laborious process of MoCap data recording and correction (see Chapter 5). In Chapter 9, we propose a first exploration of the use of deep learning technologies, and more particularly transfer learning, for the evaluation of expert gestures. Finally, we propose a generic method for visual feedback of the gesture evaluation, that can be used with any score-prediction-based evaluation model (see Chapter 10).

Part II

Data Collection and Processing

Chapter 4

Taijiquan motion capture dataset

Contents

4.1 Introduction

4.1.1 The need for a dataset

The last decades have seen a new adage becoming more and more popular, saying: "data is the new oil". Due to the recent explosion of computational power, and of new artificial intelligence algorithms, many ambitious projects can now be achieved, including beating the best chess player in the world, autonomous vehicles, realistic virtual agents, or even political manipulation. However, all these projects, achievable in theory, need a major resource to be completed: data. Depending on the objective, hours of driving data, thousands of chess games, thousands of pronounced and written sentences and their corresponding facial expressions, or millions of messages or interactions on social media must be acquired through any means to be able to design a specific model to solve a problem. Not only must they be acquired, but ideally they should be of quality, in a sufficient amount, processed, and often annotated. For instance, to train an algorithm for identification of an object in an image, a large

number of images including this object and without it must have been acquired, and labels should indicate which images contain the object to identify.

The aim of this thesis is the modeling of gestural expertise with the use of supervised machine learning algorithms. In this context, a motion dataset is needed. The recorded data must be accurate enough to provide relevant information about the expertise contained in the motion. Moreover, along with the data, labels about the quality of the gesture must be provided.

Recent technologies enable accurate recording of motion in 3D, generally using optical MoCap systems, such as Qualisys or Vicon. As an example, in Table 3.1 showing past studies on motion evaluation, most recent studies (from 2015) use 3D data generally recorded with an optical MoCap system, while many of the previous works are based on 2D data extracted from videos, or signals extracted from specific sensors (such as a wheel steering angle in Zhang et al. (2010)). A major issue of accurate MoCap techniques is that they are expensive and time-consuming. To record a performance of one participant, markers must be placed carefully on her/his body, following precise indications. An expensive MoCap system consisting of multiple cameras must be set up and calibrated. After the recording of a sequence, processing is generally required to correctly identify all the trajectories. Missing data must be reliably interpolated when possible. As a comparison, a cheap microphone is needed to record a voice with a good quality, and a cheap camera is sufficient to capture an image with a good quality, without post-processing. Moreover, a tremendous quantity of audiovisual data can be found directly on the internet.

In the literature regarding motion evaluation (see Table 3.1), a new dataset has been recorded for each research. Except for datasets of cyclic motions, 3D motion datasets include fewer than 200 samples of the same gesture. The datasets including more samples consist of cheaper sensors: Dadashi et al. (2015) used two accelerometers/-gyroscopes to record 18 swimmers performing each 300 m trials in three different conditions, leading to a total of 5981 front-crawl cycles, i.e. roughly 2000 cycles for one condition. Zhang et al. (2010) simply recorded the wheel steering angle to analyze the 1065 driving trials in two different conditions performed by 12 subjects. Chen et al. (2016) recorded 3D motions of 210 percussive gestures performed by seven subjects, using a Qualisys system. They recorded 21 markers placed on the upper-body and the mallet. However, their analysis is based on the trajectory in the sagittal plane of only one marker, placed at the extremity of the mallet. Apart from these datasets, the largest 3D dataset in Table 3.1, containing full-body 3D motion data for a rather complex gesture, was recorded by Morel et al. (2016), and contained 147 tennis serves performed by 17 subjects. Besides their small size, these datasets are generally limited to a single type of motion. Finally, these datasets, recorded for these specific studies, are generally not publicly shared to the scientific community. As a consequence, not only a limited amount of accurate 3D motion data has been recorded, but it is generally not publicly available.

To address this issue, some works led to the publication of large motion datasets. The most well-known dataset is the Carnegie Mellon University (CMU) Graphics Lab Motion Capture Database (CMU, 2003), consisting of 109 activities recorded by more than 100 subjects, with a total of 2605 motion sequences. The types of activities include basic human interactions and scenarios, locomotion, as well as physical and sports activities, generally performed by a single subject. Another large public dataset is the HDM05 Motion Capture Database (Müller et al., 2007), including about 1457 motion sequences of more than 70 activities, performed by 5 subjects. Mandery et al. (2015) recorded the KIT Whole-Body Human Motion Database, a large dataset of various locomotion and object-manipulation motions, including a total of 3704 motion sequences performed by 38 subjects. Apart from these datasets recorded with accurate optical MoCap systems, a number of lower-quality datasets were recorded for research on action recognition. Surveys on these datasets can be found in Ofli et al. (2013); Zhang et al. (2016); Mandery et al. (2016); Shahroudy et al. (2016).

However, these datasets are generally intended for animation, or general action recognition purposes. Only a few instances of each gesture are usually recorded, and with rarely more than one participant performing the same gesture. The main objective of such datasets is to include a wide range of different gestures. These data are thus not suited for the analysis of expert gestures, where a large number of instances of the same gesture by several performers is required. Moreover, little information is provided on the performers, and never on their experience in a particular activity.

Due to all these issues, a new 3D motion dataset of Taijiquan was recorded in the context of this thesis, and was made publicly available.

4.1.2 Why Taijiquan ?

Most of the gestural disciplines are usually focused on a part of the body. For instance, finger motions are crucial in piano, just as legs motions are important in football. They can also be focused on the gesture purpose, such as the aesthetics of dance gestures, the sound emitted by a musical instrument, or the speed or the spin of a tennis ball. These disciplines are thus not the best candidate to promote a general approach to study the question of motion evaluation.

Taijiquan is a martial art. However, it can be distinguished from other martial disciplines that are usually mostly fight-oriented. Taijiquan can be defined as an 'art of body awareness'. The practice of this discipline intends for the development of physical abilities such as balance, coordination, etc., as well as mental skills such as concentration. These characteristics make Taijiquan a well-suited discipline to study gesture expertise. Contrarily to most other disciplines, Taijiquan practice is focused on the quality of the motion itself, as a whole. This general motion quality learned

ID	Gender (M/F)	Age	Weight (kg)	Height (cm)	Practice (year)	Category	$Skill_1$ (0-10)	$Skill_2$ (0-10)	$Skill_3$ (0-10)	$Skill_\mu$ (0-10)
P01	M	56	95	196	32	Expert	9.3	9	10	9.43
P02	F	57	78	163	30	Expert	9.6	9.1	10	9.57
P03	F	62	58	162	24	Expert	8.5	8.5	9	8.67
P04	F	47	53	150	12	Advanced	8.2	8	8	8.07
P05	F	71	61	163	14	Advanced	6.8	7.4	7.5	7.23
P06	M	25	76	180	10	Advanced	8.4	8.6	8.5	8.5
P07	F	49	57	157	4	Intermed.	7	6.8	6.5	6.77
P08	F	34	56	158	3	Intermed.	8	7.3	7	7.43
P09	M	51	90	178	2.5	Intermed.	6.9	6.8	6.85	6.85
P10	F	59	55	163	1	Novice	6	5.8	6.5	6.1
P11	F	65	58	165	0.2	Novice	5	4.9	5	4.97
P12	M	28	96	181	0.6	Novice	5.8	6	5.75	5.85
M		50.33	69.42	168	11.11		7.46	7.35	7.55	7.45
SD		14	15.93	12.46	11.15		1.37	1.29	1.53	1.38

Table 4.1: Personal details of participants. Skill was ranked with a score between 0 and 10 by three teachers. Each one of their rankings, as well as their mean ($Skill_\mu$) is indicated in this table.

by the way of Taijiquan can often be transferred to various other sports disciplines (Caulier, 2010).

The following Sections present the multimodal Taichi dataset recorded during this thesis with both Qualisys and Kinect data. In this work, only the Qualisys data were used, and are described in the following Sections. The complete description, including Kinect data, is available under the publication Tits et al. (2018a). The dataset is available for research purpose (license CC BY-NC-SA 4.0), at:
https://github.com/numediart/UMONS-TAICHI

4.2 Participants

Twelve people volunteered to participate in the dataset recordings. All of them attended courses in the Taijiquan school of Eric Caulier, and were assigned a category according to their level: Novice, Intermediate, Advanced or Expert (three teachers of the school). Each Taijiquan teacher also provided individual rankings for each participant, on a scale of 0 to 10. These rankings were provided independently by each teacher, from their personal knowledge of all the participants during courses. Relevant personal details for each participant, including age, height, weight, gender, practice experience and skill level can be found in Table 4.1.

4.3 Recording protocol

The MoCap system (Qualisys) consisted of 11 cameras fixed on the walls and ceiling of a recording studio, leading to a recording area of 4×4 m. This system tracked 68 retroreflective markers placed on the whole body (for detailed placement, see Table 4.2), with a frame rate of 179 Hz and a spatial accuracy < 1 mm. The dextrogyre coordinate system was placed on the ground, in the middle of the recording area, with the vertical axis as the z-axis. At the beginning of each recording, the participant was standing approximately above the origin of the coordinate system facing the x-axis direction. After each gesture, the participant was again approximately facing the x-axis direction.

All participants performed 13 different techniques of the popular Taijiquan style 'Yang', all learned at the Taijiquan school Eric Caulier. These techniques are divided into two main categories: the Five Exercises (Wu gong), composed of five simple gestures, and the Eight Techniques (Bafa), composed of eight more complex gestures (see details in Table 4.3). All techniques are described in detail in Caulier (2010). Videos of the gestures performed by a teacher are included with the dataset as supplementary information. During the recording session, each participant was asked to perform three different rendition types, as described in Table 4.4.

4.4 Data processing

Qualisys MoCap data were manually corrected using the Qualisys Track Manager (QTM) software. The corrected data were then extracted in standard 3D motion data formats (C3D and TSV). All missing data (generally due to marker occlusions) were estimated with an automatic MoCap data recovery method developed during this thesis, and presented in Chapter 5.

After the recovery process, 21 joint positions and orientations were extracted from the 68 surface markers using a Visual3d^{TM} pipeline as illustrated in Fig 2.1. All the motion features used in this thesis were extracted from these data (see Chapters 2 and 6).

4.5 Manual annotation (segmentation)

All renditions were manually labeled from Qualisys data to identify beginning and ending of each instance of a gesture. To that end, the MotionMachine framework (Tilmanne and d'Alessandro, 2015) was used. The annotation software created from

Marker label	Marker placement
Head markers (left and right)	
L/RFHD	Approx. over left/right temple.
L/RBHD	Back of the head, approx. in a horizontal plane with front head markers.
Torso markers	
CLAV	Clavicles, located approx. at the jugular notch.
STRN	Sternum xiphoidal process.
CV7	7th cervical vertebrae.
TV10	10th thoracic vertebrae.
Arm and hand markers (left and right)	
L/RAC	Acromion.
L/RUA1-2	Cluster of two markers placed on the lateral surface of the upper arm.
L/R_HLE	Humerus lateral epicondyle.
L/R_HME	Humerus medial epicondyle.
L/RF1-2	Cluster of two markers placed on the lateral surface of the forearm.
L/R_RSP	Radius styloid process.
L/R_USP	Ulna styloid process.
L/R_HM1	2nd metacarpal (index).
L/R_HL5	Lateral head of 5th metacarpal (pinkie).
Pelvis markers (left and right)	
L/R_IAS	Anterior superior iliac spine.
L/R_IPS	Posterior superior iliac spine.
Leg and foot markers (left and right)	
L/R_FTC	Most lateral prominence of the greater trochanter.
L/R_TH1-4	Cluster of four markers placed on the lateral surface of thigh.
L/R_FLE	Femur lateral epicondyle.
L/R_FME	Femur medial epicondyle.
L/R_SK1-4	Cluster of four markers placed on the lateral surface of shank.
L/R_FAL	Lateral prominence of the lateral malleolus.
L/R_TAM	Medial prominence of the medial malleolus.
L/R_FCC	Aspect of the Achilles tendon insertion on the calcaneus.
L/R_FM1	Dorsal margin of the 1st metatarsal head.
L/R_FM2	Dorsal aspect of the 2nd metatarsal head.
L/R_FM5	Dorsal margin of the 5th metatarsal head.

Table 4.2: Table 2 - Marker placement. Labels and positions of 68 markers attached (scratched) to an elastic neoprene suit, according to Qualisys and C-Motion specification for standard full-body MoCap. Cluster markers (upper arm, forearm, thigh and shank) are placed approximately on the body and are only used for tracking in Visual3D™ software (C-Motion, Inc., Rockville, MD, USA).

Gesture ID	Name	Movement type
Five exercises (Wu gong)		
G01	Beginning position (Wuji)	Static posture, symmetric
G02	Tree posture (Taiji)	Static posture, symmetric
G03	Open and close lotus flower	Symmetric
G04	Bring sky and earth together	Symmetric
G05	Canalize energy	Asymmetric (left or right)
Eight techniques (Bafa)		
G06	Drive the monkey away	Asymmetric (left or right)
G07	Move hands like clouds	Asymmetric (left or right)
G08	Part the wild horse's mane	Asymmetric (left or right)
G09	Golden rooster stands on one leg	Asymmetric (left or right)
G10	Fair lady works shuttles	Asymmetric (left or right)
G11	Kick with the heel	Asymmetric (left or right)
G12	Brush knee and twist step	Asymmetric (left or right)
G13	Grasp the bird's tail	Asymmetric (left or right)

Table 4.3: Five exercises and Eight techniques of the Yang Taijiquan style.

Type ID	Description of the rendition
T01	**Five exercises** Each exercise is repeated four times in a row. After the four repetitions, a pause of 2-5 seconds is respected, before transition to the next exercise. For the fifth exercise (Canalize energy), which is the only asymmetrical gesture of the sequence, the four repetitions consist of a succession of left and right side gestures, in the order: 'left-right-left-right'.
T02	**Eight techniques** Each technique is repeated four times in a row. After the four repetitions ('left-right-left-right'), a pause of 2-5 seconds is respected, before transition to the next technique.
T03	**Chained eight techniques** Idem as the previous type, but no pause is respected during transition between two different techniques.

Table 4.4: Types of renditions performed by the participants.

Manual segmentation rules		
Gesture	**Start**	**End**
G01	(static posture)	(Static posture)
G02	(Static posture)	(Static posture)
G03	COM low[a].	COM low.
G04	COM high[b].	COM high.
G05	COM high.	COM low, foot take-off.
G06	COM low.	COM low.
G07	COM on one side[c].	COM on the other side.
G08	COM back at the center[d] (Foot take-off).	COM back at the center
G09	Foot take-off.	Foot starts to go down.
G10	COM back at the center.	COM back at the center.
G11	COM low (Just before foot take-off).	COM low.
G12	COM back at the center.	COM back at the center.
G13	Just before foot take-off.	COM back at the center.

[a]COM low: local minimum of COM z-axis.
[b]COM high: local maximum of COM z-axis.
[c]COM on one side: local extremum of COM y-axis.
[d]COM back at the center: local extremum of COM y-axis, generally near y-axis mean position.

Table 4.6: Manual segmentation rules for the 13 gestures based on visual indications on direct 3D motion and COM coordinates.

this framework allows mouse-controlled simultaneous visualization of 3D movements (Qualisys data), and 2D curves displaying temporal evolution of each coordinate of their Center Of Mass (COM), estimated from the mean position of the 68 markers. COM coordinates can be used as a global visual indication for systematic segmentation, as described in Table 4.6. Fig 4.1 shows an example of the annotation procedure. In this example, gestures G06 and G07 are being annotated.

From annotations, Qualisys data were automatically segmented using the MoCap Toolbox for Matlab (Burger and Toiviainen, 2013a) and MoCap Toolbox extension[1]. All unsegmented files were named using the convention 'PppTttCcc' (e.g. P01T01C01) for which 'pp' is the performer ID (see Table 4.1), 'tt' is the type of the sequence (see Table 4.4) and 'cc' is the number of the clip (repetition of the same sequence). All segmented files were named using the convention 'PppTttCccGggDddSss' (e.g. P01T01C01G01D01S01). 'gg' indicates the gesture (see Table 4.3), 'dd' indicates the direction (01 for left and 02 for right – symmetric gestures are denoted D01), and finally 'ss' indicates the instance of the gesture (as each gesture is repeated several times during a clip).

[1]MoCap Toolbox Extension : https://github.com/titsitits/MocapRecovery/tree/master /MoCapToolboxExtension

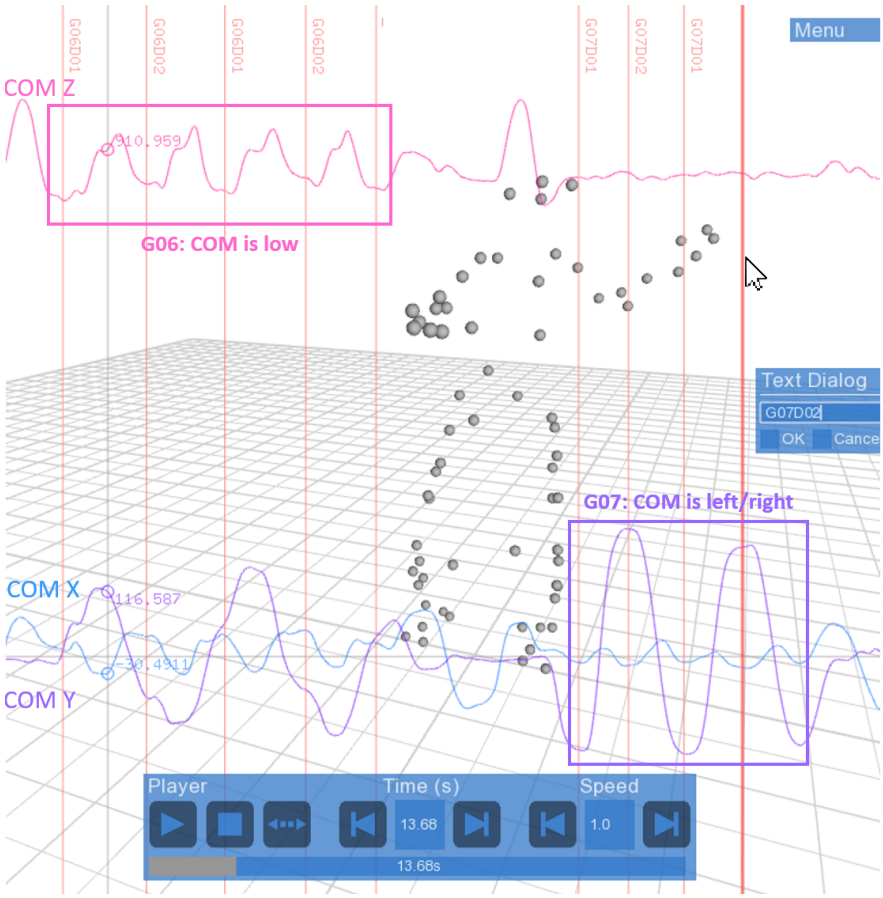

Figure 4.1: Screenshot of the annotation software. Layered display of: 1. 3D motion (gray spheres); 2. 2D-graphs showing evolution in time of the COM coordinates (blue = x, purple = y, pink = z); 3. Annotations (red vertical lines and labels). 4. GUI (blue windows, allowing navigation in the file, and label edition). In this example, G06 has been annotated, and G07 is being annotated. For G06, labels are placed when the z-axis of the COM is low, and for G07, labels are placed when the COM y-axis if low (COM is on the left) or high (COM is on the right).

4.6 Kinect data

Simultaneously with Qualisys data, a Microsoft Kinet V2 was used. The Kinect is a low-cost MoCap device, consisting of a single depth-camera, and allowing the recording 25 3D body joints at a frame rate of approximately 30Hz. The data of both systems were then synchronized manually. Alhough the Kinect data are not used in this thesis, they could be used in the future to investigate the possibility of using low-cost MoCap systems for the evaluation of expertise in gestures.

4.7 Conclusion

In this chapter, we presented a large 3D MoCap dataset of Taijiquan martial art gestures (n = 2200 samples) that includes 13 classes (relative to Taijiquan techniques) executed by 12 participants of various skill levels. Participants levels were ranked by three experts on a scale of [0–10]. The dataset was captured simultaneously with (i) a sophisticated optical MoCap system (Qualisys) consisting of 11 cameras that tracks 68 retroreflective markers at 179 Hz, and (ii) a Microsoft Kinect V2. Data of both systems were synchronized manually. The data from the Qualisys system were manually corrected, and then processed to complete any missing data (see Chapter 5). Data were also manually annotated for segmentation. The data were initially recorded for gesture recognition and skill evaluation, but they are also suited for research on synthesis, segmentation, multi-sensor data comparison and fusion, sports science or more general research on human science or MoCap.

To the authors' knowledge, it is the first dataset of sports gestures comprising simultaneously a large number of participants (12), a large number of different classes (13), and a variety of skill levels, and captured with two different MoCap systems. This dataset is used in the following chapters for validation of most of the methods proposed in the present thesis. These methods include morphology-independent feature extraction (see Chapter 7), gesture evaluation models (see Chapters 8 and 9), and a feedback model (see Chapter 10). Though only the Qualisys data are used in the present research, the Kinect data could serve in future research on gesture evaluation with low-cost sensors.

Chapter 5

Robust and automatic motion capture data recovery

Contents

5.1 Introduction

This chapter addresses a common issue of optical MoCap: missing data. The following has been partly published under Tits et al. (2018b). It must be noted that this work is not exclusively related to the field of gesture expertise evaluation, and can be useful to any application using optical MoCap data.

When using an optical MoCap system, if a marker or a body part is hidden from the cameras, its trajectory cannot be completely recorded, resulting in a gap in the MoCap data. Several issues may cause gaps, including occlusions, marker reflection quality, lighting condition, calibration or the limited area covered by the system. These gaps make it difficult and sometimes impossible to use the data (Liu and McMillan, 2006; Aristidou et al., 2008; Howarth and Callaghan, 2010). A number of methods have already been proposed to address this issue, based on various techniques. One basic method is direct interpolation. From an incomplete trajectory of a marker, the coordinates over time can be interpolated using standard methods, such as linear, spline or monotone piecewise cubic interpolation (Fritsch and Carlson, 1980), amongst others. Those methods are sufficient for small gaps (typically less than 0.5 second for human full-body motion (Liu and McMillan, 2006)), but are ineffective for larger gaps. More advanced time-series interpolation methods have been proposed, based on linear dynamic systems (Li et al., 2009), Gaussian process dynamic models (Wang et al., 2008), or Kalman filters (Aristidou et al., 2008).

Other methods are based on the fact that MoCap data generally consist of highly related trajectories of several markers, due to fixed bone length and to limited degrees of freedom in the skeleton. Expressing the incomplete trajectory using local coordinates, based on trajectories of three additional markers or based on a rigid body position and orientation, can be used to improve recovery (Howarth and Callaghan, 2010). Such coordinate transformation should reduce the variance of the trajectory representation, thereby easing the interpolation process. However, the three markers used for coordinate system transformation must have similar trajectories to the incomplete marker for the process to be efficient. This method thus highly depends on the number of complete marker trajectories available in the data.

Yet other methods for recovering missing data are based on human motion modeling, trained on a pre-recorded dataset. Liu and McMillan (2006) trained a global linear model and a set of local linear models from a training set of MoCap data. The local models are defined using segmentation with probabilistic principal component analysis (PCA), and K-means clustering. They first used the global model to recover missing data, then, from the results, they assigned a local model to each frame using a Random Forest classifier. On the other hand, Chai and Hodgins (2005) directly retrieved nearest neighbors of incomplete frames in a dataset, and trained a local linear model from these neighbors to recover the missing data. These methods are not fully automatic as they need a large dataset for the training of the models. Moreover,

the data to recover must have the same marker disposition as the one used in these pre-trained models. It means that for a new type of MoCap data (with a different marker disposition), an entire dataset must be recorded to train a new model.

Finally, some methods are based on matrix transformation techniques, using PCA (Federolf, 2013; Gløersen and Federolf, 2016), singular value thresholding (SVT) (Lai et al., 2011) or nonnegative matrix factorization (NMF) (Peng et al., 2015). These methods consider the entire motion as a matrix, with columns representing 3D components of all marker trajectories, and allow the use of information based on linear relations between the columns to reconstruct a gap in the matrix. The transformations are all based on low-rank properties of MoCap data. A key point with these methods is that a low-rank model of motion is trained on the available data of the motion sequence itself, and does not require a training dataset. These methods are thus automatic and can be used on any MoCap data (Feng et al., 2014).

A drawback of all previous methods is that the recovered trajectories may not respect human body properties, including bones' fixed lengths. Motion animations may thus lead to unrealistic results. Yet, other methods are directly based on the body skeleton, forcing marker positions to respect these properties. These constraints were successfully applied to several previously mentioned methods. Li et al. (2010) proposed the BoLeRO algorithm, combining skeleton constraints with linear dynamic systems. Tan et al. (2015) proposed a skeleton constrained SVT algorithm. Peng et al. (2015) adapted NMF to a hierarchical block-based skeleton structure model. However, such methods are generally significantly more computationally intensive as they are based on iterative optimization procedures. Moreover, they are often not automatic, as they are defined for a specific skeleton model based on a pre-defined marker set. Nonetheless, automatic procedures exist to estimate a skeleton structure in MoCap data (Kirk et al., 2005; De Aguiar et al., 2006).

Each one of the methods mentioned so far has different advantages and drawbacks, possibly making them more or less effective according to different factors, including gap length, number of markers, motion speed and complexity, and total motion sequence duration. For instance, interpolation techniques are inherently independent of the duration of the entire motion sequence and of the number of markers, unlike matrix-based and machine learning based techniques. The latter indeed require training data based on the frames of the sequence without missing markers, to model the relationship between markers. The quality of the model thus depends on the size of the training set, i.e. sequence duration, and the number of markers. On the opposite, machine-learning techniques may be more robust to gap length or motion complexity than interpolation-based methods (Liu and McMillan, 2006).

To the authors' knowledge, most previously proposed automatic MoCap data reconstruction methods are based on low-rank or temporal properties of motion, and use matrix operations to model human motion. Few papers focus on the use of machine-learning techniques such as linear and non-linear regression (Seber and Lee, 2003;

Bates and Watts, 1988) to model the motion of a missing marker. Moreover, no previous work known to the authors proposes the usage of ensemble learning to use likelihoods of different models and construct a more robust global model from the decisions of an ensemble of others (Dietterich, 2000; Hoeting et al., 1999).

Therefore, the aim of the research presented in this Chapter is to propose a probabilistic averaging method that can be used with any ensemble of recovery models, and that enforces movement constraints. This method is referred to below as *Probabilistic Model Averaging* (PMA). The averaging process is based on the posterior likelihoods of the distances between the recovered body points and other markers. To validate the method, we used existing recovery models and developed four new regression-based recovery models, which were used as inputs to the proposed probabilistic averaging method.

5.2 Method

Fig 5.1 shows the overall approach of our data recovery method, which can be divided into five main steps. First, parameters are extracted from each marker trajectory of the motion sequence. These parameters mainly represent relations between markers, and allow identification of related markers (termed as *reference markers* below) and of their distance distributions. Then, various recovery models are applied on the incomplete MoCap sequence, resulting in several candidate recovered sequences. For each candidate, a correction is applied to respect motion continuity (except for interpolation which inherently respects motion continuity). From all resulting individually recovered sequences, a weighted average is applied, in the spirit of ensemble learning systems (Dietterich, 2000). Finally, a spacing constraint is applied on the recovered trajectory, enforcing plausibility of the distance with related markers.

Note that it is good practice to center the motion sequence at the outset, by subtracting the mean position of all the markers available in each frame. This process makes it possible to reduce the component of global motion in the sequence, thus reducing motion variations. After the gap recovery process, the mean position is added back to translate the motion sequence to its original trajectories.

In the remainder of this Chapter, a motion sequence will be considered by a matrix representing all the trajectories of all the recorded markers during the entire sequence. This matrix has the dimension $N \times (3 \cdot M)$, where N is the number of frames of the sequence, and M the number of recorded markers. A marker trajectory p_j $(j \in 1, ..., M)$ is represented as an $N \times 3$ matrix.

The method was implemented with Matlab R2017a and the MoCap Toolbox (Burger and Toiviainen, 2013b). The code is available for free download at: https://github.com/numediart/MocapRecovery.

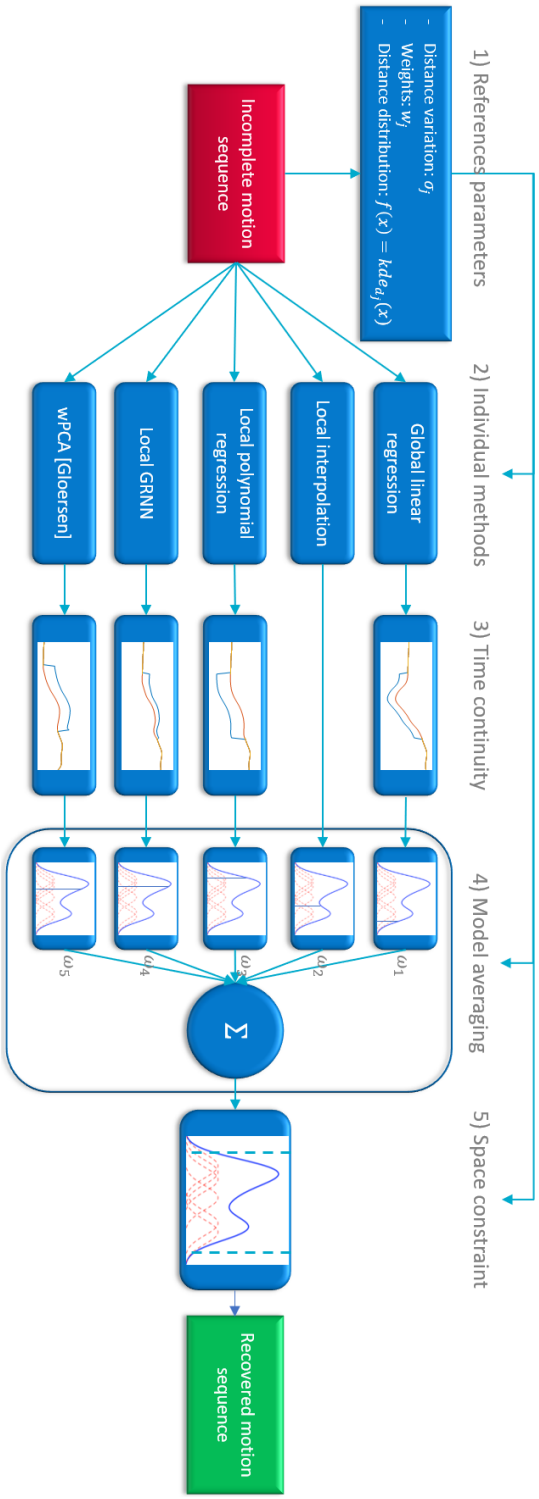

Figure 5.1: Block diagram of the proposed method. The overall process can be divided in five steps: 1) Extraction of marker trajectories parameters. 2) Individual recovery models. 3) Time constraint: trajectory continuity. 4) Distance-probability weighted averaging. 5) Spacing constraint: reference marker distance likelihoods.

5.2.1 Reference marker definition

All individual recovery models, as well as the proposed spacing constraint, depend on reference markers from which incomplete marker trajectories are to be recovered. We propose an automatic method to define such reference markers, based on inter-marker distance variations, inspired by skeleton-based methods where the intuition is based on the fact that joints of a skeleton have limited distance variations, due to bones' fixed lengths and limited degrees of freedom. We hypothesize that in many situations an incomplete marker trajectory can be recovered more effectively with methods using reference marker information, due to the close relation between the markers and their references. For instance, the wrist trajectory is more closely related to the elbow or shoulder trajectory than that of the foot. For each marker trajectory to be reconstructed, we can reorder related markers (referred to below as reference markers, or references) using distance standard variation. We denote by m the marker trajectory to recover, and p_j any other marker trajectory ($N \times 3$). The distance variation is computed as the standard deviation of the Euclidean distance between m and p_j:

$$\sigma_j = \text{std}(||(m - p_j)||) \tag{5.1}$$

The marker p_j with the smallest σ_j is the marker most related to m, i.e. the best reference for reconstruction of m. All markers can then be sorted as potential references for reconstruction of m, according to their distance variation with it.

5.2.2 Model 1: Global Linear Regression

An incomplete marker trajectory can be modeled based on trajectories of all present markers as input information. In the present context of human motion data, we have a complex problem involving underlying non-linear relationships, with potentially important quadratic or circular components, due to angular motions of skeleton segments. However, for the method to be fully automatic, the model training data are limited to the incomplete sequence to recover itself, i.e. N frames minus the missing frames.

Because many input variables are considered for the design of the model ($3 \cdot M$), the model must be simple, to avoid overfitting. Accordingly, we selected linear regression (Seber and Lee, 2003) for that task. Assuming that markers with lower distance variation σ_j are more suited for predicting the missing marker trajectory, we defined a threshold θ on σ_j, and avoided markers with $\sigma_j > \theta$ to simplify the model. This way, only relevant markers (numbered M_X) are used to model the position of the incomplete marker. This threshold was experimentally set to 50 mm in this research, as it gave the best results for the dataset, consisting of eight motion sequences with 41 markers (see Results, Table 5.1). Intuitively, a larger threshold could be considered for data with fewer available markers.

Let m be missing between frames n_1 and n_2. In other words, the rows from $n_1 + 1$ to $n_2 - 1$ of the matrix m must be recovered. A linear regression is performed on each axis i of m. Let $X = [p_j]_{j:\sigma_j<\theta, p_j \neq m}$ be the design matrix of dimension $N \times (3 \cdot M_X)$, including all marker trajectories p_j with $\sigma_j < \theta$, except the incomplete marker trajectory itself m (in practice, we add an intercept column to the design matrix, all missing frames are excluded, and we use only markers always present during the gap of m to recover). Denote by $X(n)$ as the n-th row of X, m_i the i^{th} column (or spatial axis) of m ($i \in 1, 2, 3$), and $m_i(n)$ its n-th element. The missing part of m is recovered using the following equations:

$$\beta_i = \text{argmin}_{b_i}(\sum_n (X(n) \cdot b_i - m_i(n))^2)_{i=\{1,2,3\}, n \in \{1,\dots,N\} \setminus \{n_1+1,\dots,n_2-1\}} \tag{5.2}$$

$$\tilde{m} = [X \cdot \beta_i]_{1 \leq i \leq 3} \tag{5.3}$$

where β_i is the vector of regression coefficients of m_i extracted through least square error minimization, and $\tilde{m}$ is the trajectory recovered with global linear regression. In practice, computation of $\tilde{m}$ is needed only for the missing frames.

5.2.3 Model 2: Local Interpolation

To simplify the modeling of the incomplete marker trajectory, skeletal motion properties involving strong relations between markers can be considered. A local coordinate system can be defined based on three references, and hence reduce the variance of the trajectory representation (Howarth and Callaghan, 2010).

Our second algorithm performs a local interpolation, i.e. an interpolation performed in a local reference defined by three other markers (the references).

Denote by p_1, p_2, and p_3 the first three reference markers (ordered by σ_j, see Eq. 5.1), used to recover m. Define a local coordinate system based on these markers, with three orthonormal vectors (u_1,u_2,u_3) at each time (or frame) n (here n indicates the row of a matrix ($1 \leq n \leq N$) For instance, $p_2(n)$ is the n^{th} row of p_2, i.e. a 1×3 vector):

$$v_1(n) = p_2(n) - p_1(n), \qquad u_1(n) = \frac{v_1(n)}{||v_1(n)||} \tag{5.4}$$

$$v_2(n) = v_1(n) \times (p_3(n) - p_1(n)), \qquad u_2(n) = \frac{v_2(n)}{||v_2(n)||} \tag{5.5}$$

$$u_3(n) = u_1(n) \times u_2(n) \tag{5.6}$$

where $\times$ indicates the cross product. m can then be projected into the local coordinate system:

$$P = [u_1(n)^T \; u_2(n)^T \; u_3(n)^T] \tag{5.7}$$

$$m_l(n) = (m(n) - p_1(n)) \cdot P \tag{5.8}$$

P is the projection matrix, and m_l the (projected) local trajectory.

m_l can be interpolated with simple linear interpolation. The recovered local trajectory can then be projected back into the original coordinate system:

$$\tilde{m}_l(n) = \begin{cases} m_l(n_1) + \frac{n-n_1}{n_2-n_1} \cdot (m_l(n_2) - m_l(n_1)), for\ n = \{n_1+1, ..., n_2-1\} \\ m_l(n), \text{otherwise} \end{cases} \tag{5.9}$$

$$\tilde{m}(n) = \tilde{m}_l(n) \cdot P^{-1} + p_1(n) \tag{5.10}$$

The interpolation is possible under the condition that all three references are present at frames n_1 and n_2. Also, if there are missing frames in a reference during the gap ($n \in \{n_1+1, ...n_2-1\}$), the incomplete marker trajectory m will only be partially recovered. In this case, we can iterate the process with other references on the residual gap (i.e. p_1, p_2, p_4 if p_3 was missing during the gap of m to fill, and so on) until m is completely recovered.

5.2.4 Model 3: Local polynomial regression

As just discussed, local interpolation takes advantage of markers relations by performing an interpolation in a local coordinate system. To further use that advantage, we can model and predict the position of the missing marker from its neighborhood (the local reference markers), using regression. As the number of input variables is much lower than for global regression, we can use a more complex model, able to model the non-linear relations between marker trajectories.

Our third algorithm is based on polynomial regression in the local coordinate system. First, the trajectory to recover is projected into a local coordinate system (m_l) defined by reference markers p_1, p_2, and p_3 (see Eq. 5.8). For each local coordinate of the marker to recover, a polynomial regression is performed, using reference markers local coordinates as input variables. In practice, only three input local coordinates are useful: the origin of the system is located in p_1 ($p_1 = (0,0,0)$), the new x-axis passes through p_2, giving $p_2 = (x_{p_2}, 0, 0)$, and the new y-axis is normal to the plane passing through p_1, p_2, and p_3, giving $p_3 = (x_{p_3}, 0, z_{p_3})$. Finally, the input set is composed of three variables:

$$X^l = \{x_{p_2}, x_{p_3}, z_{p_3}\} \tag{5.11}$$

For polynomial regression, the input variable set X^l is extended to quadratic polynomials in the input variables, leading to a set of 9 variables:

$$X^q = \{X_1^l, X_2^l, X_3^l, {X_1^l}^2, {X_2^l}^2, {X_3^l}^2, X_1^l \cdot X_2^l, X_1^l \cdot X_3^l, X_2^l \cdot X_3^l\} \tag{5.12}$$

The regression model is trained on the frames of the motion sequence where all markers (m_l,p_1,p_2,p_3) are present. The trained model is then used to predict all missing values of m_l:

$$\beta_i = \text{argmin}_{\beta_i}(\sum_n (X^q(n) \cdot \beta_i - m_i(n))^2)_{i=\{1,2,3\},n\in\{1,\dots,N\}\setminus\{n_1+1,\dots,n_2-1\}} \tag{5.13}$$

$$\tilde{m_l} = [X^q \cdot \beta_i]_{1\leq i\leq 3} \tag{5.14}$$

Finally, the recovered local trajectory $\tilde{m_l}$ can be projected back into the original coordinate system (see Eq. 5.10). Like local interpolation, this method is processed iteratively on sorted references until the trajectory of $\tilde{m}$ is completely recovered.

5.2.5 Model 4: Local Generalized Regression Neural Network

Generalized Regression Neural Network (GRNN) is a non-linear regression method, already used in various applications (Specht, 1991). It is a variant of an artificial neural network, consisting of four layers: the input layer, a radial basis layer, a summation layer and the output layer. GRNN allows to estimate any arbitrarily complex function, given a sufficient number of observations (generating the radial basis kernels). Comparatively to standard neural networks, GRNN does not require an iterative training. Moreover, as the output of the model is bounded by the extrema of the training dataset, a GRNN can only give physically meaningful outputs (Firat and Gungor, 2009). It means for instance that a GRNN should not predict marker positions with highly implausible distances.

The proposed algorithm applies a GRNN on local variables X^l, to model and predict the local trajectory m_l. The GRNN is thus trained with three input variables (in practice, each input variable is standardized by subtracting its mean and dividing it by its standard deviation). (X^l, see Eq. 5.11) and three output variables ($\tilde{m_l}$), according to:

$$\tilde{m_l}(n) = \frac{\sum_k m_l(k) \cdot \exp(-\frac{||X^l(n)-X^l(k)||^2}{2s^2})_{k=\{1,\dots,N\}\setminus\{n_1+1,\dots,n_2-1\}}}{\sum_k \exp(-\frac{||X^l(n)-X^l(k)||^2}{2s^2})_{k=\{1,\dots,N\}\setminus\{n_1+1,\dots,n_2-1\}}} \tag{5.15}$$

Parameter s determines the smoothness of the regression, and was experimentally set to 0.3 in this research (for standardized input variables), as it gave the best results for the dataset tested. Intuitively, a larger s could be chosen to recover slow motions, and a smaller one for sharp and fast motions.

Like the other local recovery models, local GRNN process is iterated on sorted marker references until the trajectory of $\tilde{m}$ is completely recovered.

All these individual models can be used independently to recover a trajectory, leading to several candidates. We explain in the next sections how these candidates are further processed and combined to produce a more robust recovery.

5.2.6 Time Constraint: Trajectory Continuity

Human motion time series are limited by two major constraints:

- a spacing constraint, defined by limited ranges of motion and fixed bone lengths;
- trajectory continuity, due to body inertia.

All recovery techniques that are based on interpolation intrinsically respect the continuity constraint. However, this is not the case of predictive models. To enforce continuity on recovered data, we can add a linear correction ramp:

$$\delta_{n_1} = \tilde{m}(n_1) - m(n_1) \tag{5.16}$$

$$\delta_{n_2} = \tilde{m}(n_2) - m(n_2) \tag{5.17}$$

$$\delta(n) = \begin{cases} \delta_{n_1} + \frac{n-n_1}{n_2-n_1} \cdot (\delta_{n_2} - \delta_{n_1}), for\ n = \{n_1+1, ..., n_2-1\} \\ (0\ 0\ 0),\ \text{otherwise} \end{cases} \tag{5.18}$$

$$\breve{m} = \tilde{m} - \delta \tag{5.19}$$

For each axis, we compute the difference between the real value and the predicted value at each border (δ_{n_1} and δ_{n_2}), and subtract from the recovered trajectory a linear ramp from δ_{n_1} to δ_{n_2}. An example of this correction is illustrated in Fig 5.2.

5.2.7 Probabilistic model averaging (PMA)

Depending on the context, each model can be more or less effective, making difficult the choice of the best model, and the development of a robust recovery method. To address this issue, we propose a model averaging method, based on the posterior likelihoods of the distances between the recovered body points and other markers. This method is inspired by Bayesian model averaging (Hoeting et al., 1999).

We estimate the *a posteriori* probability of each predicted location according to their distance to reference markers. For references p_1, p_2, and p_3, we estimate the distance distribution with m throughout the entire motion sequence on non-missing frames, using kernel smoothing density estimation ("kde" (Parzen, 1962) used with Silverman's rule of thumb to choose the bandwidth of the kernel estimator (Silverman, 1986)). For each recovery method k, a weight is computed:

$$d_j = ||p_j - m|| \tag{5.20}$$

$$f_j(x) = \text{kde}_{d_j}(x) \tag{5.21}$$

$$\breve{d}_{jk}(n) = ||p_j(n) - \breve{m}_k(n)|| \tag{5.22}$$

$$\omega_k(n) = f_1(\breve{d}_{1k}(n)) \cdot f_2(\breve{d}_{2k}(n)) \cdot f_3(\breve{d}_{3k}(n)) \tag{5.23}$$

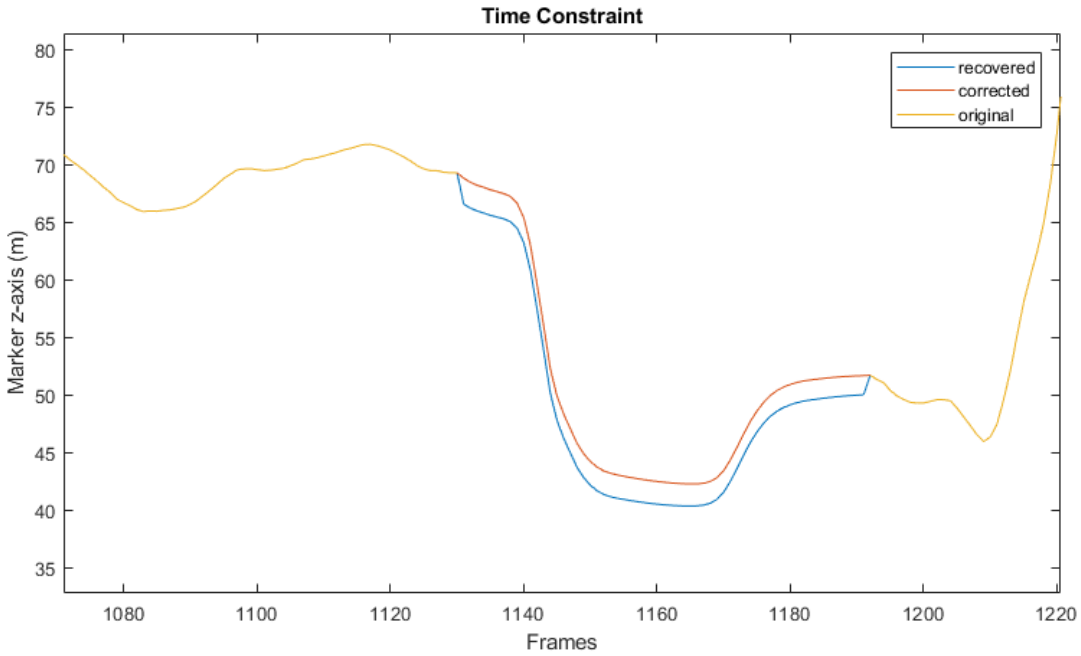

Figure 5.2: Trajectory continuity correction. The yellow curve shows incomplete data of a marker trajectory (m) on which a gap was introduced between frames 1130 and 1190 (only z-axis is shown). The blue curve represents the recovered data ($\tilde{m}$), and the red curve shows the corrected data using trajectory continuity constraint ($\check{m}$) (see Eq. 5.16-5.19).

Here ω_k is the weight of the trajectory $\breve{m}_k$ recovered with method k, and f_j is the estimated probability density function of the distance between m and the reference marker p_j.

We then compute a weighted average of all the recovered trajectories:

$$\bar{m}(n) = \frac{\sum_{k=1}^{K} \breve{m}_k(n) \cdot \omega_k(n)}{\sum_{k=1}^{K} \omega_k(n)} \tag{5.24}$$

K is the number of individual models used for the recovery.

This process allows to give more importance to most likely recovered trajectories, according to their distance with other markers. In the remainder of this Chapter, we denote this method as *Probabilistic Model Averaging* (PMA).

5.2.8 Spacing constraint: reference marker distance confidence interval

A final step is applied on the recovered trajectory. Knowing the probability density distribution of the distance $||p_1 - m||$, i.e. f_1 (see Eq. 5.21), we can check if the distance of the recovered trajectory with p_1 respects the confidence interval:

$$CI = \{x : F_1(x) \in [0.05; 0.95]\}, \tag{5.25}$$

where F_1 is the cumulated probability density function estimation of the distance $||p_1 - m||$, i.e. $F_1(x) = \int_{-\infty}^{x} f_1(\xi) d\xi$. The limits of this interval correspond to two spheres centered on p_1, with radii corresponding to:

$$r_1 = \arg\underset{x}{(F_1(x) = 0.05)} \tag{5.26}$$

$$R_1 = \arg\underset{x}{(F_1(x) = 0.95)} \tag{5.27}$$

If the recovered trajectory is outside these limits, it is projected onto the closest limit sphere. Fig 5.3 illustrates the projection of $\bar{m}(n)$ onto the limits of the confidence interval. In this example, the recovered frame $\bar{m}(n)$ is outside the confidence zone, as $x = ||p_1(n) - \bar{m}(n)|| > R_1$. The red arrow shows $\hat{m}(n)$, projection of the recovered frame $\bar{m}(n)$ onto the R_1-radius sphere centered at p_1, to fit the soft skeleton constraint. If the recovered frame is already in the confidence zone, no correction is applied: $\hat{m}(n) = \bar{m}(n)$. This process is thus used only if the recovered frame has low a posteriori probability, i.e. an unusual distance with its first reference marker p_1.

A stricter version of the spacing constraint can be applied by recursively projecting the recovered point onto the CI obtained for several reference markers (e.g. the first three references p_1, p_2 and p_3), iteratively until the point is at a plausible distance of each reference.

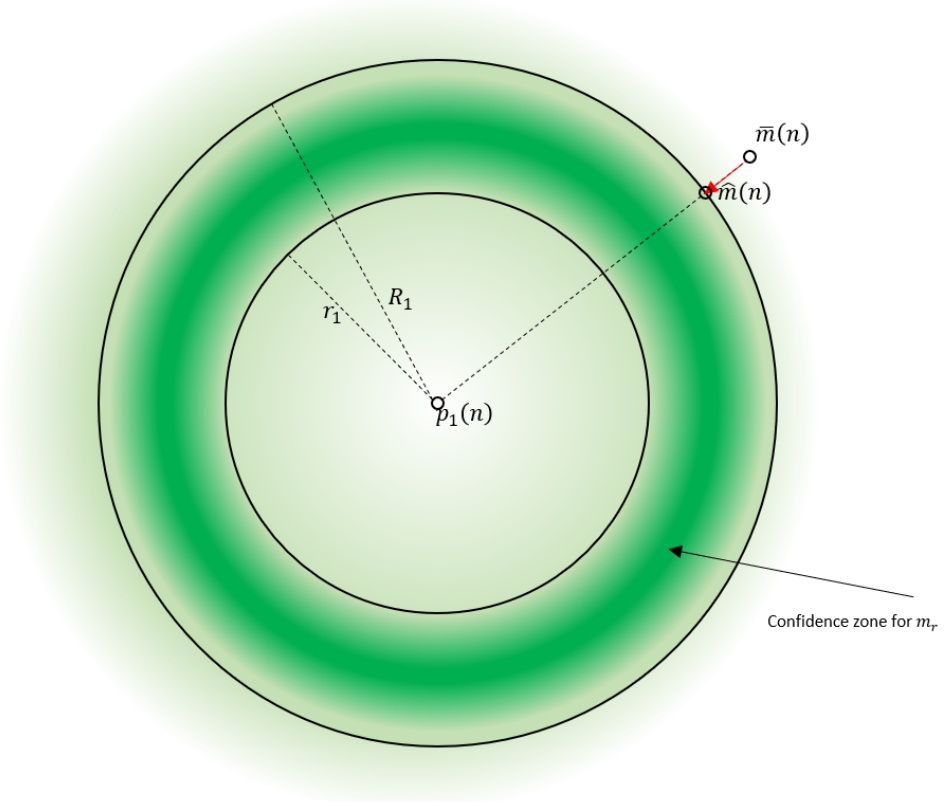

Figure 5.3: Reference distance soft constraints. The green intensity colormap indicates the probability of presence for the recovered frame. If the recovered frame $\bar{m}(n)$ is outside the confidence zone (delimited by spheres of radii r_1 and R_1), it is projected onto the closest point in this confidence zone ($\hat{m}(n)$).

File name	Type of motion	#markers	Duration (s)	#frames	fps
HDM1	Locomotion on the spot	41	33.25	3990	120
HDM2	Shelf (while walking)	41	70.83	8500	120
HDM3	Kicking and punching	41	63.93	7671	120
HDM4	Chair, table, floor	41	62.79	7535	120
HDM5	Clapping and waving	41	57.94	6953	120
CMU1	Break dance (Jumptwist)	41	6.75	810	120
CMU2	Break dance	41	37.5	4499	120
CMU3	Martial art (Empi)	41	43.35	5202	120

Table 5.1: Motion sequences used in the methods comparison.

5.2.9 Experiments

To validate our method, we tested each method used individually as well as the method combination with PMA. For such testing, we used the online CMU MoCap database (http://mocap.cs.cmu.edu/) (Hodgins) and the HDM05 database (http://resources.mpi-inf.mpg.de/HDM05/) (Müller et al., 2007). They contain a high number of various motion sequences, and they have been used by much of the related work (Liu and McMillan, 2006; Li et al., 2010; Tan et al., 2013, 2015; Peng et al., 2015; Gløersen and Federolf, 2016). Table 5.1[1] shows the motion sequences selected for the methods comparison. Motion sequences were selected to include a large variety of motions, in terms of complexity, type of motion, and duration.

The performance of the recovery method on a motion sequence may depend on different factors, including:

- The number of incomplete and complete marker trajectories
- The length of the gaps
- The sequence duration
- The motion complexity or periodicity

To analyze the performance of each method according to these factors, we introduced three concomitant gaps into our motion sequences, at random locations (uniformly

[1] The files names in this table corresponds to the following files in the respective datasets: HDM1= HDM_mm_01-02_03_120, HDM2 = HDM_mm_02-02_02_120, HDM3 = HDM_mm_03-02_01_120, HDM4 = HDM_mm_04-01_02_120, HDM5 = HDM_bd_05-01_01_120, CMU1 = 85_02, CMU2 = 85_12, CMU3 = 135_02.

distributed random markers and frames). We applied each method on these incomplete motion sequences, and extracted the recovery error for each method:

$$\epsilon = \frac{1}{g}\sum_{j=1}^{g}\frac{1}{n_{2_j} - (n_{1_j}+1)}\sum_{n=n_{1_j}}^{n_{2_j}} ||\hat{m}_j(n) - m_j(n)|| \tag{5.28}$$

g is the number of random gaps created, n_{1j} and n_{2j} delimit the location (in frames) of the randomly introduced gap j, and m_j and $\hat{m}_j$ are respectively the original and the recovered trajectories. We iterated this process 20 times with different random gap locations, and a mean recovery error was extracted from all iterations to estimate the general performance of each method. To analyze the influence of the sequence duration, fragments with different duration were extracted from each motion file.

Our method performances were compared to related work available online, namely the BoLeRo algorithm from Li et al. (2010) (Matlab code available for download at : https://github.com/lileicc/dynammo) and the weighted PCA-based reconstruction method from Gløersen and Federolf (2016) (Matlab code available for download at : https://doi.org/10.1371/journal.pone.0152616). In the sequel, these methods will be identified with the following numbers and acronyms:

1. Global Linear Regression (GLR)
2. Local Interpolation (LI) (Howarth and Callaghan, 2010)
3. Local Polynomial Regression (LPR)
4. Local GRNN (LGRNN)
5. weighted PCA-based method (PCA) (Gløersen and Federolf, 2016)
6. BoLeRo algorithm with soft bone constraints (BoLeRo) (Li et al., 2010)

We used the soft bone constraints version of the BoLeRo algorithm, with 16 hidden dimensions as proposed by Li et al. (2010). The PCA-based method was used with the parameters proposed by their authors, using the consecutive reconstruction strategy for multiple gaps (Gløersen and Federolf, 2016).

All results were obtained in MATLAB R2017a on a computer with Intel Core i7-4712HQ 2.3 GHz and 16 GB RAM running Windows 10.

5.3 Results

In this section, we present the results of the recovery on different simulated incomplete motion sequences. We analyze the influence of gap length, motion sequence duration, the number of incomplete marker trajectories, and the influence of the motion type.

5.3.1 Gap length

Fig 5.4 shows the results for two different motion sequences (respectively CMU1 and CMU3). In both cases, BoLeRo gives higher errors than all other methods. Moreover, the processing time, due to the iterative optimization process of the method, is significantly higher than others. For instance, the process duration is above the minute for filling three gaps of 2 seconds in the file CMU3, against less than a second for all other individual methods. For these reasons, and for better graphics readability, BoLeRo is left out in the remaining of the results.

On the right graphs, we can see results for all individual methods, except BoLeRo. Concerning the MoCap sequence CMU1 (top graph), Fig 5.4 shows a clear separation of each method accuracy, where LGRNN seems to reach the best accuracy (20.2 mm mean error for three random gaps of 5 seconds). Accuracy of all methods seems to decrease with gap size. Concerning the MoCap sequence CMU3 (bottom graph), GLR seems to give the best results, with a mean recovery error of 12.7 mm for three random gaps of 5 seconds.

Fig 5.4 also shows results of model averaging of several methods (dashed lines). In general, each combination of individual methods (all but 5 (PCA), all but 2 (LI), and all methods) seems to lead to an error comparable to that of the best individual methods in general. Our PMA method thus seems robust to gap size.

5.3.2 Motion sequence duration

Except for our most basic method based on interpolation (LI), each individual method performance may depend on the motion sequence duration. Indeed, more frames in the sequence means more information (more possible data variation), and more samples for model training.

To illustrate the influence of sequence duration on performance of gap recovery methods, fragments with different durations were extracted from each motion file. Fig 5.5 shows the mean recovery error for different sequence durations, for different motion sequences. We can see on each graph that all methods follow similar patterns, showing that their performance highly depends on the specific motion. Nonetheless, for most graphs (except for HDM5), the mean recovery error seems to be higher for sequence durations of 5 seconds, and decreases for a sequence duration of 10 seconds. Beyond that duration, the recovery is not much improved. Concerning individual methods, LGRNN seems to be more robust to sequence duration, compared to other regression methods. For long durations, GLR seems to give the best results of all individual methods.

For all durations and all motion sequences, PMA effectively weights each individual method, hence providing optimal recovery in any context. The best combination is the averaging of all methods but LI (dark red dashed line). Our PMA method is thus robust to motion duration.

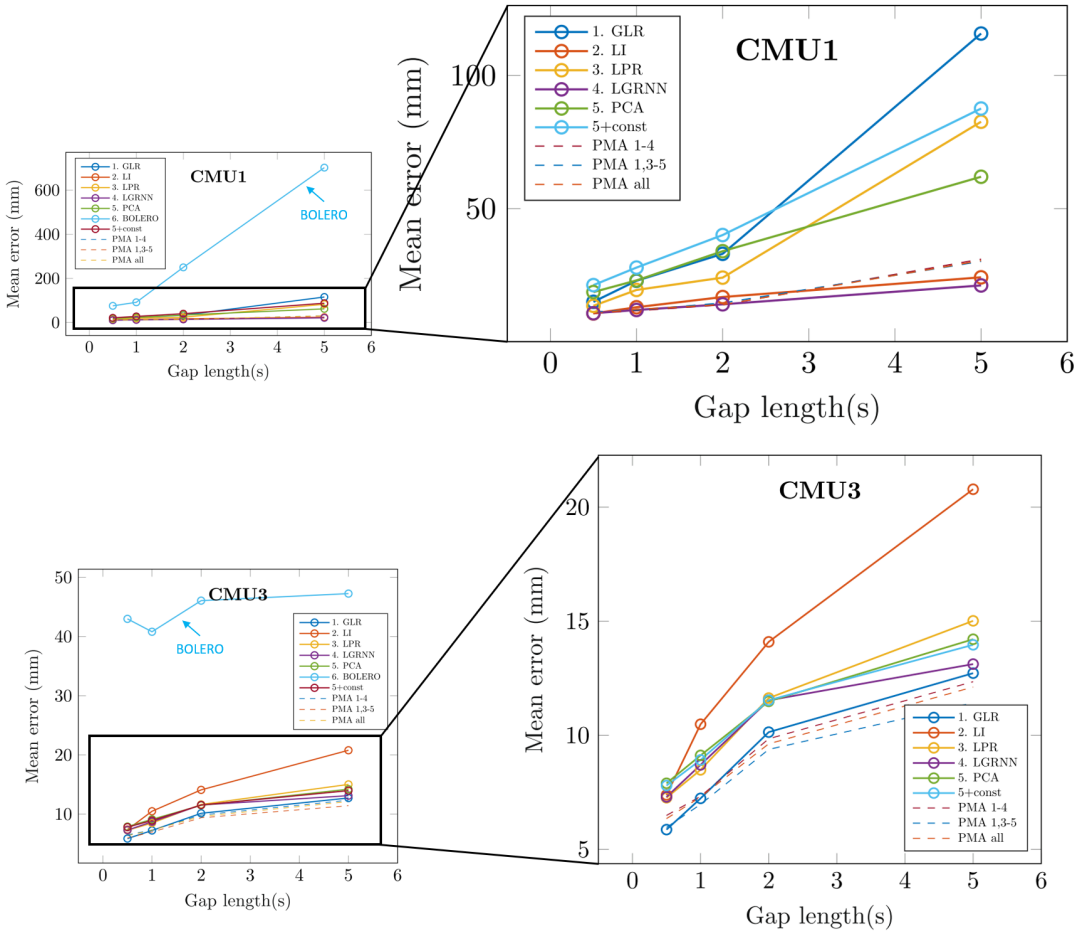

Figure 5.4: Mean recovery error for different gap sizes and gap recovery methods. Top: CMU1. Bottom: CMU3. Left: results including BoLeRo method. Right: results without BoLeRo method. Each point represents the mean of recovery errors, computed with 20 iterations, of three randomly created gaps of the same length (0.5, 1, 2 or 5 seconds). Solid lines show results for each individual method. Dashed lines show results for distance-probability averages of various combinations of individual methods.

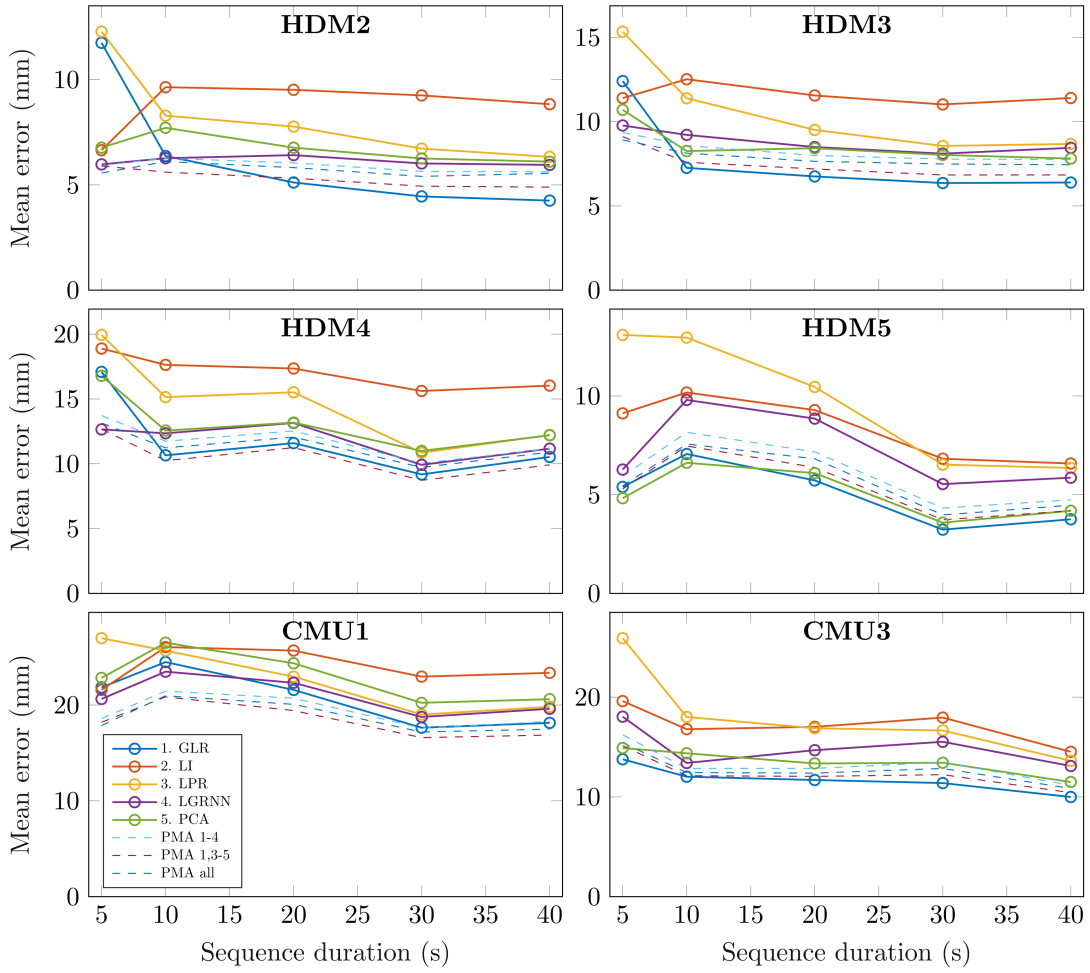

Figure 5.5: Mean recovery error for different sequence durations and gap recovery methods. To illustrate the influence of sequence duration on performance of gap recovery methods, fragments of different durations were extracted from each motion file. Each point represents the mean of the recovery errors computed on 20 iterations of three randomly created gaps of 1 second. Continuous lines show results for each individual method. Dashed lines show results for PMA with various methods combinations.

Sequence	$\bar{\epsilon}$**, no constraint (mm)**	$\bar{\epsilon}$**, constraint (mm)**	**Difference (mm)**	**t-test p-value**
HDM1	8.1	6.6	1.5	$p < 10e-5$
HDM2	4.8	4.4	0.4	$p = 0.46$
HDM3	5.5	5.2	0.3	$p = 0.07$
HDM4	8.5	7.4	1.2	$p = 0.08$
HDM5	4.5	3.0	1.5	$p = 0.001$
CMU1	17.1	15.7	1.5	$p = 0.008$
CMU2	14.4	13.5	1.0	$p < 10e-3$
CMU3	10.5	8.4	2.1	$p < 10e-10$

Table 5.2: Effect of constraints on mean recovery error (t-test, n = 200; conditions: 3 gaps of 1 seconds). Paired t-test (n = 200) on constraints effect on PMA for the reconstruction of 3 gaps of 1 second, introduced into different motion sequences. Individual methods 1 to 4 were used in this test.

5.3.3 Number of concomitant gaps

Except for basic interpolation or dynamic filtering methods, the reconstruction quality of one marker trajectory depends on the presence of reference markers. If several markers are missing at the same time during the motion sequence, less information is available for reconstruction. According to the method, the reconstruction quality may be influenced differently. Fig 5.6 shows the mean recovery error of individual methods and their PMA combinations for different motion sequences, and for different numbers of markers missing at the same time (gaps of one second). We can see in general that for all motion sequences and for all methods, the recovery error grows with the number of concomitant gaps. Again, PMA generally give the best results.

5.3.4 Constraints effect

For all previous results, time and spacing constraints were applied for all individual methods and model averages. To verify the effectiveness of these constraints, PMA reconstruction was tested with and without constraints for each motion sequence, with 200 iterations of three gaps of one second. For each motion sequence, a paired t-test was performed on the mean recovery error of the 200 iterations with and without constraint, as shown in Table 5.2.

Results show that for almost all tested motion sequences, PMA yields a significant improvement of the recovery method. The constraints did not improve the recovery for the sequence HDM2 (larger p-value), but this may be due to the fact that the recovery error is already low without constraint ($\bar{\epsilon} = 4.8$ mm).

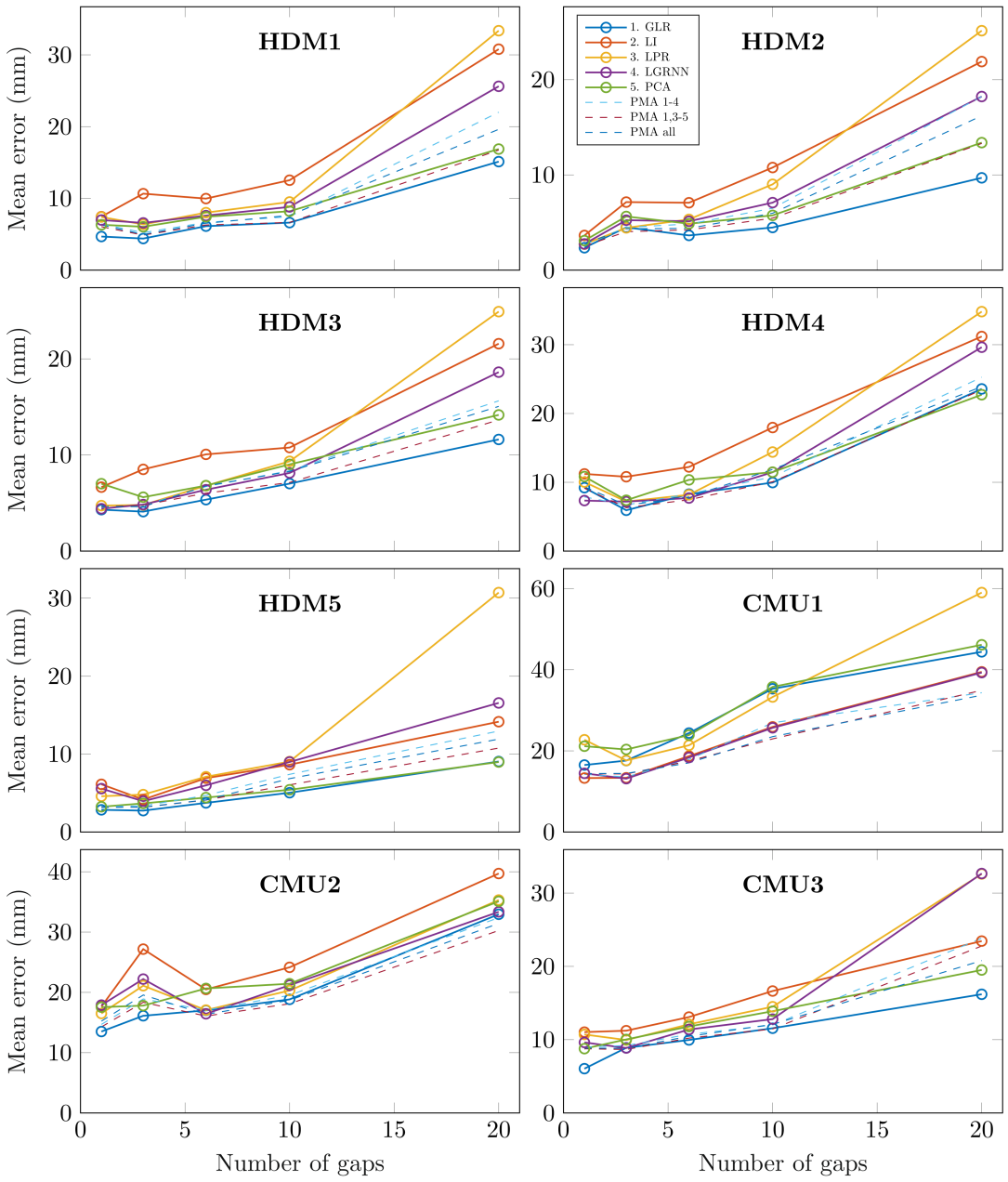

Figure 5.6: Mean recovery error for different numbers of missing markers and gap recovery methods. Each point represents the mean of recovery errors computed over 20 iterations of a number of randomly created gaps of 1 second (1, 3, 6, 10 or 20 gaps). Solid lines show results for each individual method. Dashed lines show results for distance-probability averages of various methods combinations.

Sequence	$\bar{\epsilon}$**, no constraint (mm)**	$\bar{\epsilon}$**, constraint (mm)**	**Difference (mm)**	**t-test p-value**
HDM1	40.9	13.5	27.4	$p < 10e-10$
HDM2	14.4	8.4	6.0	$p < 10e-3$
HDM3	32.2	10.8	21.4	$p < 10e-10$
HDM4	29.8	16.1	13.7	$p < 10e-8$
HDM5	25.0	8.6	16.4	$p < 10e-10$
CMU1	80.8	39.1	41.7	$p < 10e-10$
CMU2	27.7	26.2	1.5	$p = 0.16$
CMU3	22.1	19.6	2.6	$p = 0.02$

Table 5.3: Effect of constraints on the mean recovery error (t-test, n = 200; conditions: 10 gaps of 5 seconds). Paired t-test (n = 200) on constraints effect on PMA for the reconstruction of 10 simultaneous gaps of 5 seconds, introduced into different motion sequences. Individual methods 1 to 4 were used in this test.

Table 5.3 shows a similar analysis in a situation of low marker presence. In this case, 10 simultaneous gaps of 5 seconds were introduced into each motion sequence. We can see that in such situation, PMA's mean recovery error is much higher, and constraints always improve it significantly, up to 40 mm for CMU1.

5.3.5 Synthesis - mean results

As a synthesis, fig 5.7 shows the mean results of each method, obtained from the mean of the recovery errors on all the selected motion sequences. It can be seen that the various PMA combinations give more robust reconstruction regardless of the type of motion, the gap length (left graph), the duration of the motion sequence (center graph) and the number of incomplete marker trajectories (right graph). Among the individual methods, there is no clear difference of performance according to gap length. The local GRNN method seems more robust to gap length: it allows to recover three concomitant gaps of 5 seconds with a mean error of 10 mm. All other methods lead to a mean error above 15 mm. The local GRNN seems to be more robust to sequence duration. A duration of 5 seconds (with 41 markers, 120 fps) allows to train an effective model to reconstruct three concomitant gaps of 1 second with a mean error of 12 mm. However, for a longer sequence (40 seconds), GLR gives the best results, with a mean error of 9 mm. All individual methods are highly sensitive to the number of concomitant gaps, and thereby to the number of markers available to predict the missing trajectories. Finally, PMA systematically improves the gap recovery, independently of motion type, gap length, sequence duration or number of missing markers. As the local interpolation method seems to be the less effective, the best method combination is the averaging of methods 1, 3, 4 and 5.

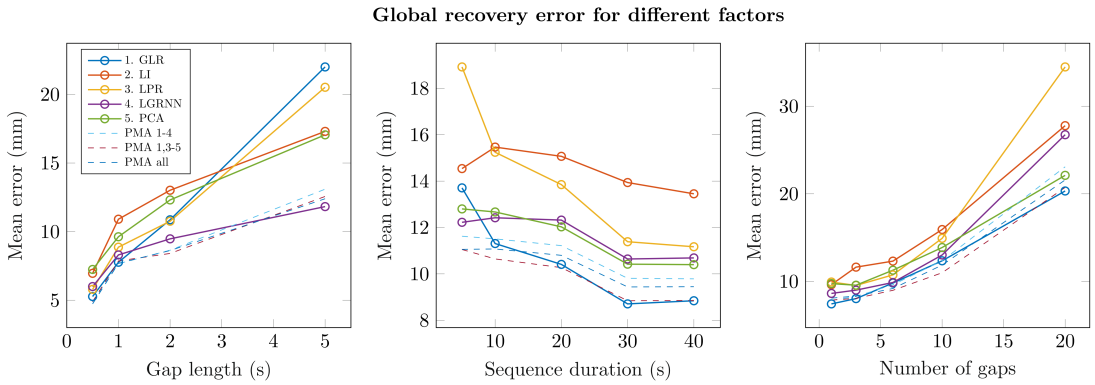

Figure 5.7: Mean recovery error for different recovery methods, for all test motion sequences. Left: different gap lengths (3 concomitant gaps, total sequence duration); Center: different motion durations (3 concomitant gaps of 1 second); Right: different numbers of concomitant gaps (gaps of 1 second, total sequence duration). Each point represents the mean of recovery errors computed over 20 iterations of a number of randomly created gaps. Solid lines show results for each individual method. Dashed lines show results for PMA with various individual methods combinations.

5.3.6 Visual results

Besides error minimization in MoCap data reconstruction, the visual result is another important criterion, especially if data are used for an animation purpose. Moreover, simulated gaps allow for a comparison of the different methods based on objective and quantitative results, but may not completely represent the ground truth. Gaps occurring in real situation may be more complicated to recover than randomly simulated ones. For instance, when a marker is occluded in a real situation, it is likely that the neighbor markers are also occluded. It is also very likely that a marker is recurrently occluded during a MoCap session, due to its placement on the body. Fig. 5.8 shows visual results of gap recovery with pchip interpolation (baseline), weighted PCA (Gløersen and Federolf, 2016) and our method, for three MoCap files originally recorded for various research and applications[2]. These motion sequences were selected as they are ground truth data, involving a high number of marker occlusions. A baseline method such as pchip interpolation (Fritsch and Carlson, 1980) could not recover missing data with sufficient accuracy for effective use of these files:

- A karate MoCap sequence with 25 optical markers, recorded at 250 fps (duration: 53 s, minimum marker presence: 86%, maximum number of simultaneous gaps: 12). This motion sequence involves sharp movements, difficult to recover with interpolation.
- A contemporary dance performance with a fall, recorded with 68 optical markers at 180 fps (duration: 57 s, minimum presence: 65%, maximum number of

[2]Data can be retrieved at: https://github.com/titsitits/MocapRecovery

simultaneous gaps: 26). This motion sequence involves long and simultaneous gaps on a large number of markers.

- A contemporary dance performance where the dancer performs a roulade on the ground, recorded with 68 optical markers at 180 fps (duration: 13 s, minimum presence: 34%, maximum number of simultaneous gaps: 30). This motion sequence has a poor quality for all markers throughout the whole sequence, as a roulade leads to many and frequent occlusions.

For better visualization, videos are supplied in supporting information (see the videos provided at https://github.com/numediart/MocapRecovery). All screen-shots and videos were created with the MotionMachine framework (Tilmanne and d'Alessandro, 2015).

Though objective results cannot be obtained in this context, we can observe that the movements reconstructed with our method generally seem realistic and respect movement constraints (for continuity, see the videos). Highly unrealistic results are obtained with pchip when gaps occur at the beginning or the end of a sequence. This is due to the fact that the recovered data are extrapolated and not interpolated.

The method proposed by Gløersen and Federolf (2016) obtained particularly bad results on the roulade motion sequence. This may be due to the fact that the method assigns weights to trajectories based on mean distance between markers. However, with roulade or similar movements, all limbs are gathered and their distance is reduced. In this case, distance standard deviation would be a more robust indicator of the relation between neighbor markers.

Videos show results with all individual methods, without constraints, and with PMA and spatial and temporal constraint. We can observe that PMA and the constraint improve the stability of the recovered movement, avoiding glitches and discontinuities in marker trajectories.

Finally, unrealistic results may appear for each method. In this case, it is possible that the error is due to corrupted original data, such as a swapping of marker trajectories. Several MoCap data softwares enable MoCap data post-processing, allowing to correct swapped trajectories, or simply to remove incoherent data. This step is hence recommended before recovery to obtain better results.

5.4 Discussion

Our PMA method presents several advantages compared to the available state of the art. It is fully automatic and does not require any prior knowledge or any pre-trained model. It can be used on MoCap data recorded with any marker set. Graphical

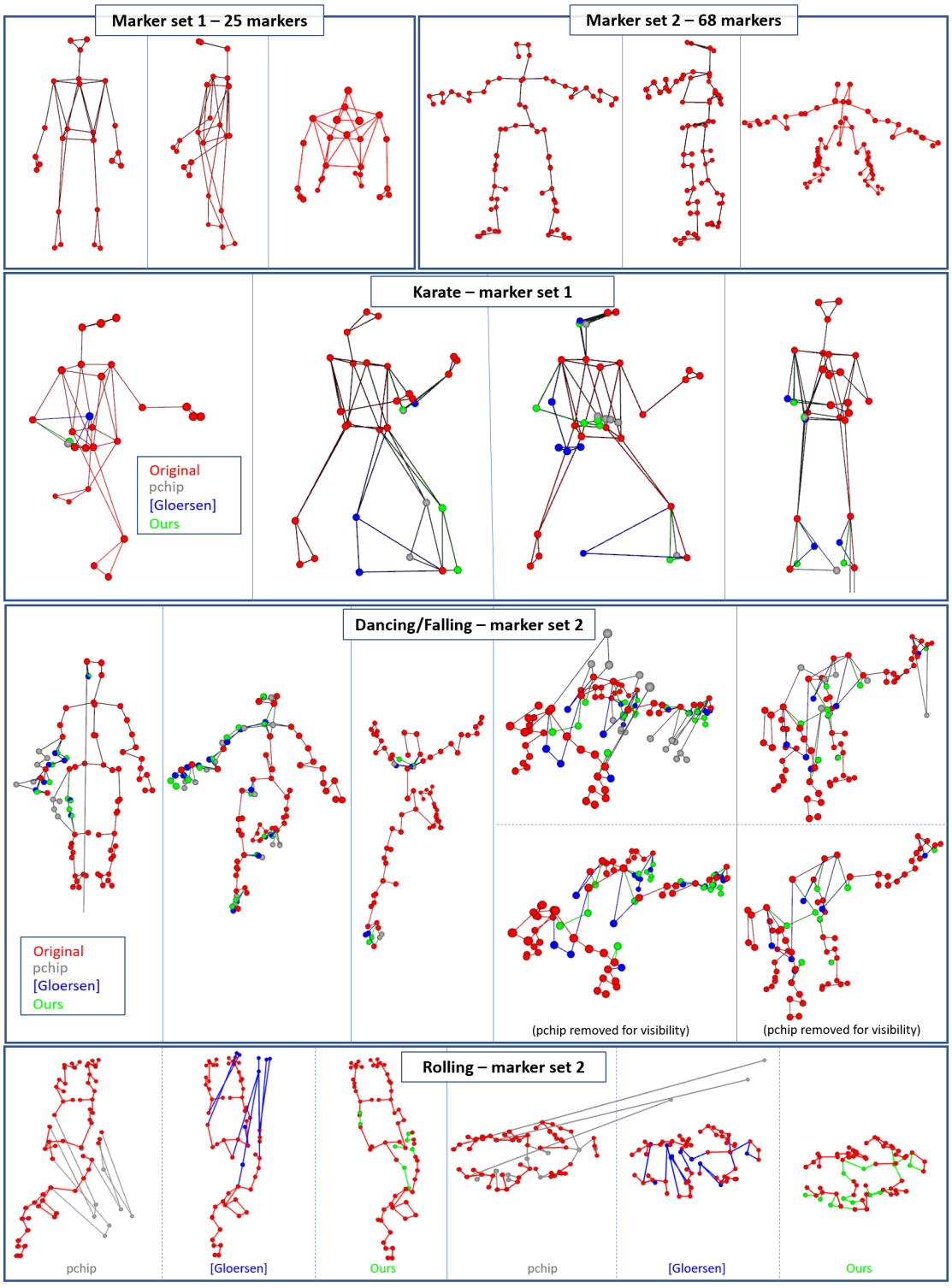

Figure 5.8: Visual comparison of different gap recovery methods on different motion sequences, with different marker sets. Red: original data. Gray: pchip interpolation (baseline) (Fritsch and Carlson, 1980). Blue: Gløersen and Federolf (2016). Green: our algorithm (PMA with constraints).

results show that PMA is robust to various factors, including gap length, sequence duration, the number of simultaneous gaps, and the type of motion. Additionally, the use of temporal and spacing constraints significantly improves the reconstruction, especially in challenging conditions (see Table 5.3).

Figs 5.4-5.7 show results where spacing and time constraints were applied to each individual method. These constraints may indeed be applied afterwards to any prediction method. The improvement of recovery after method combination is hence exclusively due to PMA, and confirms its effectiveness. In terms of quantitative results, no individual method shows better performance in general. All individual methods are more or less sensitive to the context, including motion type, gap length, sequence duration or the number of simultaneous missing markers. In contrast, PMA seems to take advantage of every individual method, improving the robustness of the recovery algorithm. Moreover, as averaging is based on distance with reference markers, PMA partly takes skeleton constraints into account.

Our methods could not be compared to some recently proposed methods, due to the unavailability of the code. However, the recent skeleton-constrained SVT method from Tan et al. (2015) (not included in our study), based on both skeleton constraints and low rank properties, achieved similar results to BoLeRo (Li et al., 2010): their improvement over BoLeRo mainly lies in execution time, as explained in Tan et al. (2015). Their constraint-fitting optimization method converges significantly faster than BoLeRo.

The effectiveness of the method proposed in the present study, including PMA and time and spacing constraints, is independent of the individual recovery model. It can theoretically be applied to any other set of individual recovery models in the future, possibly leading to better performance.

5.4.1 Limitations

The methods included in the present study rely on several parameters, including the threshold for reference marker selection in GLR, the smoothing parameter in LGRNN, parameters from w-PCA (Gløersen and Federolf, 2016), as well as confidence interval thresholds for spacing constraints. All these parameters were experimentally chosen in this research, as they gave the best results for the dataset tested (see Table 5.1). The user should adapt these parameters for her/his own data if necessary. Optimal parameters could depend on MoCap data, including for instance the number of recorded markers, their particular placement, the complexity or speed of the motion, data accuracy, or noise due to marker vibrations or camera quality. For instance, a larger threshold for the linear regression model could be considered for MoCap data with fewer available markers, and the optimal smoothness parameter for LGRNN could depend on the smoothness of the motion itself.

For the validation of the proposed methods, gaps were introduced into motion sequences at random locations. It is possible that in some particular cases, a marker can be isolated, without any highly related reference. If this marker is missing, it could lead to a poor recovery. This issue thus depends on the placement of markers It is hence relevant to consider this aspect when defining marker placement, to avoid isolated markers. On the other hand, markers placed too close to each other risk to be occluded simultaneously. A trade-off must thus be considered for their placement.

PMA has some limitations in comparison to a method such as BoLeRo (Li et al., 2010). In case of a blackout, i.e. when all markers disappear at the same time, a method based on a predictive filter such as Kalman filter can reconstruct an entire frame, and then use gradient descent or a similar optimization method to fit skeleton constraints, whereas PMA needs at least three present markers as references to evaluate distance probabilities. However, this is an extreme case, which can generally be avoided with an efficient use of the MoCap system.

5.4.2 Improvement prospects

Our methods could be improved in various ways. First, human motion is not a stationary process (Chiari et al., 2005). Each individual model might be made more efficient by giving more importance to motion data that are close to the gap to reconstruct. For instance, for each gap, a local model could be trained on a limited time window centered on that gap. The distance probabilities could also be locally defined on a time window. However, this would limit the number of available data for model training.

Secondly, we could use a more complex constraint fitting method, making use of a dynamic model such as the Kalman filter (Kalman et al., 1960) to ensure trajectory continuity. Additionally, an optimization procedure could be used instead of projection for skeleton constraints fitting whenever the inter-marker distance is outside the confidence interval. However, this could significantly increase execution time.

Finally, an original interest in the distance variation density estimation is the possibility to assess the quality of the reconstruction. It could further be used as an indication to identify and verify the most sensitive parts of the data, and possibly reject them and reprocess them with another configuration or method.

5.4.3 Processing time consideration

It is interesting to note that in human motion data, adjacent frames are very similar if the frame rate is high enough. In this case, data can be easily subsampled without losing much information for model training. This subsampling can drastically

	Nb. files	Nb. frames	Incomplete frames	Incomplete trajectories	Missing positions
Average	N/A	38131	5177	10.73	6417
Maximum	N/A	61663	17804	21	23788
Total	104	3965668	538396	1116	667408

Table 5.4: Taijiquan MoCap dataset recovery.

decrease computation time, either for computing reference weights (Eq. 5.1), for individual model training, as well as for kernel smoothing density estimation (Eq. 5.21).

Though it is not the initial goal of the proposed algorithm, each individual method based on regression, as well as their combination with PMA could be adapted for real-time purpose. Each individual model and distance distribution estimation can be trained on previously recorded data, and can be effective after a few seconds of recording. In this case, the time constraint would be limited to information about previous data. A Kalman filter would be appropriate for this task.

5.5 Taijiquan dataset recovery

In the present thesis, the proposed method has been used to process the Taijiquan dataset presented in Chapter 4. This process ensures the use of high-quality data in the next steps of the framework proposed in this thesis for gesture evaluation.

The method was applied on the unsegmented data files, allowing a finer modeling of the distances between markers, as well as a more efficient training of the various algorithms used in the method. As illustrated in Fig 5.7, a longer motion sequence allows a more robust recovery. Moreover, the data recovery process was applied on the positions of the 68 surface marker presented in Table 4.2, before the extraction of joint positions and orientations with Visual3D^{TM}. A larger number of markers implies more highly related trajectories, allowing a finer recovery of the missing marker trajectories. Table 5.4 summarizes the recovery of missing data in the Taijiquan dataset. The dataset consists of 104 unsegmented files of 38131 frames in average (one file per recording). The average number of incomplete frames per file was 5177, and the average number of incomplete trajectories per file was 10.73. For the entire dataset, a total of 667408 positions were missing and recovered with the proposed method. The joint positions and orientations were then extracted from the recovered data.

5.6 Conclusion

In this chapter, we proposed an original automatic method, Probabilistic Model Averaging (PMA), for robust reconstruction of missing MoCap data. The robustness of our method relies on two major steps:

1. The weighted combination of several models, based on the posterior likelihoods of inter-marker distances.
2. The application of simple but effective constraints, enforcing trajectory continuity and plausible distance of reconstructed trajectories with related markers.

To support and validate our model-averaging method, several reconstruction methods based on regression and local coordinates were proposed, and were found to compete with state-of-the-art methods. Results show that PMA used with the constraints outperforms individual methods in various conditions, including various gap lengths, motion sequence durations and numbers of simultaneous gaps.

Our method has the advantage of being fully automatic. The algorithm is data-driven, and does not need any prior knowledge or any pre-trained model. Moreover, the model averaging and the proposed constraints are general and can be used with any other individual reconstruction method, leading to possible future improvement.

In the present thesis, the proposed method was used for the processing of the Taijiquan MoCap dataset.

Chapter 6

Taijiquan ergonomic principles: a new set of features

Contents

6.1 Introduction

In Chapter 2, we presented various features allowing a low-level or high-level representation of motion. Among these features, we presented the category of ergonomic features. Ergonomics is closely related to skill, as it is the study of motor control effectiveness while minimizing energy expenditure and risks of injury. However, no previous work known to the author used this type of feature for evaluation of expertise. Andreoni et al. (2009) proposed a method based on perceived discomfort (see Section 2.3.6.4) to automatically assess the ergonomy of a posture from MoCap data, showing a potential use of ergonomic features in quantitative motion quality assessment.

In this Chapter, present a new set of ergonomic features is presented, inspired by Taijiquan. One major component in the learning of Taijiquan is focused on the ergonomy of motion, as extensively developed in the work of the ergonomist and Taijiquan teacher Eric Caulier (Caulier, 2010). A collaboration with him led to the development of a new set of ergonomic features, inspired by principles largely taught in the school of Taijiquan Eric Caulier. The proposed feature set can be divided into four main categories, as presented in the following sections:

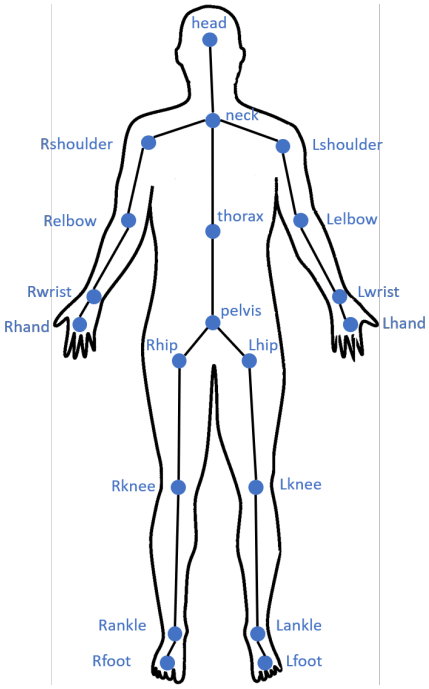

Figure 6.1: Body joint representation and naming convention.

- Stability (see Section 6.2)
- Joint Alignments (see Section 6.3)
- Favorable angles (see Section 6.4)
- Fluidity (see Section 6.5)

The body joints representation and naming convention used in the remaining of this chapter is illustrated in Fig 6.1.

6.2 Stability

In the practice of Taijiquan, the body must always remain stable during the motion, in terms of balance, and balance variation. The trunk must remain vertical, and segments between symmetric joints must remain in the horizontal plane. The CoM must be close to the pelvis, and above the support base. When a limb moves away from the CoM, another one must be used in synchrony as a counterweight to maintain balance, and reduce CoM effective quantity of motion. The heel kick technique is a good example of this notion, as illustrated in Fig 6.2. During the gesture, arms move in synchrony with the foot, and are used as counterweights for a better stability.

To evaluate stability, four features have been implemented:

Fig. 197 Donner un coup de pied.

Fig. 198 Donner un coup. Ouvrir coudes et genoux ensemble.

Fig. 199 Donner un coup de pied. Le bras droit est au-dessus de la jambe. Etendre bras et jambes ensemble.

Figure 6.2: Heel kick technique. During the gesture, arms move in synchrony with the foot, and are used as counterweights for a better stability. Reproduced from Caulier (2010).

- Static stability: the static stability is evaluated by computing the Euclidean distance between the components of the pelvis and the CoM in a horizontal plane (x, y):

$$F_1(t) = \sqrt{(x_{pelvis}(t) - x_{CoM}(t))^2 + (y_{pelvis}(t) - y_{CoM}(t))^2} \tag{6.1}$$

- Dynamic stability: the dynamic stability is computed as the time derivation of the static stability:

$$F_2(t) = \frac{dF_1(t)}{dt} \tag{6.2}$$

- Verticality: the verticality of the trunk is computed as the Euclidean distance between the components of the pelvis and the neck in a horizontal plane (see Fig 6.3 (a)):

$$F_3(t) = \sqrt{(x_{pelvis}(t) - x_{neck}(t))^2 + (y_{pelvis}(t) - y_{neck}(t))^2} \tag{6.3}$$

- Horizontality: the horizontality of the body is computed as the mean absolute difference between the height (z) of main joint pairs (shoulders, hips and knees, see Fig 6.3 (a)):

$$F_4(t) = \frac{\sum_j^3 \left| z_{L_j}(t) - z_{R_j}(t) \right|}{3} \tag{6.4}$$

where j stands for the joint index in $L = \{Lshoulder, Lhip, Lknee\}$ and $R = \{Rshoulder, Rhip, Rknee\}$.

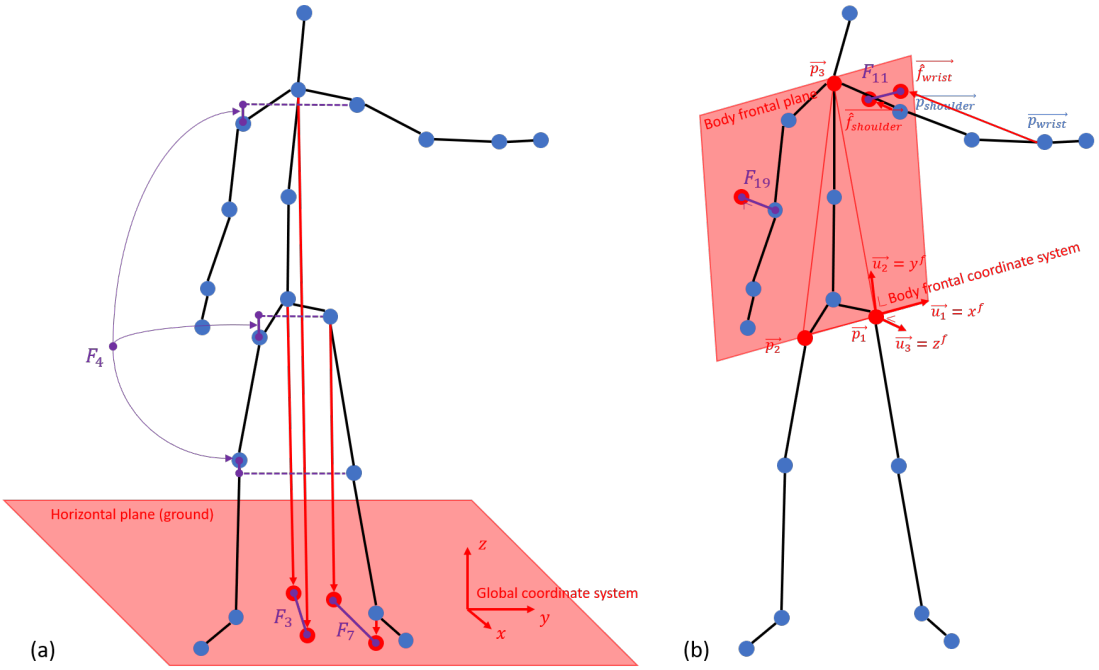

Figure 6.3: Visualization of some stability and alignment features inspired by Taijiquan ergonomic principles. (a): F_3 (verticality), F_4 (horizontality) computed in the horizontal plane, and F_7 (vertical alignment of left hip and left ankle). (b): F_{11} (frontal alignment of left shoulder and left wrist) and F_{19} (right elbow not behind body) computed in the body frontal plane.

Figure 6.4: Joint alignments in tree posture (Wuji). Reproduced from Caulier (2010).

6.3 Joint alignments

Joints must as often as possible remain aligned, vertically or horizontally. Fig 6.4 illustrates this notion. Joint alignments allow a better stability, but also a better transmission of the forces from the ground to the body extremities.

To evaluate joint alignments, nine features have been proposed:

- Joints vertical alignments: computed as the Euclidean distance between the horizontal components of two joints:

$$F_{4+j}(t) = \sqrt{(x_{U_j}(t) - x_{D_j}(t))^2 + (y_{U_j}(t) - y_{D_j}(t))^2} \tag{6.5}$$

  for $U = \{Lshoulder, Rshoulder, Lhip, Rhip, Lknee, Rknee\}$ and
  $D = \{Lhip, Rhip, Lankle, Rankle, Lfoot, Rfoot\}$.

- Shoulder-wrist frontal alignment: computed as the Euclidean distance between the coordinates of the left (resp. right) wrist and the left (resp. right) shoulder into the body frontal plane. At each time, the body frontal plane is defined by three points placed on both hips ($\overrightarrow{p_1}$and $\overrightarrow{p_2}$) and the neck ($\overrightarrow{p_3}$). A local coordinate system (referred to as *body frontal coordinate system* below, see Fig 6.3

(b)) is then defined using these points:[1]

$$\overrightarrow{v_1} = \overrightarrow{p_1} - \overrightarrow{p_2}, \quad \overrightarrow{u_1} = \frac{\overrightarrow{v_1}}{||\overrightarrow{v_1}||} \tag{6.6}$$

$$\overrightarrow{v_3} = \overrightarrow{v_1} \times (\overrightarrow{p_3} - \overrightarrow{p_1}), \quad \overrightarrow{u_3} = \frac{\overrightarrow{v_3}}{||\overrightarrow{v_3}||} \tag{6.7}$$

$$\overrightarrow{u_2} = \overrightarrow{u_3} \times \overrightarrow{u_1} \tag{6.8}$$

where $\times$ is the cross product, $\overrightarrow{u_1}$, $\overrightarrow{u_2}$and $\overrightarrow{u_3}$ are three orthonormal vectors: $\overrightarrow{u_1}$ and $\overrightarrow{u_2}$ are in the body frontal plane, and $\overrightarrow{u_3}$ is normal to the body frontal plane (see Fig 6.3 (b)). The origin of the *body frontal coordinate system* is defined as $\overrightarrow{p_1}$. The coordinates of a joint in the *body frontal coordinate system* are thus defined as:

$$P = \quad [\overrightarrow{u_1}^T \ \overrightarrow{u_2}^T \ \overrightarrow{u_3}^T] \tag{6.9}$$

$$\overrightarrow{f_j} = (\overrightarrow{p_j} - \overrightarrow{p_1}) \cdot P \tag{6.10}$$

$$\overrightarrow{\hat{f}_j} = (x_j^f, y_j^f, 0) \tag{6.11}$$

where we replace the third coordinate (corresponding to $\overrightarrow{u_3}$) by 0. Finally, the frontal alignment is computed as Euclidean distance between the components of the shoulder and the wrist in the body frontal plane:

$$F_{11} = \left\| \overrightarrow{\hat{f}_{Lshoulder}} - \overrightarrow{\hat{f}_{Lwrist}} \right\| \tag{6.12}$$

$$F_{12} = \left\| \overrightarrow{\hat{f}_{Rshoulder}} - \overrightarrow{\hat{f}_{Rwrist}} \right\| \tag{6.13}$$

- Feet alignment: feet must often be parallel in Taijiquan. To assess this alignment, the absolute difference is computed between the Euclidean distance between heels and the Euclidean distance between toes:

$$F_{13} = \left| \left\| \overrightarrow{p_{Lankle}} - \overrightarrow{p_{Rankle}} \right\| - \left\| \overrightarrow{p_{Lfoot}} - \overrightarrow{p_{Rfoot}} \right\| \right| \tag{6.14}$$

6.4 Favorable angles

Each body joint has an optimal flexion, leading to both suppleness and robustness. Joints must never be fully stretched nor too bent. An equilibrium must be found between stretching and tenseness at any time of the motion. This notion is illustrated in Fig 6.5. This property can be related to optimal muscle lengths allowing the highest force production: the force production of a muscle is at its minimum when it is fully stretched or tense, as shown in Fig 6.6 (Gordon et al., 1966).

[1]Note that the time variable t is not indicated for readability.

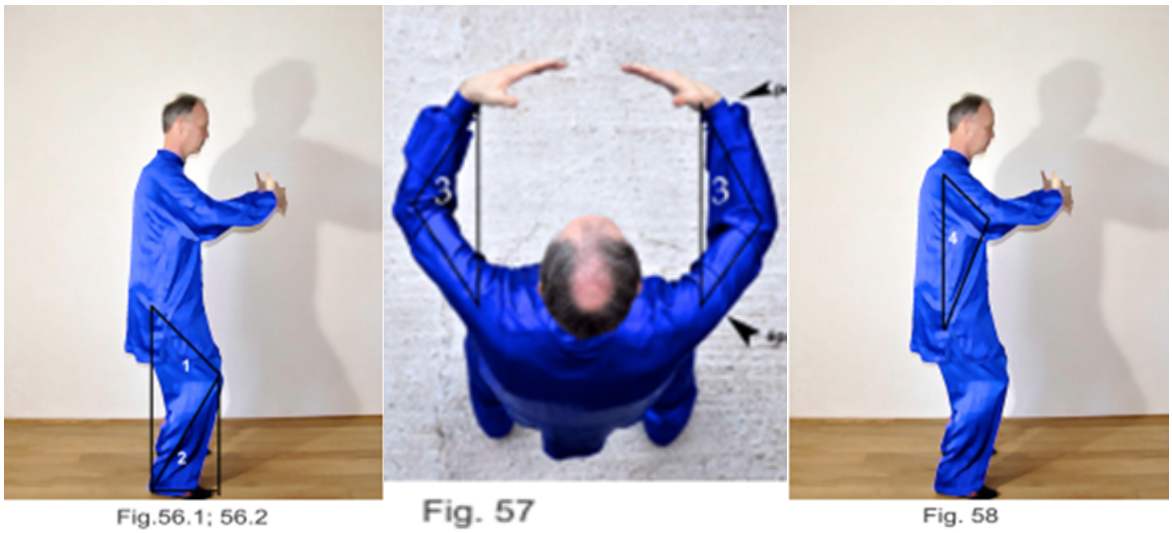

Figure 6.5: Favorable joint angles in tree postures (Wuji). No joint is fully stretched nor fully bent. Reproduced from (Caulier, 2010).

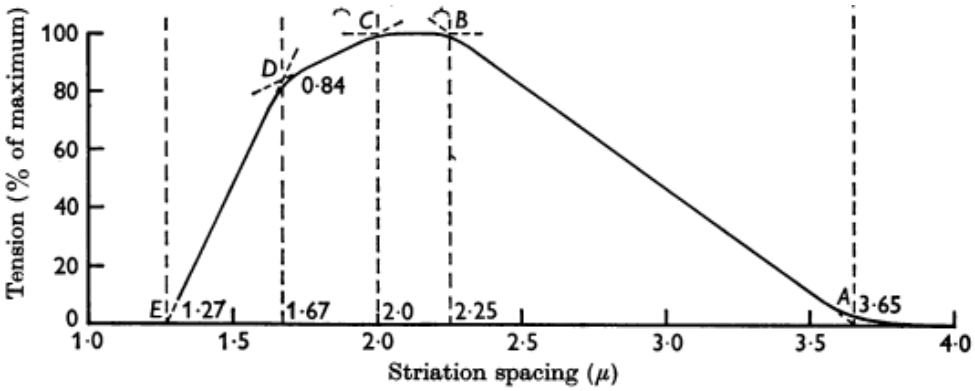

Figure 6.6: Length-tension relation of a sarcomere in a muscle fiber. Reproduced from (Gordon et al., 1966).

To evaluate favorable angles, eight features have been proposed:

- The shoulders must be low. To avoid useless tension and stiffness of the shoulders, they must remain as low as possible during the gesture. To evaluate if the shoulders are low, the angle between the shoulder, the neck and the thorax is extracted. The angle between three joints j, k and l is calculated as:

$$\overrightarrow{v_{j,k}} = \overrightarrow{p_j} - \overrightarrow{p_k} \tag{6.15}$$

$$\overrightarrow{v_{l,k}} = \overrightarrow{p_l} - \overrightarrow{p_k} \tag{6.16}$$

$$(\widehat{j,k,l}) = \frac{\arccos(\overrightarrow{v_{j,k}} \cdot \overrightarrow{v_{l,k}})}{\|\overrightarrow{v_{j,k}}\| \times \|\overrightarrow{v_{l,k}}\|} \tag{6.17}$$

  The angles between shoulders, the neck and the thorax are obtained using eq. 6.17, leading to (see Fig 6.7):

$$F_{14} = (Lshoulder, \widehat{neck}, thorax) \tag{6.18}$$

$$F_{15} = (Rshoulder, \widehat{neck}, thorax) \tag{6.19}$$

- Elbow flexion deviation from optimal angle: elbow flexion angle should remain approximately between 90° and 135°. To evaluate the quality of the elbow flexion (computed as the angle between the wrist, the elbow and the shoulder), the optimal angle is estimated as 112.5°, and the deviation from this angle is calculated as (see Fig 6.7):

$$F_{16} = \left|112.5° - (Lwrist, \widehat{Lelbow}, Lshoulder)\right| \tag{6.20}$$

$$F_{17} = \left|112.5° - (Rwrist, \widehat{Relbow}, Rshoulder)\right| \tag{6.21}$$

- The elbows must not be behind the body. To evaluate if the elbow is behind the body, the elbow z-coordinate in *body frontal coordinate system* is extracted (cf. eq. 6.10, see Fig 6.3):

$$F_{18} = z^f_{Lelbow} \tag{6.22}$$

$$F_{19} = z^f_{Relbow} \tag{6.23}$$

- The elbows must not be too low (against the body), nor too high. The optimal abduction of the elbow (estimated as the angle between the elbow, the shoulder and the hip) is about 67.5°, and the deviation from this angle is calculated as (see Fig 6.7):

$$F_{20} = \left|67.5° - (Lelbow, \widehat{Lshoulder}, Lhip)\right| \tag{6.24}$$

$$F_{21} = \left|67.5° - (Relbow, \widehat{Rshoulder}, Rhip)\right| \tag{6.25}$$

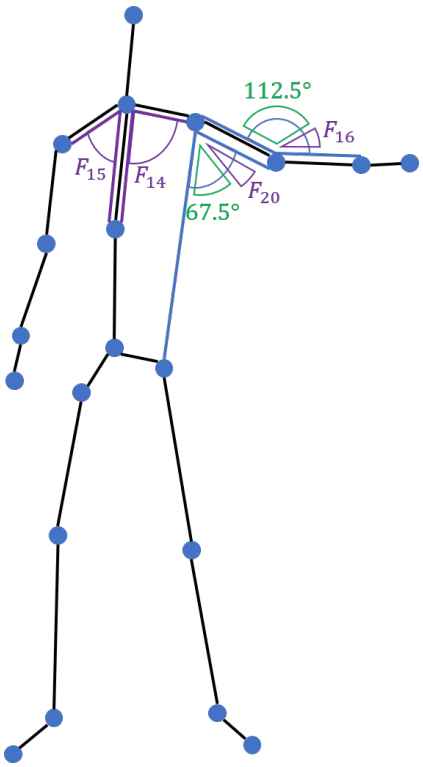

Figure 6.7: Some favorable angles, inspired by Taijiquan ergonomic principles.

6.5 Fluidity

During a gesture, the body must always be in motion, in order to keep a continuous kinetic energy, to keep joints warm and supple. On the other hand, the motion must remain smooth, avoiding jerks: jerky motions lead to more impact on the joints, and more energy expenditure. To evaluate the fluidity of each limb (arms and legs) and of the trunk, the normal speeds, accelerations and jerks of their CoMs are computed:

$$\overrightarrow{S_{Limb_j}}(t) = \frac{d\overrightarrow{CoM_{Limb_j}}(t)}{dt}, \quad F_{21+j}(t) = \left\|\overrightarrow{S_{Limb_j}}(t)\right\| \tag{6.26}$$

$$\overrightarrow{A_{Limb_j}}(t) = \frac{d\overrightarrow{S_{Limb_j}}(t)}{dt}, \quad F_{26+j}(t) = \left\|\overrightarrow{A_{Limb_j}}(t)\right\| \tag{6.27}$$

$$\overrightarrow{J_{Limb_j}}(t) = \frac{d\overrightarrow{A_{Limb_j}}(t)}{dt}, \quad F_{31+j}(t) = \left\|\overrightarrow{J_{Limb_j}}(t)\right\| \tag{6.28}$$

where $\overrightarrow{S_{Limb_j}}(t)$, $\overrightarrow{A_{Limb_j}}(t)$, and $\overrightarrow{J_{Limb_j}}(t)$ are respectively the speed, acceleration and jerk of the CoM of $Limb_j$, in $Limb = \{Larm, Rarm, Lleg, Rleg, Trunk\}$.

6.6 Summary and conclusion

In this chapter, a new feature set was presented, inspired by Taijiquan ergonomic principles. This set of 36 features can be divided into four main types: stability, joint alignments, favorable angles and fluidity features. A summary is provided in Table 6.1.

Although these features are inspired by Taijiquan specific rules, most of them are generic and can be applied to other disciplines. However, according to the specific discipline, some features are more relevant than others. Some gestures might require specific alignments or joint angles. For instance, in a tennis serve, the tossing arm is usually highly extended above the shoulder, and the striking arm is placed behind the body during the preparation step, discrediting F_{18}/F_{19} (elbow bot behind the body) and F_{20}/F_{21}(elbow abduction around 67.5°). Nonetheless, to optimize the gesture, experienced players will not exaggerate these positions. They will bend their knees to avoid an exaggerate abduction of the tossing-arm elbow, and their body will be in profile (i.e. not facing the court) to avoid a position of the striking-arm elbow behind the body. In other words, it is assumed that most of these ergonomic features must generally be respected, to the extent allowed by the specific constraints of the gesture. For validation purpose, this feature set will be tested in Part III, either for gesture evaluation (see Chapters 8 and 9), and for gesture feedback (see Chapter 10).

Index	Definition	Computation
Stability		
F_1	Static stability	Distance between the horizontal projections of CoM and pelvis.
F_2	Dynamic stability	Time-derivation of static stability.
F_3	Verticality	Distance between the horizontal projections of neck and pelvis.
F_4	Horizontality	Mean of vertical distances of joint pairs (shoulders, hips, knees)
Alignments		
F_5 / F_6	Shoulder-hip vertical alignment	Distance between the horizontal projections of shoulder and hip.
F_7 / F_8	Hip-ankle vertical alignment	Distance between the horizontal projections of hip and ankle.
F_9 / F_{10}	Knee-foot vertical alignment	Distance between the horizontal projections of knee and foot.
F_{11} / F_{12}	Wrist-shoulder frontal alignment	Distance between the frontal projections of shoulder and wrist.
F_{13}	Feet alignment	Difference between length of segments $(Lankle - Lfoot)$ and $(Rankle - Rfoot)$.
Favorable angles		
F_{14} / F_{15}	Shoulders low	Angle between shoulder, neck and thorax.
F_{16} / F_{17}	Elbows optimal flexion	Deviation from 112.5° of the angle between wrist, elbow and shoulder.
F_{18} / F_{19}	Elbows not behind body	Elbow z-coordinate in *body frontal coordinate system*.
F_{20} / F_{21}	Elbows optimal abduction	Deviation from 67.5° of the angle between elbow, shoulder and hip.
Fluidity		
$F_{22} - F_{26}$	Limb speed	Normal speed of limb CoM (arms, legs and trunk).
$F_{27} - F_{31}$	Limb acceleration	Normal acceleration of limb CoM (arms, legs and trunk).
$F_{32} - F_{36}$	Limb jerk	Normal jerk of limb CoM (arms, legs and trunk).

Table 6.1: Features inspired by Taijiquan ergonomic principles.

Chapter 7

Morphology-independent residual feature extraction (MIRFE)

Contents

7.1 Introduction

The following is partly reproduced from Tits et al. (2017).

In Chapter 2, various features were presented, allowing a representation of different aspects of motion. Chapter 6 presented a new feature set inspired by Taijiquan ergonomic principles, aiming for a relevant description of expertise in sports gestures. More features means more information about the various aspects of motion, and possibly about expertise. However, it also adds complexity in the data, due to the possible noise, redundancy, as well as irrelevant information present in the features. Motion features can indeed be influenced by many factors, including psychological, social or physiological factors. As a consequence, these features are therefore not

optimal for expertise modeling. To solve this issue, a common step in machine learning research is the use of feature selection techniques, allowing for the selection of a subset of features keeping only the most relevant and reliable information regarding the targeted task (in the present case, expertise modeling). The problem of feature selection is that it discards all information contained in the non-selected features. Another solution is to post-process the features, in order to reduce the influence of some factors on them, and hence to improve the information related to the targeted factor.

Morphology is a factor that has a direct influence on motion, making comparison between gestures of different individuals difficult. For instance, during a kick gesture the foot of a tall person will generally move higher than the foot of a short person. On the contrary, if a particular height of the kick is aimed, then the hip angle of the taller person will be smaller. In both cases, some features of both individuals will be very different (either foot height, or hip angle), without any indication about the quality of the performance, and making the comparison between gestures difficult. In this sense, the information contained in any feature about morphology can be considered as noise and should therefore be reduced as much as possible. Moreover, this information is generally redundant as it is contained in many features.

To alleviate this issue, different motion data representations have been proposed. Sie et al. (2014) proposed a simple skeleton scaling method, by placing the coordinate system on a reference node of the body (i.e. on the pelvis), and dividing all nodes coordinates by the torso height. Features can then be extracted on these scaled data. This method was later used by Morel et al. (2016) for gesture evaluation. It has the advantage of being very simple, but has many limitations. It is based on the simplistic hypothesis that the movement of a short individual should be an homothety of that of a tall individual. However, weight, height of the center of mass, shoulder width and hip width, among others, may also influence movement in different ways including inertia, balance, speed and power. These characteristics will be altered by such a basic scaling.

As described in Chapter 2 (see Section 2.3.3), Müller et al. (2005) proposed a specific feature set based on 40 binary relational features, originally developed for whole-body motion classification and retrieval. These relations may for instance correspond to a foot being raised, a hand being in front of the body, legs being crossed, etc. The thresholds for the binary decisions are defined by different body segment lengths such as the humerus length or the shoulder width, so that each feature is scaled by a custom pre-defined body characteristic. However, this method is limited to this type of feature, and does not allow the extraction of new features afterwards.

Kulpa et al. (2005) developed a morphology-invariant representation of motion, originally developed for animation, where they defined limbs with variable lengths. Each limb (legs and arms) is defined by the position of its end-effector and by a plane where the middle joint (knee and elbow) is located. The spine is represented as a

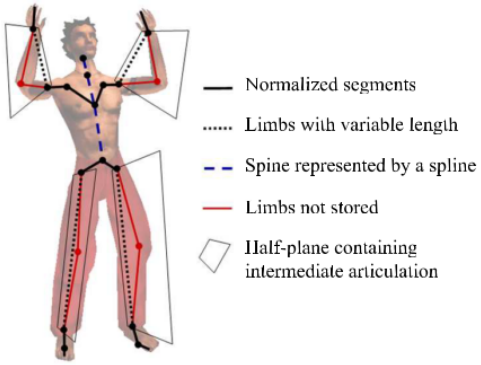

Figure 7.1: Kulpa et al. (2005) method for morphology-invariant representation of motion. Reproduced from Kulpa et al. (2005).

spline (see Fig 7.1). This representation allows reconstruction of the movement to fit specific constraints. However, it does not fully store the actual movement, and it modifies it to fit these constraints. It is relevant for animation and motion retrieval, but is not suited for movement analysis, which can require details of the movement that are lost in this representation.

In this Chapter, a new method is proposed for reducing the influence of morphology on any motion feature. It estimates and removes the correlation between a feature and a morphology factor, avoiding direct manipulation of the spatial skeleton data that would alter other body characteristics. The estimated relationship is based on a linear regression of individual means and standard deviations of features with a morphological factor. The proposed method can be seen as a tuning of each feature, independently removing the influence of morphology on each feature post-processed through our method. It is also more general than the related work, as it could theoretically be used with any factor and on any feature, whereas the literature known to the author is limited to body size.

7.2 Method

7.2.1 Morphology Independent Residual Feature Extraction (MIRFE)

The objective of the proposed method is to remove from the feature distribution the component resulting from the influence of inter-individual factors such as morphology. In order to assess the influence of a morphological factor on a feature, the best linear combinations of this factor to approximate the statistics of the feature are identified. For that purpose, constrained linear regression is used.

Let $x_n(t)$ be a temporal feature extracted from a motion sequence $n \in \{1, ..., N\}$, in a dataset of N motion sequences. Let $\mu(n) = mean(x_n(t))$ and $\sigma(n) = std(x_n(t))$ be

respectively the mean and standard deviation of this feature over time. Both of these statistics can be considered as non-temporal features.

Let $m(n)$ be a variable corresponding to a morphological factor, known for each motion sequence n. For each statistic (μ, σ), a linear regression is performed with the regressor m, leading to the following predictive equations:

$$\mu_{pred}(n) = \beta_{0,\mu} + \beta_{1,\mu} \cdot m(n) \tag{7.1}$$
$$\sigma_{pred}(n) = \beta_{0,\sigma} + \beta_{1,\sigma} \cdot m(n) \tag{7.2}$$

where $\beta_{0_{\mu,f}}$ and $\beta_{1_{\mu,f}}$ (resp. $\beta_{0_{\sigma,f}}$ and $\beta_{1_{\sigma,f}}$) are the intercept and slope of the linear regression of means μ_f (resp. individual standard deviations σ_f).

To ensure positive (and hence physically meaningful) predictions for the standard deviation (σ_{pred}) , the slope parameter ($\beta_{1,\sigma}$) is constrained:

$$\begin{aligned} n_{min} &= argmin_n(\sigma_{pred}(n)),\ n \in \{1, ..., N\} \\ if\ \sigma_{pred}(n_{min}) < 0 :\ \beta_{1,\sigma} &= \frac{-\beta_{0,\sigma}}{m(n_{min})} + \epsilon \end{aligned} \tag{7.3}$$

where ϵ is a small positive number (arbitrarily defined as 0.0001 in our work). If the lowest value of σ_{pred} is negative, the application of the constraint on $\beta_{1,\sigma}$ defined in Eq. 7.3 to Eq. 7.2 leads to:

$$\sigma_{pred}(n_{min}) = \beta_{0,\sigma} + (\frac{-\beta_{0,\sigma}}{m(n_{min})} + \epsilon) \cdot m(n_{min}) = \epsilon \tag{7.4}$$

The method is then based on the hypothesis that the prediction of the linear regression is the part of the statistic that can be fully described by the factor, while the residue is the uncorrelated part of the statistic, i.e. the part that is not influenced by the morphological factor. The residues μ_{res} and σ_{res} of the predictions are expressed as:

$$\mu_{res}(n) = \mu(n) - \mu_{pred}(n) \tag{7.5}$$
$$\sigma_{res}(n) = \frac{\sigma(n)}{\sigma_{pred}(n)} \tag{7.6}$$

The following equation is then used to compute a version of the feature x_n corresponding to these residual statistics, where the influence of morphology has been removed:

$$x_{res,n} = \frac{(x_n - \mu(n))}{\sigma(n)} \cdot \sigma_{res}(n) + \mu_{res}(n) \tag{7.7}$$

Eq. 7.7 can be interpreted as follows: for each motion sequence, the feature x_n is first standardized to remove its initial mean ($\mu(n)$) and standard deviation ($\sigma(n)$), and then scaled and translated so that its mean and standard deviation correspond to $\mu_{res}(n)$ and $\sigma_{res}(n)$.

Figure 7.2 illustrates the extraction of such residues, on an example dataset consisting of six motion sequences ($N = 6$), each one from a different participant. This number was arbitrarily used for illustration, although a larger number of participants should be recorded to avoid overfitting of the linear regression. The first graph (a) displays a dummy feature for the 6 sequences, with means and standard deviations (μ and σ) for each sequence. The second graph (b) shows the morphological factor m, representing in this example the size of the participant for each sequence. The third and fourth graphs (c and d) respectively show the results of linear regression of μ and σ with the regressor m. The blue curve is the regressand (μ or σ), the red curve is the prediction (Eq. 7.1 and 7.2), and the green curve is the residue (Eq. 7.5 and 7.6). The final graph (e) then displays the result of the residual feature extraction, where means (μ_{res}) and standard deviations (σ_{res}) are now independent of the morphology (Eq. 7.7).

When the residual features are extracted, a correlation analysis can be performed with the factor of interest, e.g. skill. As skill is supposedly not correlated to morphology, it is assumed that the process will not decrease the correlation between features and skill, but on the contrary, could increase it. Moreover, as the same information is removed from the features, their redundancy should decrease, allowing a more efficient use in machine-learning based modeling in general.

If the features were standardized before the process, they should be standardized again afterwards, as the overall mean and standard deviation may have changed. In the remainder of the paper, our method will be referred to as *Morphology Independent Residual Feature Extraction* or MIRFE.

7.2.2 Experiments

The goal of the MIRFE method is to reduce the influence of the morphology on motion features, while preserving information about other independent factors (i.e. expertise in the present research). To verify the effectiveness of the method, the relation between features and factors can be analyzed, with or without the use of this method, and results can be compared with a baseline method, i.e. skeleton data scaling (Sie et al., 2014). In the skeleton scaling, all global joint coordinates are divided by the size of the individual before feature extraction. As a benchmark, the eight Bafa techniques of the Taijiquan dataset presented in Chapter 4 are used (see Table 4.3). Different types of features are extracted from this dataset, including:

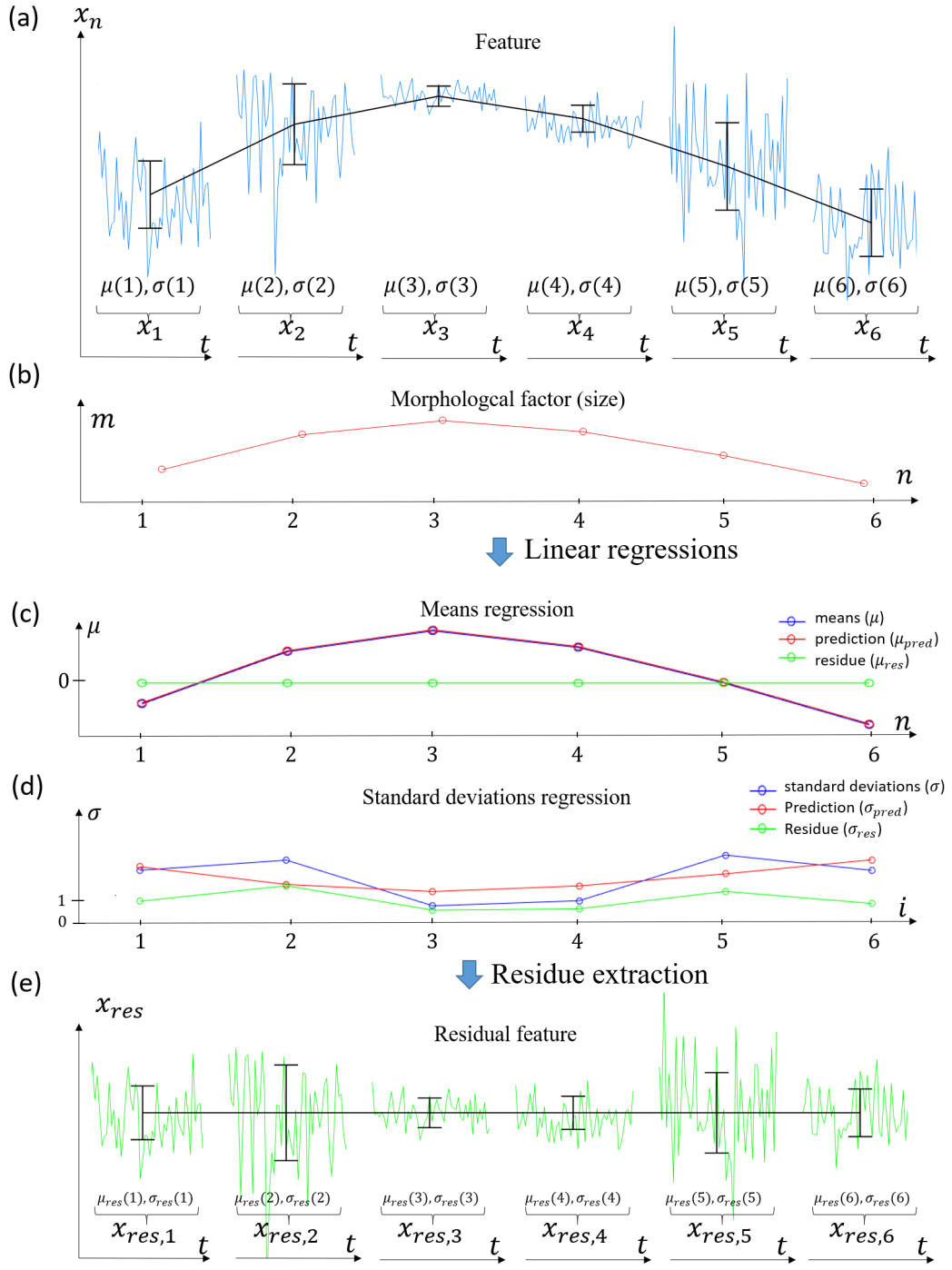

Figure 7.2: Inter-individual factor independent residual feature extraction. (a): feature and statistics (μ and σ). (b): individual morphology (size). (c) and (d): linear regression of means and standard deviations. The blue curve is the regressand (μ or σ), the red curve is the prediction (Eq. 7.1 and 7.2), and the green curve is the residue (Eq. 7.5 and 7.6). (e): residual feature extraction (Eq. 7.7).

1. Joint 3D global coordinates (reference placed on the hips) ($F = 61$)[1]
2. Joint 3D local coordinates (reference placed on each parent joint) ($F = 53$)[2]
3. Joint global quaternions (reference placed on the hips) ($F = 64$)[3]
4. Joint local quaternions (reference placed on each parent joint) ($F = 64$)
5. Continuous relational features (Müller feature set without thresholding, see Section 2.3.3) ($F = 40$)
6. Ergonomic features ($F = 93$), including :
    - ROM (in degrees, see Table 2.1) ($F = 32$)
    - Joint perceived discomfort (see Section 2.3.6.4) ($F = 14$)[4]
    - CoM kinematics (including CoM 3D coordinates, 3D speeds, 3D accelerations, as well as normal speed and acceleration) ($F = 11$)
    - Taijiquan ergonomic principles (see Chapter 6) ($F = 36$)

The entire feature set hence comprises $F = 375$ features.

To analyze the relations between features statistics (μ and σ) and factors (morphology m and skill s), absolute correlation will be used:

$$\Phi_{a,b} = |R(a,b)| \tag{7.8}$$

where $R(a,b) \in [-1,1]$ denotes the correlation between a statistic variable a and a factor variable b, and $\Phi_{a,b} \in [0,1]$ denotes their absolute correlation**.** The mean of the absolute correlations ($\overline{\Phi_{a,b}}$) for the feature sets will then be extracted and compared according to the method used, i.e. (i) without processing, (ii) with skeleton scaling and (iii) with MIRFE.

Not only the removal of morphology influence could allow for a improvement of the relation with other factors, but it could also reduce the redundancy of the features, as the same redundant information is removed from them. To analyze the redundancy of the motion features in a feature set, the mean of their absolute pairwise correlations will be computed, and will be denoted as $\overline{\Phi_f}$ (and referred to below as *redundancy*), where the index $f \in (1,...,6)$ stands for the identifier of the feature type in the above list.

[1] 3D coordinates of 21 joints (pelvis, thorax, neck, back head and forehead, both shoulders, elbows, wrists, hands, hips, knees, ankles, feet) but due to the placement of the reference on the hips, the x-coordinates of both hips are always zeros, leading to only 61 useful features.

[2] 53 non-zero features from 3D coordinates of 21 joints.

[3] Quaternions of 16 segments: head, thorax, both arms, forearms, hands, hips, thighs, calves, feet.

[4] Perceived discomfort extracted for 14 joints: neck, pelvis, both shoulders, elbows, wrists, hips, knees and ankles.

7.2.3 Factor definition

Morphology can be defined with numerous variables tightly linked together, such as the size or weight of each body segment. To extract the most relevant variable to represent morphology, a PCA was performed on several variables, including individual segment lengths (foot, calf, leg, trunk, arm, forearm, hand and head), hip width, shoulder width, size from feet to head, and size from feet to fingers.

The first principal component alone explained 76% of the data variance, and it was almost equivalent to the size from feet to fingers ($R = 0.9932$, $p = 1.15 \times 10^{-15}$). As a consequence, the size from feet to fingers is chosen as the morphology factor m for the experiment. The use of only one variable to represent morphology limits the complexity of the regression model, and thus the risks of overfitting due to the small number of individuals in the benchmark dataset ($I = 12$).

The skill factor s in the gesture will be estimated as the mean level of the individual's skill as annotated by the experts (see Table 4.1, last column). To verify the independence of both factors (m and s), their correlation was computed for the 12 participants of the benchmark dataset, resulting in $R = -0.03$ ($p = 0.92$).

7.3 Results

7.3.1 Validation on the eight Bafa techniques

To test the effectiveness of MIRFE in various situations, this section presents a correlation analysis of feature statistics and individual factors for various gestures (the eight Bafa techniques) extracted from the Taijiquan MoCap dataset. For each feature, absolute correlations of individual statistics (μ and σ) with morphology (m) and with skill (s) were computed, without processing, after scaling, and after MIRFE. The mean of these absolute correlations was then extracted for all features. Results are displayed in Fig 7.3.

The upper graphs (a and b) show the mean absolute correlations of m with features statistics (μ on the left graph and σ on the right graph). Without any processing, $\overline{\Phi_{m,\mu}}$ averages to 0.28 for all the analyzed features and techniques, and $\overline{\Phi_{m,\sigma}}$ averages to 0.22. It can be observed that the baseline, i.e. the scaling method only partially reduces the correlation between feature statistics and morphology ($\overline{\Phi_{m,\mu}} = 0.22$ and $\overline{\Phi_{m,\sigma}} = 0.19$), whereas MIRFE removes it almost completely. With MIRFE, the absolute correlation with m is null for all features and for each technique ($\overline{\Phi_{m,\mu}} = 0$), and $\overline{\Phi_{m,\sigma}}$ averages to 0.018 for all features and techniques. As the results are similar for the eight Bafa techniques, it seems therefore that MIRFE allows for an efficient removal of the morphology influence on the features, independently of the type of motion.

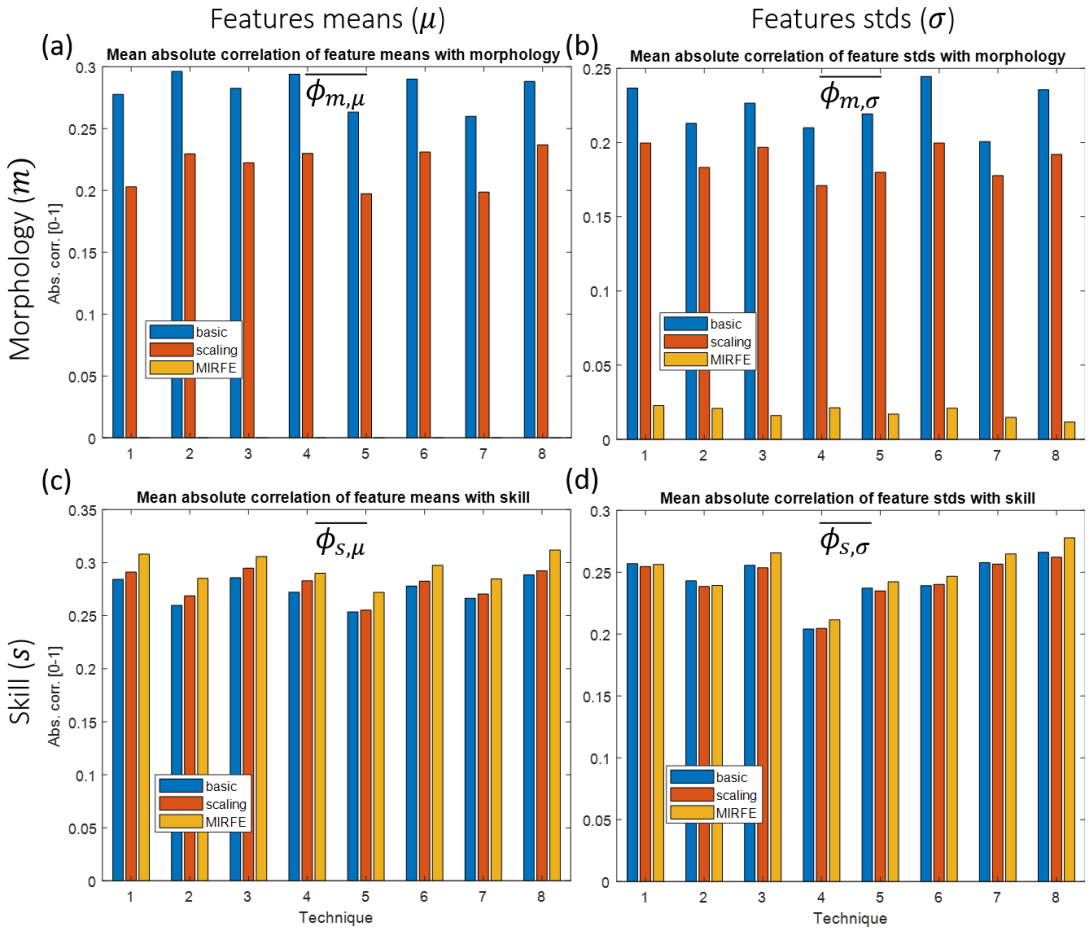

Figure 7.3: Absolute correlation analysis between feature statistics and motion factors, for the eight Taijiquan Bafa techniques: (a) morphology and features means; (b) morphology and features standard deviations; (c) skill and features means: (d) skill and features standard deviations.

The lower graphs, c and d in Fig 7.3, respectively show the mean absolute correlations of the participant skill s with features statistics μ and σ. It can be seen that for each technique, MIRFE yields the best results. Concerning $\overline{\Phi_{s,\mu}}$, the use of MIRFE significantly increases the mean absolute correlation with s by $[.0017 - 0.025]$ for the eight Bafa techniques.[5] Concerning $\overline{\Phi_{s,\sigma}}$, the use of MIRFE leads to a significant improvement for all techniques except the first two ('Drive the monkey away' and 'Move hands like clouds'), up to 0.012 for the last technique ('Grasp the bird's tail'). On the opposite, no significant improvement could be observed for the baseline method, both for $\overline{\Phi_{s,\mu}}$ and $\overline{\Phi_{s,\sigma}}$.

[5]A significant difference between $\Phi_{s,\mu}$ with and without MIRFE is assumed if the p-value of the Student's t-test on their difference for all features respects the condition: $p < 0.005$.

7.3.2 Feature set correlation analysis

In Section 7.3.1, the effectiveness of MIRFE was demonstrated for various types of gestures from the Taijiquan dataset. In this section, factor effects are analyzed for various feature types, without processing, with skeleton scaling and after MIRFE post-processing.

Fig 7.4 shows the mean absolute correlations computed on each feature set (see Section 7.2.2) for the eight Bafa techniques. Without any processing, global and local joint coordinates seem to be the most related to morphology ($\overline{\Phi_{m,\mu}} = 0.40$ for global coordinates and $\overline{\Phi_{m,\mu}} = 0.43$ for local coordinates). As already seen in Section 7.3.1, the use of MIRFE allows an effective removal of the correlation of feature statistics with m, while skeleton scaling only reduces it partly for some features (see Graphs (a) and (b) in Fig 7.4).

Graphs (c) and (d) respectively display the relations between feature statistics (μ on the left graph and σ on the right graph) and skill (s). It can be observed that for any type of feature, the use of MIRFE leads to a significant improvement of their correlation with s by $[0.01 - 0.04]$ for $\overline{\Phi_{m,\mu}}$, and by $[0.004 - 0.013]$ for $\overline{\Phi_{m,\sigma}}$ (except for relational and ergonomics features). On the opposite, skeleton scaling yields significant improvement only for $\overline{\Phi_{m,\mu}}$ on global positions (improvement of 0.02) and on relational features (improvement of 0.018).

From these graphs, it seems that after the process with MIRFE the joint 3D global coordinates are the most linearly related features with s. However, it does not indicate that they provide the best description of skill, as joint global coordinates may also be highly redundant. To analyze the redundancy within the feature sets, absolute pairwise correlations were computed for each feature set ($\overline{\Phi_f}$), as illustrated in Fig 7.5. It can be observed in this figure the use of MIRFE significantly reduces the redundancy of the features for all the feature sets by $[0.035 - 0.071]$, except for local quaternions. On the opposite, skeleton scaling reduces the redundancy only for the joint local coordinates. It can also be observed from this graph that global coordinates are the most redundant features in every case. Their pairwise absolute correlations average to $\overline{\Phi_1} = 0.27$ without processing and with skeleton scaling, and are decreased to $\overline{\Phi_1} = 0.23$ after MIRFE.

7.4 Discussion

The results show that our MIRFE method allows to remove almost completely the morphological influence on the features (at least for one morphological factor). Moreover, it seems that, by removing morphology influence, the redundancy between the features is decreased, and that their relations with other factors such as skill can be improved. This method could be used to improve the analysis of the influence of

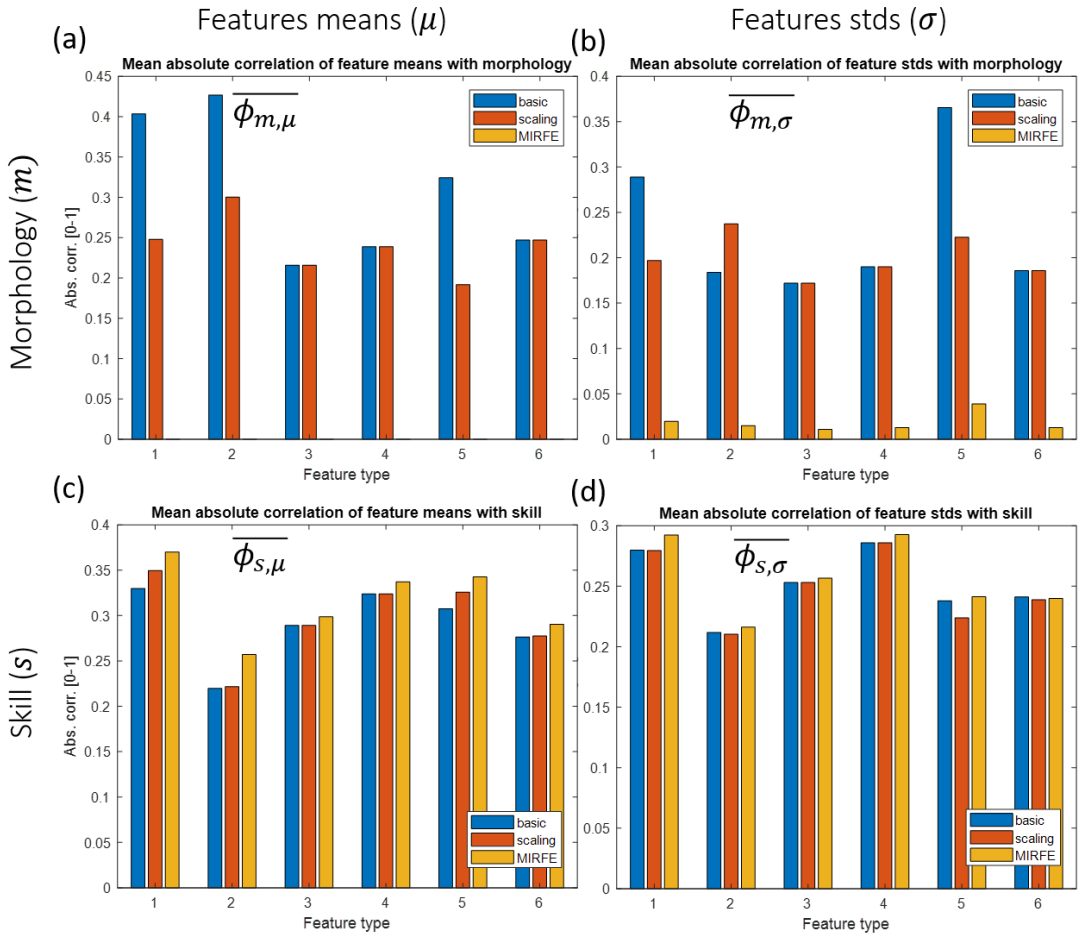

Figure 7.4: Absolute correlation analysis between feature statistics and motion factors, for each feature type: (a) morphology and features means; (b) morphology and features standard deviations; (c) skill and features means: (d) skill and features standard deviations. The indices of feature type correspond to the list in Section 7.2.2.

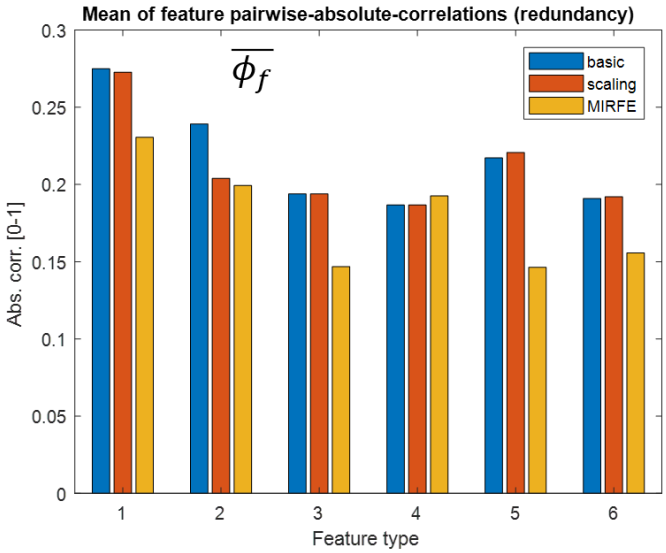

Figure 7.5: Absolute pairwise correlations between features, for each type of feature, without processing, after skeleton scaling, and after MIRFE). The indices of feature type correspond to the list in Section 7.2.2.

inter-individual factors other than skill on movement, such as expression, fatigue, illness, age, etc. Unlike skeleton scaling, MIRFE could easily be used with any morphological factor, such as weight, hip width or shoulder stature, in order to reduce their influence on motion. Moreover, factor-independent residual feature extraction could also be generalized for reducing unwanted influence of other factors, such as age or sex. However, morphology has a direct influence on motion and is thus more appropriate in our case.

Another drawback of direct skeleton data manipulation, such as basic scaling or skeleton representation adaptation, is that they only consider spatial variability due to morphology. In fact, motion is a spatiotemporal series, and morphology may also have an influence on time, because of body inertia for instance. As MIRFE can be applied to any feature, it can also be used on kinematic or kinetic features, and hence reduce morphology influence on their variability.

MIRFE is based on simple linear regression, but could be generalized to non-linear regression, or multiple regression using several morphological or non-morphological factors as predictors. It could theoretically model multiple relations and interactions between feature statistics and motion factors. However, a more complex model would probably need a larger dataset than the one used in this study. As a reminder, the dataset includes only 12 different participants, i.e. 12 unique values for each factor variable. This study of a more complex model is thus left as an improvement prospect.

7.5 Conclusion

In this Chapter, an original method was proposed for extraction of morphology-independent features (MIRFE). This method is based on constrained linear regression with a morphological factor on features statistics (means and standard deviations). The residues of these regressions allow for the computation of residual features, independent of the morphology. Results showed that MIRFE efficiently removes the influence of morphology and improves the relation with the participant skill. MIRFE outperforms skeleton scaling both for morphology independence and skill relation improvement. Its effectiveness will be further tested in various gesture evaluation models presented in the following chapters (see Chapters 8 and 9). MIRFE also has as advantage that it could be used with any inter-individual factor, and on any feature. MIRFE could also be adapted with more complex models than linear regression, but would probably require a larger dataset in that case.

Part III

Gesture Evaluation: a case study on Taijiquan

Chapter 8

Gesture evaluation: a statistical-based approach

Contents

8.1 Introduction

This third part of the present thesis focuses on the evaluation of the expertise within a gesture. In the following chapters, several models will be proposed and tested. As a reminder, previous works concerning gesture evaluation are presented and discussed in Chapter 3. All these studies proposed an original method generally using a single type of motion features, and tested it on a specific motion capture dataset generally consisting of a single gesture type. Moreover, these methods were rarely compared, as their code was generally not provided for reproducibility, and because of the lack of any available benchmark dataset.

This chapter presents a model based on feature statistics and classical machine learning for evaluating the expertise within a gesture. In the taxonomy presented in Chapter 3, the proposed method can be ranked as a "score prediction" method (see Section 3.4.3), based on a regression model. The regression is performed on Principal Components (PCs) extracted on means and standard deviations over time of a large set of motion features. The originality of the proposed method essentially resides in (i) the variety of features that can be used as input, allowing the combination of different types of features, and (ii) the feature post-processing step (MIRFE, see Chapter 7). In this chapter, the proposed method will be tested with various feature combinations, and with different regression models. The use of MIRFE will be assessed, using the proposed evaluation model as a validation procedure. In the next chapters, two original exploratory studies will be presented, proposing a gesture evaluation model based on deep learning (see Chapter 9), and an original and generic visual feedback method (see Chapter 10).

8.2 Methods

8.2.1 Model design

Fig 8.1 illustrates the general workflow of the proposed gesture evaluation model. A set of motion features is provided as the input of the workflow, containing N samples (for N sequences of the same gesture), and F temporal features (extracted from the corresponding motion sequences). Any type of temporal feature can be used, including both low-level or high-level motion representations. From these features, statistics are extracted for each sample, including the mean (μ) and the standard deviation (σ) over time, leading to $2 \times F$ variables. A PCA is then performed on these variables, allowing a reduction of the number of variables while keeping the largest variance possible. Finally, a regression model is used on the extracted PCs to predict the skill level of the participant. The proposed workflow is generic, in the sense that it can be used with any number and type of motion features, and with any regression model.

8.2.2 Experiment

The proposed model will be evaluated using various sets of features. The efficiency of the feature post-processing step with MIRFE will also be evaluated. As a benchmark, the eight Bafa techniques of the Taijiquan dataset presented in Chapter 4 will be used (see Table 4.3). Six different types of features will be used in this experiment, including:

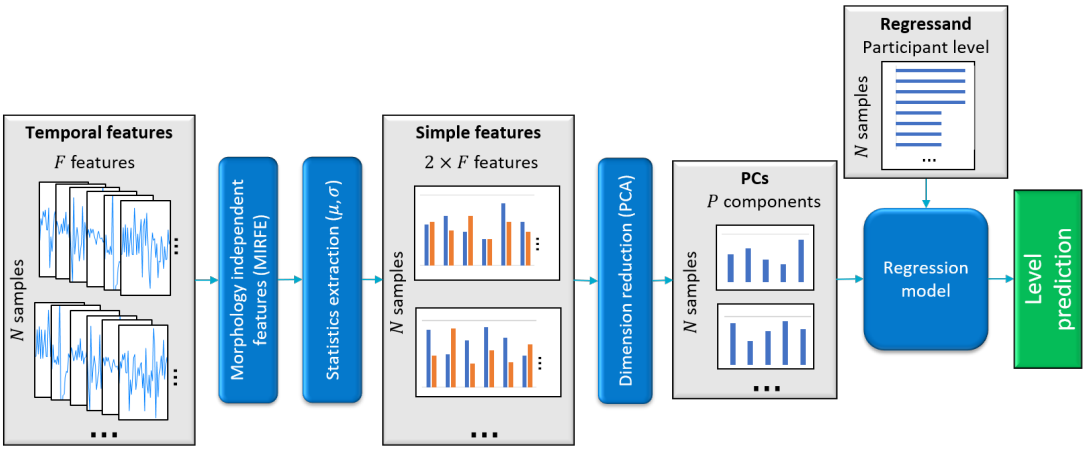

Figure 8.1: Generic workflow of the statistical-based gesture evaluation model.

1. Joint 3D global coordinates (reference placed on the hips) ($F = 61$)[1]
2. Joint 3D local coordinates (reference placed on each parent joint) ($F = 53$)[2]
3. Joint global quaternions (reference placed on the hips) ($F = 64$)[3]
4. Joint local quaternions (reference placed on each parent joint) ($F = 64$)
5. Continuous relational features (Müller feature set without thresholding, see Section 2.3.3) ($F = 40$)
6. Ergonomic features (including 32 ROM from Table 2.1, 36 Taijiquan ergonomic principles from Table 6.1, and 11 CoM kinematic features) ($F = 79$)[4]

These categories of features will be used separately as well as in combination, in order to determine which combination provides the best description of expertise. Finally, various commonly used regression models will be tested, including:

- Linear regression
- Elastic Net regularized linear regression (EN) (Zou and Hastie, 2005)
- Linear Support Vector Regression (L-SVR) (Drucker et al., 1997)
- Gaussian Support Vector Regression (G-SVR) (Aizerman, 1964; Smola and Schölkopf, 2004)

[1] 3D coordinates of 21 joints (pelvis, thorax, neck, back head and forehead, both shoulders, elbows, wrists, hands, hips, knees, ankles, feet) but due to the placement of the reference on the hips, the x-coordinates of both hips are always zeros, leading to only 61 useful features.

[2] 53 non-zero features from 3D coordinates of 21 joints.

[3] Quaternions of 16 segments: head, thorax, both arms, forearms, hands, hips, thighs, calves, feet.

[4] The CoM kinematic features include its 3D coordinates, 3D speeds, 3D accelerations, as well as its normal speed and acceleration.

- Generalized Regression Neural Network (GRNN) (Specht, 1991)
- Multi-Layer Perceptron (MLP) (2 hidden layers with 2 neurons each, trained for 20 epochs) (Rumelhart et al., 1985)
- A tree ensemble using Least Square Boosting (LSB) (20 trees) (Friedman, 2001)

For each test configuration, including various feature types and various regression models, the following procedure will be used: for each Bafa technique, a gesture evaluation model will be tested, following a leave-one-participant-out (LOPO) cross-validation procedure. It means that to test the model efficiency on one participant (the left-out participant), the model will be trained on the other eleven participants, and used to predict the skill level of the left-out participant. Finally, by gathering all the predictions ($N = 1660$ for the 8 Bafa techniques and the 12 participants of the dataset), the Pearson's correlation (R) with the reference skill level (the annotation) will be extracted, as well as the mean absolute error (ϵ) of the prediction.

8.2.3 Comparison with related work

Two methods presented in the recent literature will be compared to the proposed method:

- The first method is using eigenmovement decomposition. Eigenmovements have been used to evaluate the skill in various motion disciplines, as presented in Chapter 3. The method used in this work is similar to Young and Reinkensmeyer (2014) and Zago et al. (2016) (see Section 3.4.3). The first eigenmovement were extracted for each sequence, and a linear regression model was trained on their weights to predict the participant skill.
- The second method is the spatial error computation method from Morel et al. (2017) (see Section 3.4.2). On motion sequences temporally aligned with DTW, a spatial error was computed for each limb (both arms, both legs and trunk). As this score is computed frame-by-frame, the mean over time was extracted, leading to a total of five variables (the mean spatial error for each limb). A linear regression was then used with these five variables to predict the participant skill.

8.3 Results

8.3.1 Feature comparison

This section presents a comparison of the effectiveness of various feature types for the modeling of expertise with the proposed model. In this experiment, linear regression

is used, and is tested with various numbers of PCs. The MIRFE post-processing step is not used yet, in order to analyze the initial representation power of each feature set.

Fig 8.2 (a) shows the results for the prediction of the participant skill level for the eight Bafa techniques. Among the six compared features, the best results were obtained with the ergonomic features, leading to a correlation of $R = 0.73$, using only the first two PCs as predicting variables. Based on the assumption that the most diversified feature set would allow the best modeling of expertise, a larger feature set was used by combining global positions, local quaternions, relational and ergonomic features. This feature set led to even better results with a correlation of $R = 0.79$ with two PCs as predicting variables. With this configuration, the participant level can be predicted with a mean absolute error of $\epsilon = 0.72$, for all participants and the eight Bafa techniques.

Fig 8.2 (b) shows the variance accumulated by the PCs extracted from each feature set. All curves seem similar, except for the relational features, where the first PCs accumulate more variance. This is probably due to the fact that this feature set is smaller than others ($F = 2 * 40$ for μ and σ).

The best results were obtained with only two PCs, corresponding to 40% of the total cumulated variance for the relational features and 30% for the combined feature set. The correlation then generally decreases while the number of PCs increases for most feature sets. The remaining information is either not relevant for the modeling of expertise or is not properly interpreted by the modeling algorithm (linear regression in this case).

The previous chapter analyzed the impact of MIRFE on the feature correlation with skill. In the next section, models will be trained with features post-processed using MIRFE, allowing an analysis of its real interest in the modeling of expertise.

8.3.2 MIRFE validation

In this section, the same comparison as in Section 8.3.1 is proposed.Various features types are tested for the modeling of expertise with the proposed model, used with linear regression and various numbers of PCs. However, in this case, the MIRFE post-processing step is applied on the features, allowing a comparison with the previous results.

Fig 8.3 shows the results for the prediction of the participant skill level for the eight Bafa techniques. It can be observed in the left graph (a) that results include several differences with the ones presented above (see Fig 8.2 (a)). Firstly, for most feature sets the correlation seems to increase with the number of PCs used as predicting variables, up to about 45 PCs, corresponding to 99% of the cumulated variance for

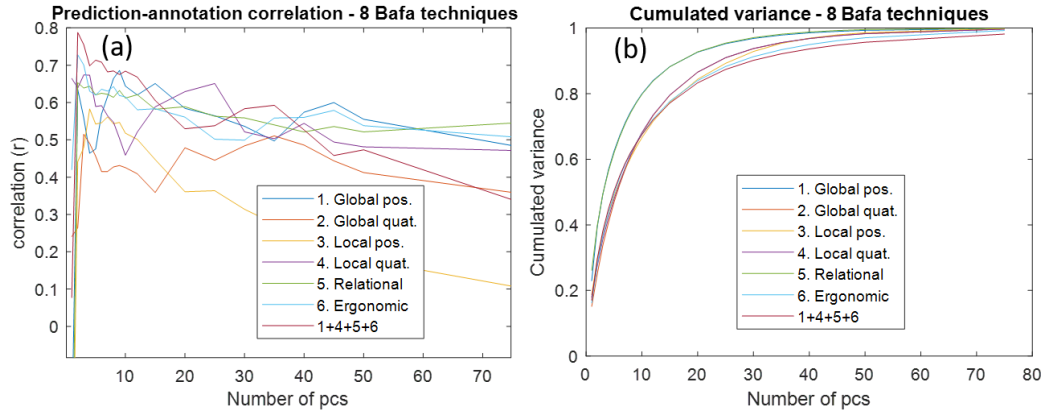

Figure 8.2: Participant skill prediction using linear regression on PCs extracted on various features sets (no MIRFE post-processing).

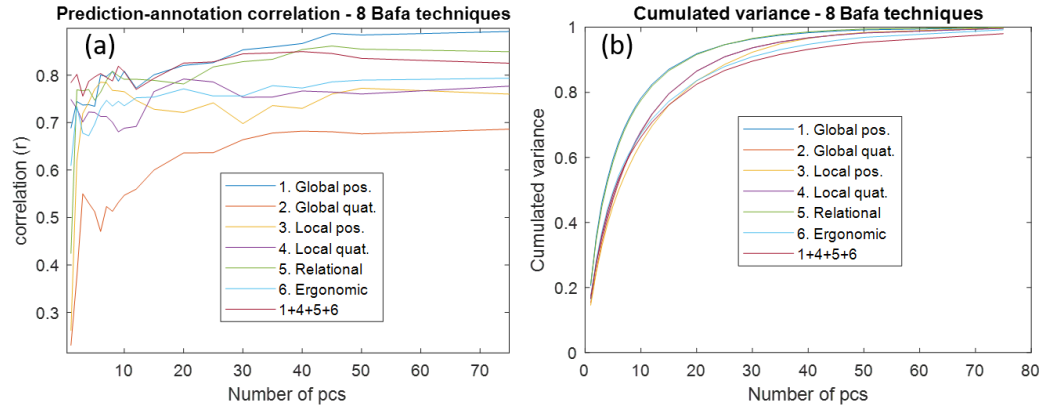

Figure 8.3: Participant skill prediction using linear regression on PCs extracted on various features sets, after MIRFE post-processing.

relational features and 94% for the combined feature set. Secondly, the best results are not obtained with high-level features nor with the combined feature set, but with the global positions ($R = 0.88$ with 45 PCs). With this configuration, the participant level could be predicted with a mean absolute error of $\epsilon = 0.52$, for all participants and the eight Bafa techniques.

8.3.3 Model comparison

In this section, various commonly used regression models are compared for the modeling of the expertise. Each regression model is tested with various feature types and with various number of PCs extracted on features processed with MIRFE. Table 8.1 presents the best results for each regression model. The parameters of each model were optimized for the eight Bafa techniques dataset. Best results were obtained with the three linear models (EN, L-SVR, linear regression), with 60 PCs extracted from

Model	Feature types	N_{pcs}	R
Linear regression	1+5	60	0.904
EN	1+5	60	0.909
L-SVR	1+5	60	0.890
G-SVR	1+4+5+6	1	0.789
MLP	1+4+5+6	1	0.788
LSB	1+4+5+6	1	0.769
GRNN	1+4+5+6	1	0.800

Table 8.1: Participant skill prediction using various regression models: best results. The numbers used to identify the features correspond to the list presented in Section 8.2.2.

global positions and relational features. For these three models, the correlation with annotations generally increased with the number of PCs until 60 PCs (similarly to Fig 8.3 (a)), up to $R = 0.909$ for EN[5]. A clear gap can be observed with non-linear models (G-SVR, MLP, LSB and GRNN). With all these models, the correlation with annotations decreased with the number of PCs. For these models, the best results were obtained with a single PC on the large feature set combining global positions, local quaternions, relational and ergonomic features.

8.3.4 Comparison with related work

Table 8.2 shows the results for two methods from the related work: eigenmovements, and limb spatial errors from Morel et al. (2017). These methods were used on features that were either extracted on a scaled skeleton, or post-processed with MIRFE. These results are compared with the best results obtained for the method proposed in this chapter. For both methods, the feature post-processing with MIRFE leads to better predictions. The score regression with limb spatial error leads to the worst correlation ($R = 0.425$ using MIRFE). Regression with 60 eigenmovement weights leads to a correlation of $R = 0.849$. The proposed method outperforms both methods, with $R = 0.909$.

8.3.5 Synthesis

From all the tested models, the best results (of $R = 0.909$) were obtained for a model based on:

- A feature set combining global positions and relational features

[5]With optimal parameters, the EN corresponded to a L2-regularization with a regularization parameter $\alpha = 0.0017$

Method	$N_{variables}$	R
Eigenmovement weights (with scaling)	10	0.709
Eigenmovement weights (with MIRFE)	60	0.849
Morel et al. (2017) limb spatial error (with scaling)	(5)	0.319
Morel et al. (2017) limb spatial error (with MIRFE)	(5)	0.425
Proposed method (EN-regression on 60 PCs extracted from statistics (μ and σ) of global positions and relational features, with MIRFE)	60	0.909

Table 8.2: Participant skill prediction using two methods from the literature.

- A feature post-processing step using MIRFE
- The extraction of 60 PCs from the feature means and standard deviations
- An EN-regularized linear regression model

Fig 8.4 shows the predictions for all the samples of the dataset including the eight Bafa techniques. The smallest correlation was obtained for the gesture G11 ('Kick with the heel'), with $R = 0.828$, and the largest one was obtained for G06 ('Drive the monkey away') and G13 ('Grasp the bird's tail'), with $R = 0.953$. The mean absolute error of all the predictions is $\epsilon = 0.424$.

8.4 Discussion

In this chapter, various skill evaluation methods were tested and compared with two methods of the literature. The best results were obtained with the proposed method, configured with an EN-regression model, and using a combination of low-level features (global positions) and higher-level features (relational features), with a correlation of $R = 0.909$, and a mean absolute error of $\epsilon = 0.473$. The proposed method outperformed two methods from the literature, including regression on eigenmovement weights (Young and Reinkensmeyer, 2014), and regression on spatial errors (Morel et al., 2017).

Two major advantages of the proposed method may explain these results: first, it can be used with any type of feature, and the PCA allows extraction of a compact representation while including various information due to the different types of features used. The combination of global positions with relational features, providing a higher-level representation of motion, allowed the extraction of the most relevant PCs for the description of skill in the present study.

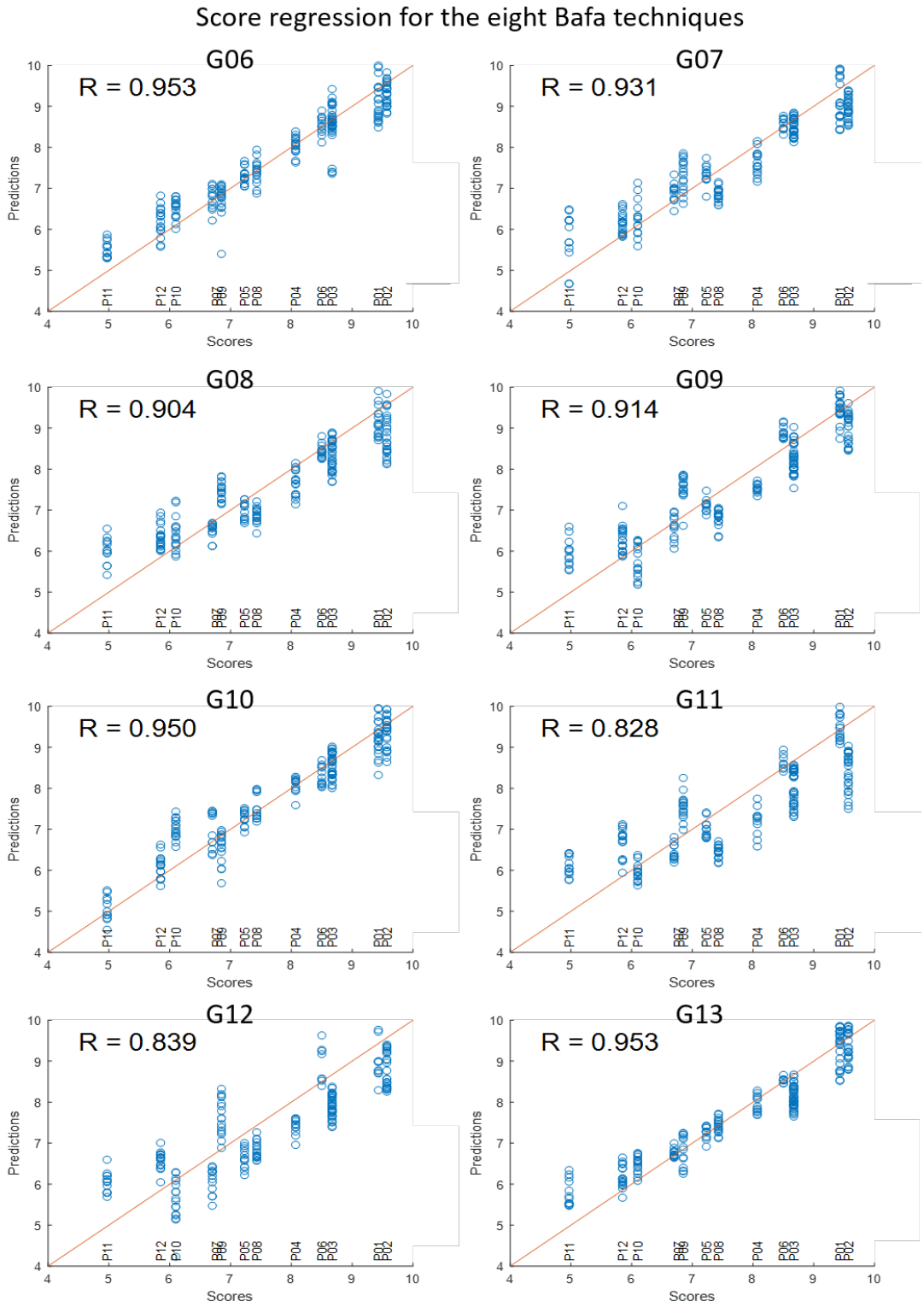

Figure 8.4: Score predictions for the eight Bafa techniques. Model: EN-regression on 60 PCs from μ and σ of global positions and relational features, post-processed with MIRFE.

Secondly, another possible advantage of the proposed method is that it is based on the comparison of statistics extracted from the gestures. On the opposite, eigenmovements and spatial errors are based on frame-by-frame relations between gestures. This type of comparison is thus highly sensitive to the temporal synchronization of the gestures to compare. Though temporal alignment methods such as DTW can be used to deal with synchronization, they only allow a global alignment of the motion sequence, and do not ensure an optimal alignment of each marker position or each feature independently. Moreover, they alter the original gestures and do not ensure that the aligned movement still corresponds to the same skill level as the original.

Nevertheless, frame-by-frame relations could be relevant for other disciplines where the timing of different motion features in a gesture is particularly important. In that case, a statistical representation of the features would be inappropriate.

Another limitation arises from this same characteristic of the proposed model: as the skill is evaluated from statistics extracted on the entire motion sequence, a score can only be predicted for the entire motion sequence. On the contrary, a method based on frame-by-frame comparison such as Morel et al. (2017) provides a spatial error for each frame of the gesture. A frame-by-frame score allows feedback information on which part of the sequence was performed more or less efficiently. Nonetheless, a new generic feedback method is presented in Chapter 10, allowing wider feedback information, independently of the gesture evaluation model.

For each model tested in this chapter, the feature post-processing with MIRFE led to significantly better score predictions. With MIRFE, the quality of the model seems to increase with the number of PCs (see Fig 8.3). As a comparison, without MIRFE, the best results are obtained for one or two PCs for any feature set, as illustrated in Fig 8.2. It seems therefore that MIRFE allows the extraction of features that are easier to interpret by the proposed models. The same observation was made for two methods of the related work (see Table 8.2). It is probably due to the fact that the processed features are less dependent on morphology and are thus more comparable between participants, but also less correlated between each other.

These results can be compared to the differences between teachers' annotations (see Table 4.1). Tables 8.3 and 8.4 respectively show the correlations and mean absolute differences between these annotations. The largest difference is observed for $Skill_1$ and $Skill_3$ with a correlation of $R = 0.949$ and a mean absolute difference of $\epsilon = 0.392$. The variability of the annotations enlightens some subjectivity in the teachers' perception of each participant's skills. As the model is trained to predict these annotations, the results are limited to the teachers' own perception of expertise. The prediction results ($R = 0.909$ and $\epsilon = 0.473$) should hence be compared to annotation variability (represented by $R = 0.949$ and $\epsilon = 0.392$).

Besides this annotation variability due to teachers' subjectivity, the performer's own variability must also be highlighted: each rendition of a gesture can be performed by

R	$Skill_1$	$Skill_2$	$Skill_3$	$Experience$
$Skill_1$	1	0.972	0.949	0.831
$Skill_2$	0.972	1	0.963	0.854
$Skill_3$	0.949	0.963	1	0.934
$Experience$	0.831	0.854	0.934	1

Table 8.3: Correlations of the annotations of the three teachers with each other and with participant experiences (years of practice).

ϵ	$Skill_1$	$Skill_2$	$Skill_3$
$Skill_1$	0	0.275	0.392
$Skill_2$	0.275	0	0.358
$Skill_3$	0.392	0.358	0

Table 8.4: Mean absolute difference between the annotations of the three teachers.

the same performer with a different quality. For instance, athletes never throw a disk at the same distance nor jump over the same distance in two consecutive attempts. This variability can be due to various factors, including the performer's fatigue or concentration. To assess this variability in this dataset, the first teacher was asked to annotate every rendition of the 'Kick with the heel' gesture (G11) for each participant. On these annotations, the mean absolute difference to the average annotation of each participant was calculated, resulting in $\epsilon_{rendition} = 0.402$. According to the teacher, G11 is the most difficult technique. It requires more balance, accuracy and synchronization than others, probably leading to more variability between renditions.

Finally, the same target variable (the global participant level) was used to model the skill for each technique. However, each participant can master some techniques better than others. This leads to another type of skill variability that is not learned by the model.

A higher prediction accuracy could hence be obtained with individual annotations for each rendition, or at least for each type of gesture. However, such an annotation would require an extensive work, and would still be limited to the annotators' subjectivity. The use of global scores for each participant is a simplifying assumption, which still leads to interesting results ($R = 0.909$ and $\epsilon = 0.473$). In other words, the global participant level can be predicted with a relative error (ϵ_{rel}) of 10.28%:

$$\epsilon_{rel} = \frac{\epsilon}{(\max(Skill_\mu) - \min(Skill_\mu))} = \frac{0.473}{9.57 - 4.97} = 0.1028 \tag{8.1}$$

8.5 Conclusion

In this chapter, a new gesture evaluation method was presented. The method is presented as a generic model which can be used with various types of motion features,

and any type of regression algorithm. From motion features post-processed with MIRFE, means and standard deviations are extracted for each gesture of the dataset. PCA is applied on the means and standard deviations, allowing a dimensionality reduction to a smaller set of PCs. A regression model is then used on these PCs to predict the skill level annotated by experts. The proposed model was tested on the eight-Bafa techniques dataset, leading to a correlation of $R = 0.909$ with the best configuration, i.e. with global positions and relational features post-processed with MIRFE, 60 PCs , and EN-regression (actually corresponding to a ridge regression with a L2-regularization parameter of $\alpha = 0.017$). The proposed method outperformed other methods from the literature, including regression on eigenmovement weights (Young and Reinkensmeyer, 2014), and regression on spatial errors (Morel et al., 2017). Results show a significant effect of the MIRFE processing method presented in Chapter 7. This post-processing step allows better interpretation of the features by all of the models tested in this chapter, including methods from the related work. The results are limited to the subjectivity of the annotations provided by the teachers, and better prediction could be obtained if annotations were provided by a larger number of experts for each sample of the dataset.

Chapter 9

Towards a deep-learning-based gesture evaluation model

Contents

9.1 Introduction

In the previous chapter, we presented a statistical-based gesture evaluation method using classical machine learning algorithms (PCA and EN-regression). Recent developments in machine learning include Deep Neural Networks (DNN), and more particularly Convolutional Neural Networks (CNN), originally designed for image analysis. These techniques dramatically outperform classical machine learning in more and more disciplines including speech processing, image processing, and many

other domains (LeCun et al., 2015). This type of model has recently been successfully adapted in Laraba et al. (2017) for gesture recognition in MoCap data. The following subsections briefly explain the basics of deep learning, which are necessary to understand the method proposed in Laraba et al. (2017). An adaptation of this method is then proposed in Section 9.2, and serves as a first proof-of-concept in the exploration of the use of deep learning in gesture evaluation.

9.1.1 Neural network

A Neural Network (NN) is a machine learning system, consisting of connections between learning units, trained together with a dataset to learn a task such as regression or classification. The basic unit of a NN, the neuron, is a non-linear function (usually a logistic sigmoid), called the *activation function* linking a set of input variables with an output variable. A common type of NN is the Multilayer Perceptron (MLP), or feedforward NN (see Fig 9.1). In this type of network, layers of neurons are connected in series. If neurons or hidden layers are added to the network, it allows modeling of more complex relations between the input variables (x) and the output variable (y). However, it also adds a large number of parameters (weights w), increasing the risk of overfitting or convergence failing. A larger dataset is then needed to train the network efficiently.

A NN is trained by minimizing a cost function defined by the prediction errors of the training set. This optimization procedure is generally performed using the *backpropagation* technique. With this technique, the prediction error is computed at the output and distributed back to the previous layers. An iterative gradient-descent with a specified step size (called the *learning rate*) is used to reach a local minimum of the cost function. This procedure can be performed either with a *batch process*, using the entire training set at each iteration for computing the gradient, or with a *mini-batch* process, using smaller sets of training samples at each iteration, generally allowing a faster convergence to a local minimum with an adapted learning rate. The mini-batch process is also called *stochastic gradient descent* (SGD) as the gradient computed with the fewer training samples can be seen as a noisy estimation of the gradient computed on the entire training set (LeCun et al., 2015).

A simple DNN can be defined as a NN with a large number of layers. As the number of layers (and hence complexity) increases, the convergence of the model is harder to reach. Due to the almost horizontal shape of the sigmoid function at its tails, the gradient is close to zero, and eventually becomes zero through backpropagation into several layers, due to computer limited floating-point precision. To deal with the vanishing gradient, different activation functions have been used, such as the Rectified Linear Unit (ReLU, Glorot et al. 2011), following the equation:

$$y = max(0, x) \tag{9.1}$$

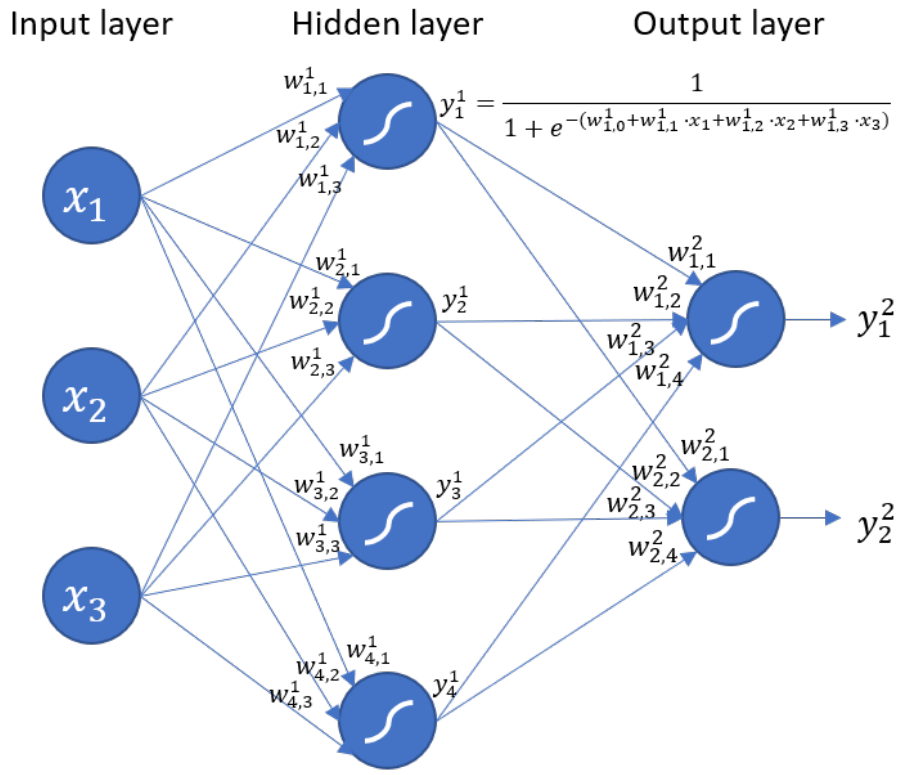

Figure 9.1: Feedforward neural network with one hidden layer.

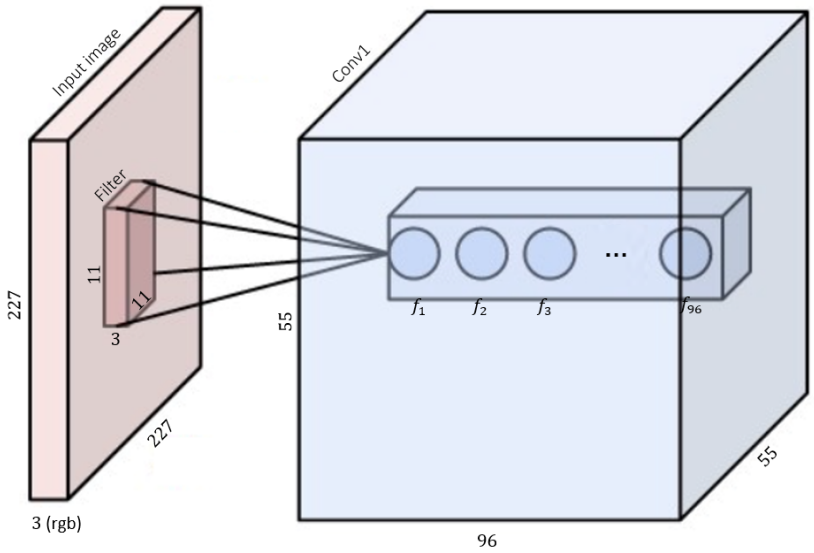

Figure 9.2: CNN convolution layer (AlexNet first convolution layer).

9.1.2 Convolutional neural network

As a variation of DNN, CNN have been designed for the specific task of image analysis. In this type of network, a layer does not consist of logistic functions on all previous neurons weighted sums, but of convolutional filter banks instead. These filter banks replace the weights of the fully-connected (FC) layers from the MLP, and have the advantage of being more adapted to higher relations between neighbor pixels. Fig 9.2 shows an example of convolutional layer. In this example (which corresponds to the first convolutional layer of the CNN *AlexNet* (Krizhevsky et al., 2012)), 96 3D filters of dimensions $11 \times 11 \times 3$ are convoluted on an input image of dimension $227 \times 227 \times 3$, leading to $96 \times 11 \times 11 \times 3 = 34848$ weights. As a comparison, a single neuron connected to all pixels of the input image would lead to $227 \times 227 \times 3 = 154587$ weights.

The output of the convolutional layer can be seen as 96 filtered and scaled grayscale versions of the input image. A non-linear function (ReLU) is then applied on this output.

Just as MLPs, some CNNs consist of several layers in series (Krizhevsky et al., 2012). Others are designed using more complex architectures with parallel layers or cycles (Szegedy et al., 2015).

Fig 9.3 illustrates the convolutional layers of AlexNet (Krizhevsky et al., 2012), and a visualization of the neurons based on Wei et al. (2017). It can be observed that the first convolutional layers are trained to extract general low-level features from images such as blobs and edges. The following layers allow extraction of higher-level features, such as texture patterns and even objects parts, more dependent on the specific training images.

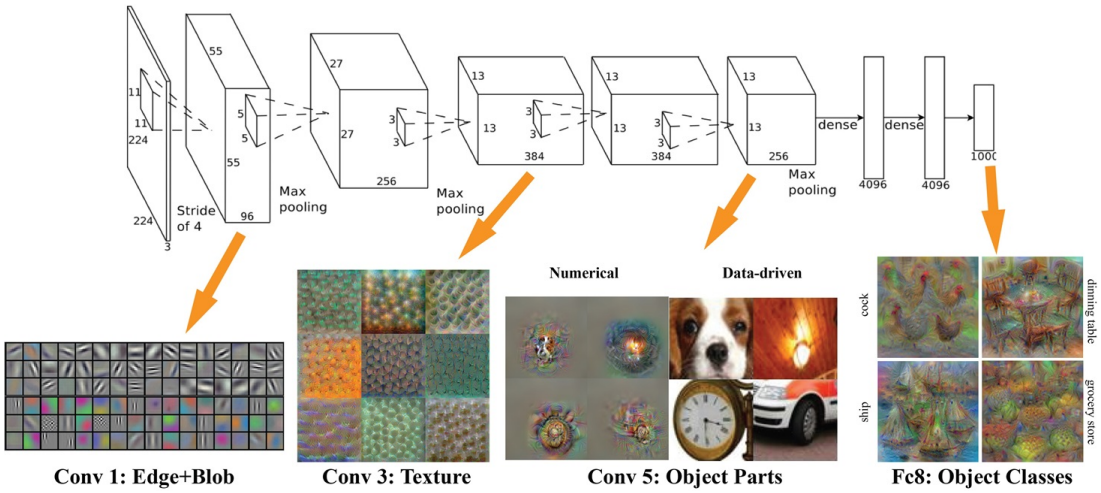

Figure 9.3: Convolutional layers and neuron visualization of AlexNet trained on the ImageNet dataset. Reproduced from Wei et al. (2017).

9.1.3 Transfer learning and fine-tuning

In practice, large image-recognition CNNs are rarely trained from scratch. Their training has an extensive computational cost, which can be several weeks with a single GPU (You et al., 2017). Moreover, for an efficient training, a large dataset is needed. As an example, the state-of-the-art image-recognition DNNs are generally benchmarked at the 'ImageNet Large Scale Visual Recognition Challenge' (Krizhevsky et al., 2012). They are trained on a dataset of 1.2 million images labeled into 1000 categories, a subpart of the ImageNet dataset (Deng et al., 2009).

Instead of recording and labeling a large dataset, and training a CNN on it for weeks for each new specific object-recognition task, an existing pre-trained CNN can be adapted for that task. As explained above, the first layers of a CNN encode generic features which can be relevant for a wide range of tasks. On the opposite, the last layers encode more specific features, allowing classification into categories, such as the 1000 categories of the ImageNet dataset. To avoid a complete retraining of a CNN, the first layers of a pre-trained model can be used as a basis, and the last layers can be either replaced by new ones or fine-tuned by pursuing their training on a smaller dataset including new images. The final layer can also be replaced, to classify a few types of object rather than the 1000 categories, or for a regression task. The number of parameters to train is much lower in those cases, decreasing the risk of overfitting and the computational cost.

9.1.4 Representing a MoCap sequence as an image

Laraba et al. (2017) used transfer learning and fine-tuning on a pre-trained CNN model (*GoogleNet*, Szegedy et al. 2015), and adapted it for the classification of MoCap

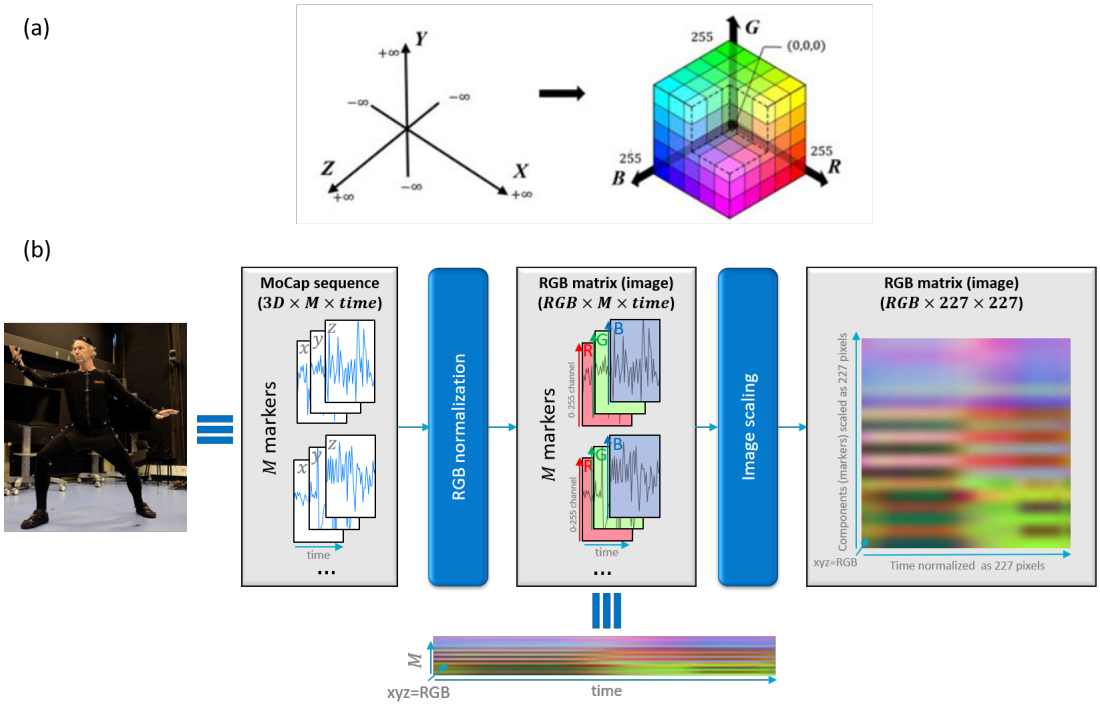

Figure 9.4: (a) Similarity between 3D axes and RGB channels (Reproduced from Laraba et al. 2017). (b) representation of a MoCap sequence as an RGB image.

sequences. As the input layer of the CNN needs an RGB image, they proposed a representation of a MoCap sequence as an image, as illustrated in Fig 9.4. They used the similarity between the 3D axes of MoCap data (x, y and z) and RGB images (red, green and blue channels, as illustrated Fig 9.4 (a)). They transformed MoCap data into the RGB scale with a discretization and normalization of each 3D axis on a [0-255] scale (see Fig 9.4 (b)), and resized the image linearly to fit the input size of the model (227 × 227 pixels). The result is a striped abstract image where each stripe corresponds to a marker trajectory. Using this technique, they fine-tuned GoogleNet to recognize actions in various3D MoCap datasets, and outperformed the recent literature in most cases.

9.2 Methods

In this Section, a method inspired by Laraba et al. (2017) is proposed for gesture evaluation, using transfer learning and a final regression layer on participant levels. The proposed method is generalized to any type of motion representation as explained in Section 9.2.1, and relies on two principal transfer learning steps (see Section 9.2.2). The pre-trained CNN used as a basis was AlexNet (Krizhevsky et al., 2012), trained on a subset of the ImageNet dataset (1.2 million images labeled into 1000 categories).

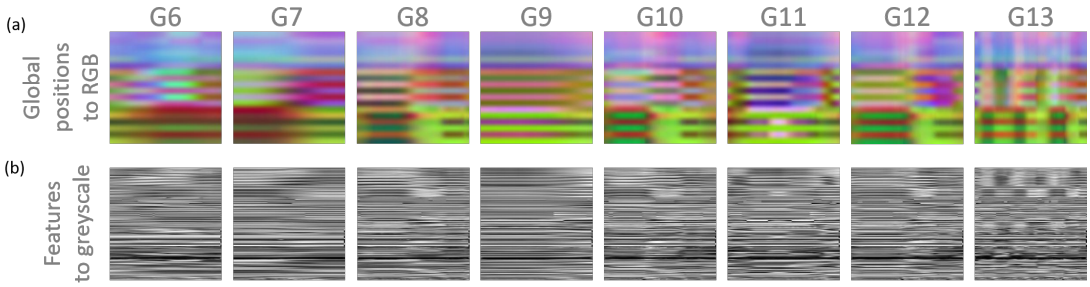

Figure 9.5: Eight Bafa techniques represented as abstract images. (a) Global positions represented as RGB images. (b) Features (global positions, local quaternions, relational and ergonomic features) represented as grayscale images.

AlexNet was selected as proof-of-concept for its relative simplicity and small number of parameters compared to other available pre-trained models. Note that GoogleNet (Szegedy et al., 2015) and SqueezeNet (Iandola et al., 2016) were also tested with various learning parameters without success (random results for the regression of the skill level).

9.2.1 Representing features as an image

As discussed in the previous chapters, different low or higher-level representations of motion can be used, and their use may have various advantages and drawbacks. Higher-level features are assumed to be easier to interpret, while low-level features provide a complete description of motion. Any type of motion feature could be used as an abstract image for the training of a CNN. However, as features are not necessarily three-dimensional, the abstract images would be grayscale (i.e. with only one color channel). As large pre-trained CNNs (including AlexNet) generally use RGB images, a 1D feature is simply replicated on the three channels. Apart from this, the same channel-normalization and image-scaling procedures as Laraba et al. (2017) are applied, resulting in abstract images as illustrated in Fig 9.5.

9.2.2 Two-step transfer learning

Abstract images are inevitably different from the images used in the training of the original pre-trained model. The model should hence be first adapted for interpretation of this new type of image. Secondly, the goal of the final model is the regression of the participant level from these abstract images. The proposed method relies on a two-step transfer learning process, taking advantage of the multiple information available on the dataset, including both the gesture category and the participant level. Fig 9.6 illustrates the two-step transfer learning process:

- First, a pre-trained AlexNet CNN is adapted for the classification of the gesture category. The last two FC layers of AlexNet are replaced by two smaller ones, with 64 neurons for the first one, followed by a ReLU function, and 8 neurons for the last one, followed by a softmax function, allowing classification of the eight Bafa techniques. This model is trained with the eight Bafa technique dataset ($n = 1660$), using a large learning rate for the last two layers (0.01) and a learning-rate 50 times smaller (0.0002) for fine-tuning of the AlexNet pre-trained layers. For the training, mini-batches of 512 images are used. This first step aims at a gross interpretation of the abstract images, in order to distinguish different types of gestures.
- Secondly, the model is further adapted for the prediction of the participant level. To that end, the last layer is replaced by a FC layer with a single output, allowing the regression of the participant skill level. For each Bafa technique, a model is trained of the corresponding data subset ($n \sim 208$), using a larger learning rate for the last layer (0.0025) than for the rest of the network (0.00005). For this training, mini-batches of 20 images are used. This second step allows a finer interpretation of the images of a single category, aiming at extracting discriminant features for different levels of expertise.

The optimal learning rates and batch sizes were determined with a manual iterative procedure.

9.2.3 Experiments

To test the validity of the proposed method, various experiments are conducted. First, the two-step transfer learning procedure are first tested in various configurations, including different feature types and post-processing (MIRFE), either for the classification of Bafa techniques (Step 1) or for regression of the skill level (Step 2). The results of the regression are then compared with those of the statistical-based method presented in Chapter 8.

9.2.3.1 Step 1: Bafa classification CNN

To test the validity of the first step of the transfer learning procedure, the classification model (see Fig 9.6) will be trained with a LOPO procedure, and the validation accuracy will be calculated from the left-out-participant predictions. To that end, 12 models (one for each participant) will be trained with various input feature sets:

1. Joint 3D global coordinates (reference placed on the hips)[1]

[1] 3D coordinates of 21 joints (pelvis, thorax, neck, back head and forehead, both shoulders, elbows, wrists, hands, hips, knees, ankles, feet) but due to the placement of the reference on the hips, the x-coordinates of both hips are always zeros, leading to only 61 useful features.

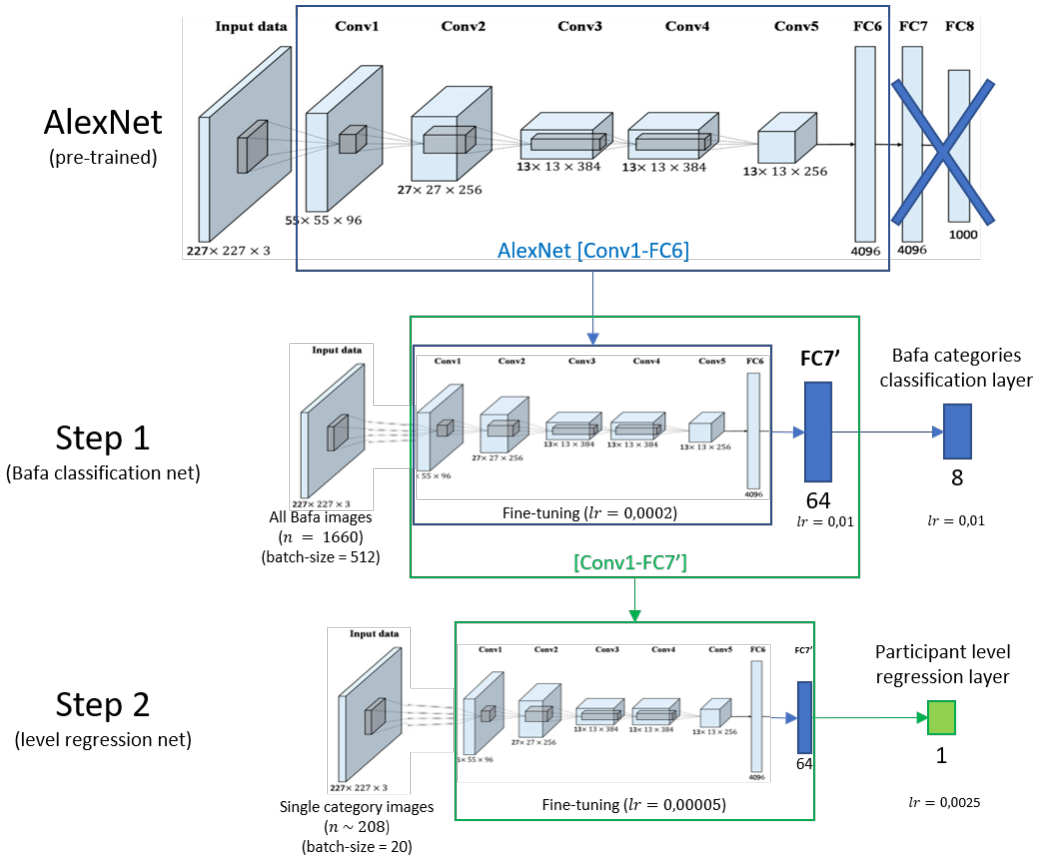

Figure 9.6: Two-step transfer learning procedure. Step 1: a CNN is designed for classification of the eight Bafa techniques. Step 2: a regression CNN is designed for the prediction of the participant skill level on one Bafa technique. (AlexNet image adapted from Han et al. 2017).

   (a) represented as grayscale images ($F = 61$ 1D features)
   (b) represented as colored images ($F = 21$ 3D features)
2. Joint 3D local coordinates (reference placed on each parent joint)[2]
   (a) represented as grayscale images ($F = 53$ 1D features)
   (b) represented as colored images ($F = 21$ 3D features)
3. Joint global quaternions (reference placed on the hips) ($F = 64$ 1D features)[3]
4. Joint local quaternions (reference placed on each parent joint) ($F = 64$ 1D features)
5. Continuous relational features (Müller feature set without thresholding, see Section 2.3.3) ($F = 40$ 1D features)
6. Ergonomic features (including 32 ROM from Table 2.1, 36 Taijiquan ergonomic principles from Table 6.1, and 11 CoM kinematic features) ($F = 79$ 1D features)[4]
7. The combination of joint 3D global coordinates and relational features ($F = 101$ 1D features)
8. The combination of joint 3D global coordinates, joint local quaternions, relational features and ergonomic features ($F = 244$ 1D features)

In each case, the resulting images are linearly resized to fit the input size of the model (227×227 pixels).

To verify the effectiveness of MIRFE, this experiment will be conducted both with and without the MIRFE feature post-processing step.

9.2.3.2 Step 2: level regression CNN

To test the validity of the two-steps transfer learning procedure, CNNs for the regression of the participant level will be trained, with and without a pre-training of the intermediate layers with step one (see Fig 9.6). The models will be trained with a LOPO procedure, and the correlation between the annotations and the predictions of the left-out-participants will be extracted.

As a comparison with the two-step transfer learning procedure, a direct transfer learning procedure will be used. In this procedure, the pre-trained model is directly adapted for the prediction of the participant level. To that end, layers FC7 and FC8 of AlexNet are removed, and a final FC layer with a single output is directly added after FC6. A learning rate of 0.001 is used for the last layer, and 0.00005 for the rest of the network.

[2] 53 non-zero features from 3D coordinates of 21 joints.

[3] Quaternions of 16 segments: head, thorax, both arms, forearms, hands, hips, thighs, calves, feet.

[4] The CoM kinematic features include its 3D coordinates, 3D speeds, 3D accelerations, as well as its normal speed and acceleration.

Feature type	No MIRFE	MIRFE
1.(a) Joint 3D global coordinates (grayscale)	90.65%	95.98%
1.(b) Joint 3D global coordinates (RGB)	95.55%	97.5%
2.(a) Joint 3D local coordinates (grayscale)	62.49%	94.74%
2.(b) Joint 3D local coordinates (RGB)	74.06%	95.81%
3. Joint global quaternions	90.78%	94.18%
4. Joint local quaternions	88.72%	93.23%
5. Relational features	**96.15**%	97.69%
6. Ergonomic features	91.76%	95.18%
1.(a) and 5	90.60%	**97.82**%
1.(a), 4, 5 and 6	84.39%	95.26%

Table 9.1: Validation of the first transfer learning step: classification accuracy of the Bafa techniques.

9.3 Results and discussion

9.3.1 Step 1: Bafa classification CNN

Table 9.1 displays the results obtained for the first transfer learning step, i.e. the prediction accuracy for classification of the Bafa techniques. It can be observed that for any type of feature used by the model, the use of the MIRFE procedure, providing features less dependent on morphology, yields better prediction accuracy. With the MIRFE process, the best results were obtained for the combination of global positions and relational features with an accuracy of 97.82%. This model is thus able to classify the eight Bafa techniques from abstract images including both global positions and relational features processed with MIRFE, with an error of 2.18%.

9.3.2 Step 2: level regression CNN

Table 9.2 displays the correlations between the prediction of the models with the annotations, using the combination of global positions and relational features. It can be observed again that the use of MIRFE yields better correlations either with a direct transfer learning from AlexNet to a regression CNN, or with a two-step transfer learning with an intermediate step on the classification of the Bafa techniques. The best result is obtained with the two-step transfer learning, leading to a correlation of 0.518. On the opposite, the worst result is obtained for the two-step transfer learning without the use of MIRFE, leading to a correlation of 0.329. This might be due to an ineffective intermediate training of the model. It can be noticed that the classification accuracy in this configuration was only 90.60% (see Table 9.1), i.e. a classification error of 9.4%, against 2.18% with the use of MIRFE.

Method	**No MIRFE**	**MIRFE**
Direct transfer learning	0.357	0.454
Two-step transfer learning	0.329	**0.518**

Table 9.2: Prediction correlations with annotations for various skill-level-regression CNNs.

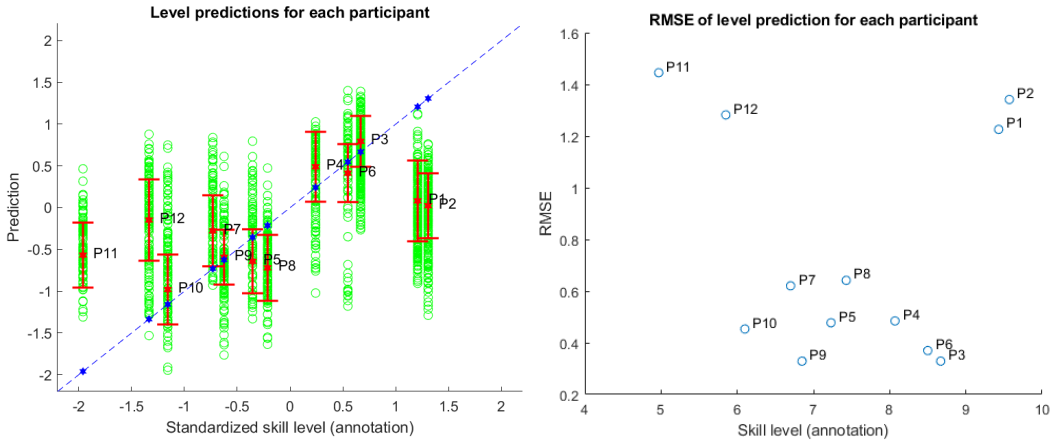

Figure 9.7: Left: predictions of all motion sequences for each participant, against their skill level. Right: prediction RMSE for each participant, against their skill level.

These correlations are low in comparison with the results obtained with the method proposed in Chapter 8. As a reminder, the best correlation ($R = 0.909$) was obtained with EN-regression on 60 PCs extracted from means and standard deviations of motion features. However, as shown in Fig 9.7, the prediction error seems to depend on the skill level of the participant, and is thus not fully random. It can be observed in Fig 9.7 (right graph) that for all the participants with a skill level between 6 and 9, the prediction RMSE is below 0.65, and does not seem dependent of the skill level. On the opposite, for the extreme participants (the two best ones, P1 and P2 and the two worst ones, P11 and P12), the prediction RMSE is above 1.2, i.e. worse than chance.[5] A possible interpretation is that the model is not suited for generalization on unseen skill level values. In other words, it cannot predict values far from the ones used for its training, as the corresponding training sample is missing.

Based on these considerations, Tables 9.3 and 9.4 respectively show the correlations between annotations and predictions, without two extreme participants (P2 and P11) and without four extreme participants (P1, P2, P11 and P12). It can be observed in these tables that the predictions lead to a correlation of 0.573 without two extreme participants (see Table 9.3), and 0.809 without four extreme participants (see Table 9.4), both with the two-step transfer learning and MIRFE methods. These results suggest that with a larger dataset, including a large number of beginners and experts, the overall accuracy of the model could be improved.

[5] RMSE would be about 1.0 for $\mathcal{N}(0,1)$-random predictions for a standardized skill-level.

Method	**No MIRFE**	**MIRFE**
Direct transfer learning	0.322	0.461
Two-step transfer learning	0.319	**0.573**

Table 9.3: Prediction correlations with annotations for various skill-level-regression CNNs, all participants except P2 and P11.

Method	**No MIRFE**	**MIRFE**
Direct transfer learning	0.755	0.780
Two-step transfer learning	0.756	**0.809**

Table 9.4: Prediction correlations with annotations for various skill-level-regression CNNs, all participants except P1, P2, P11 and P12.

9.3.3 Comparison with the statistical-based model

In Chapter 8, a method was proposed for gesture evaluation, based on an EN-regression on PCs extracted from means and standard deviations of motion features. For comparison purposes, Fig 9.8 shows predictions results and RMSE for this model. This figure can be compared with Fig 9.7. It can be observed that for the statistical-based model, the two largest RMSE are obtained for the lowest-skilled participant (P11), and the best one (P2). However, all RMSE are below 0.7, showing that the statistical-based model can generalize from a small dataset consisting of only 12 participants. The next lowest/highest-skilled participants (P1 and P12 respectively) are evaluated as well as most of the other participants, with an RMSE of about 0.4.

9.3.4 Limitations and improvement prospects

In this chapter, the use of deep learning for gesture evaluation is explored. An adaptation of the method proposed in Laraba et al. (2017) is presented, and serves as a proof-of-concept. For this purpose, AlexNet was used as a basis. However, various architectures could be used instead, and vast exploration is still needed on the subject. Various types of intermediate layers could also be used, including combinations of convolutional and FC layers with different hyper-parameters.

However, a more in-depth exploration on the subject would require a larger dataset. In the present work, the number of samples for the training of the regression network was in average 208 for each Bafa technique. Nonetheless, a first insight of the interest of the proposed technique was highlighted with this dataset. It was shown that the performance of the model was significantly better than random predictions, and that the results were better for mid-level participants. It seems to show that better results could be achieved with a larger dataset, including more experts and beginners, allowing a larger coverage of the variability of skill levels in Taijiquan.

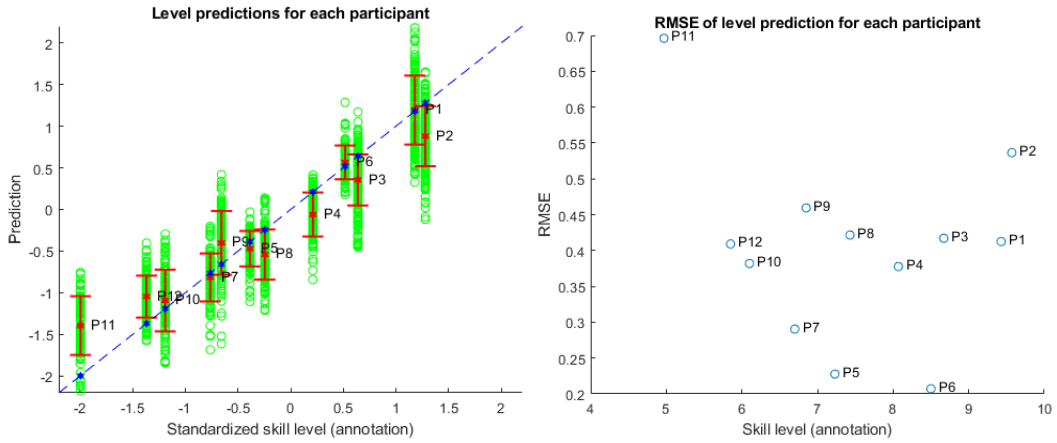

Figure 9.8: Prediction results for the statistical-based model from Chapter 8, based on EN-regression on 60 PCs extracted from means and standard deviations of global positions and relational features. Left: predictions of all motion sequences for each participant, against their skill level. Right: prediction RMSE for each participant, against their skill level.

With a larger dataset, various prospects for improvement could be explored. First, the use of more complex model architectures than AlexNet could be tested and compared. Secondly, a very large dataset would allow the development of new types of deep-learning-based models trained from scratch. These models could take advantage of the particular type of data that are motion sequences, including the spatial aspect, the higher relations between joints of the same limb, as well as the temporal aspect. For instance, the use of Recurrent Neural Network, particularly adapted for the modeling of time series, could be explored (LeCun et al., 2015). Particular convolutional filters could also be developed, taking into account the type of input data. For instance, particular CNNs with 1D filters have recently been used in speech processing for automatic translation (Tachibana et al., 2017). The input is a spectrogram, i.e. a large set of temporal variables, similar to a motion sequence. Finally, multitask learning could be explored. Multitask learning allows the simultaneous modeling of different types of output from a single type of input. The use of a shared representation in the model for the learning of different tasks can result in an improved efficiency for each separate task (Caruana, 1997). For instance, the use of information on the category of gesture or on the participants, including morphology, age or sex, would allow the model to take these factors into account in the modeling of expertise. This approach can be compared to MIRFE, though the process would be handled by the model itself.

In the proposed image representation of motion data, the order of the features or coordinates is arbitrarily defined. However, different results could be obtained according to the order of the variables, i.e. of the stripes. As convolutional filters are applied on neighbor pixels, the relations between neighbor variables are more implied at the beginning of the modeling process. In the present study, left and right

pairs of joints are alternated in a specific order from the head to the feet. It could be interesting to group the stripes of the joints that are closer in the skeleton. However, a new type of model specifically designed for motion data could directly take account of the skeleton topology of the data. For instance, convolution filters in the first layer could be applied separately on the joints of each limb or segment of the body, independently of the order of the variables.

9.4 Conclusion

In this chapter, a new gesture evaluation model based on deep learning is presented. For that purpose, motion sequences are represented as RGB images, for their use with pre-trained image classification models. The proposed model is based on AlexNet, and a two-step transfer learning procedure is applied in order to adapt it for skill level regression. First, the model is modified to allow classification of the eight Bafa techniques, and is trained on the entire dataset. Secondly, the classification layer is replaced by a regression layer, allowing the prediction of the skill level. A model is trained separately for each Bafa technique. The results showed that the two-step transfer learning procedure coupled with the use of MIRFE on the input features allowed better predictions than direct transfer learning. Although the results are significantly lower than those of the method presented in Chapter 8, it could be observed that the predictions were better for mid-level participants than for extreme ones (the lowest-skilled and the highest-skilled). A larger dataset including more experts and beginners could hence lead to an improvement of the performance of the model. These results show the potential of the use of deep learning for gesture evaluation applications.

Chapter 10

Towards a generic visual feedback model for gesture evaluation

Contents

10.1 Introduction

Gesture evaluation has applications in various areas, such as medicine, education and serious gaming. It can provide a quantified measure of the quality of a gesture, allowing both the tracking of a patient's progression, and the objectivity of the evaluation. However, more information could be obtained from data than a single score. Feedback information could be extracted from the evaluation model about how the gesture was performed, and how it should be modified to produce a better skill level prediction. Users of such a feedback system could benefit from this more interpretable information, allowing a more efficient training. This system could be used either as an automated supervisor, or as a tool for teachers.

However, due to the newness of MoCap technologies, the small but emerging literature on gesture evaluation as presented in Chapter 3 is not yet focusing on the use

of the proposed evaluation models for actual training and supervision. Nonetheless, the few works mentioned in Section 3.4.4 establish some reference points in that direction. Young and Reinkensmeyer (2014) proposed an original method for synthesis of a new motion corresponding to a particular skill level. From a score, they generate features (eigenpostures and PMs) that would predict this score (see Section 3.4.3), and then synthesize the corresponding gesture from a 'reciprocal process' of the eigenpostures (following a reciprocal procedure of Troje (2002) for eigenpostures extraction). Although they used this system to produce general interpretations on the analyzed discipline (competitive diving), the synthesized gestures could be compared to the user's own performance, allowing feedback. In that direction, Pirsiavash et al. (2014) used the same idea of 'reciprocal process' of their evaluation model (based on pose-DCT and L-SVR , see Section 3.4.3) to generate new features corresponding to the gradient of the model. They first compute the gradient of the L-SVR model, and then perform an inverse-DCT on this gradient, providing information on the modification of the joint positions that would improve the score.

In this chapter, a novel generic method for visual feedback based on synthesis is proposed. From a given score, a new motion sequence (or a new feature set of any type) is synthesized from a weighted average of gestures from a pre-processed dataset. The synthesis method, based on GRNN, does not require the 'reciprocal process' of any specific model or features, and can hence be used with any gesture evaluation model that allows the prediction of a continuous score. The method could even be used for direct interpretation on experts' annotations, without any gesture evaluation model.

From the synthesized data (either a motion sequence or a feature set), three types of feedback information are generated: (i) a visual feedback, based on the synchronized visualization of the user's performance with the synthesized motion sequence, (ii) a striped-image representation of the differences between the user's features and the synthesized ones, and (iii) a striped-image representation of the Euclidean distances between the corresponding markers of both motion sequences.

10.2 Method

10.2.1 Synthesis-based feedback loop

Fig 10.1 shows the diagram of the proposed feedback method. First, a motion sequence performed by a user is recorded and the skill level of the gesture is predicted using an evaluation model. Any of the regression model presented above can be used. An increment (a *hop*) is then applied on the predicted score, resulting in an improved score. From this improved score, a synthesizer generates a new artificial motion sequence and/or a new feature set corresponding to this score. The synthesis is based on a weighted average of sequences/features extracted from the dataset.

The synthesized data are then compared with the original ones, allowing extraction of various types of feedback:

1. The synthesized sequence is superimposed with the original one, and both can be visualized synchronously. It allows a direct visual comparison showing the differences between both sequences to the user.
2. The difference between the original features and the synthesized ones is computed, and is illustrated as a striped image with scaled colors, where each stripe corresponds to the temporal evolution of a feature. Note that the features can be joint global 3D positions. A raw motion sequence can be seen as a particular type of feature set.
3. The Euclidean distances between each corresponding marker of the synthesized and original sequences are computed, and are also illustrated as a striped image (one stripe per joint).

10.2.2 Skilled-gesture synthesis

The key part of the proposed feedback system relies on the synthesis of a gesture corresponding to a particular skill level. To synthesize new motion data corresponding to a given skill level, a GRNN (or Gaussian-kernel regression) is used for the regression of a motion sequence with the skill level as input variable. Fig 10.2 illustrates the underlying process of the proposed synthesis method. For each sequence of a pre-recorded dataset, scores are computed using the evaluation model from which feedback is requested. Annotations could be used instead if they are provided for each sequence.[1] For a new sequence performed by the user of the system (referred to below as the *test sequence*), the skill is evaluated, and a hop (h) is applied to this score, as explained in Section 10.2.1. All these scores, denoted by s_i, are then used as input for the regression. The output of the regression is either a motion sequence (termed below as the *feedback sequence*), denoted by X_s, or any type of feature set F_s. Note that a motion sequence can be seen as a particular type of feature set. The same approach can thus be used for both.

The dataset of motion sequences is adapted for a more efficient weighted combination of several motion sequences, and for optimal comparison with the user's gesture, as illustrated in Fig 10.3:

1. First, all motion sequences are scaled to the size of the user using skeleton scaling.[2]

[1] If only general annotations are given for the participants of the dataset, only a few different values (12 in the case of the Taijiquan dataset) would be used as input for the training of the GRNN (i.e. as kernels), resulting in a sparse regression space.

[2] MIRFE cannot be used for this purpose as it only extracts residual features that would result in an irrelevant 3D visualization. Nevertheless, MIRFE is used on the features used for the stripe image displaying feature differences.

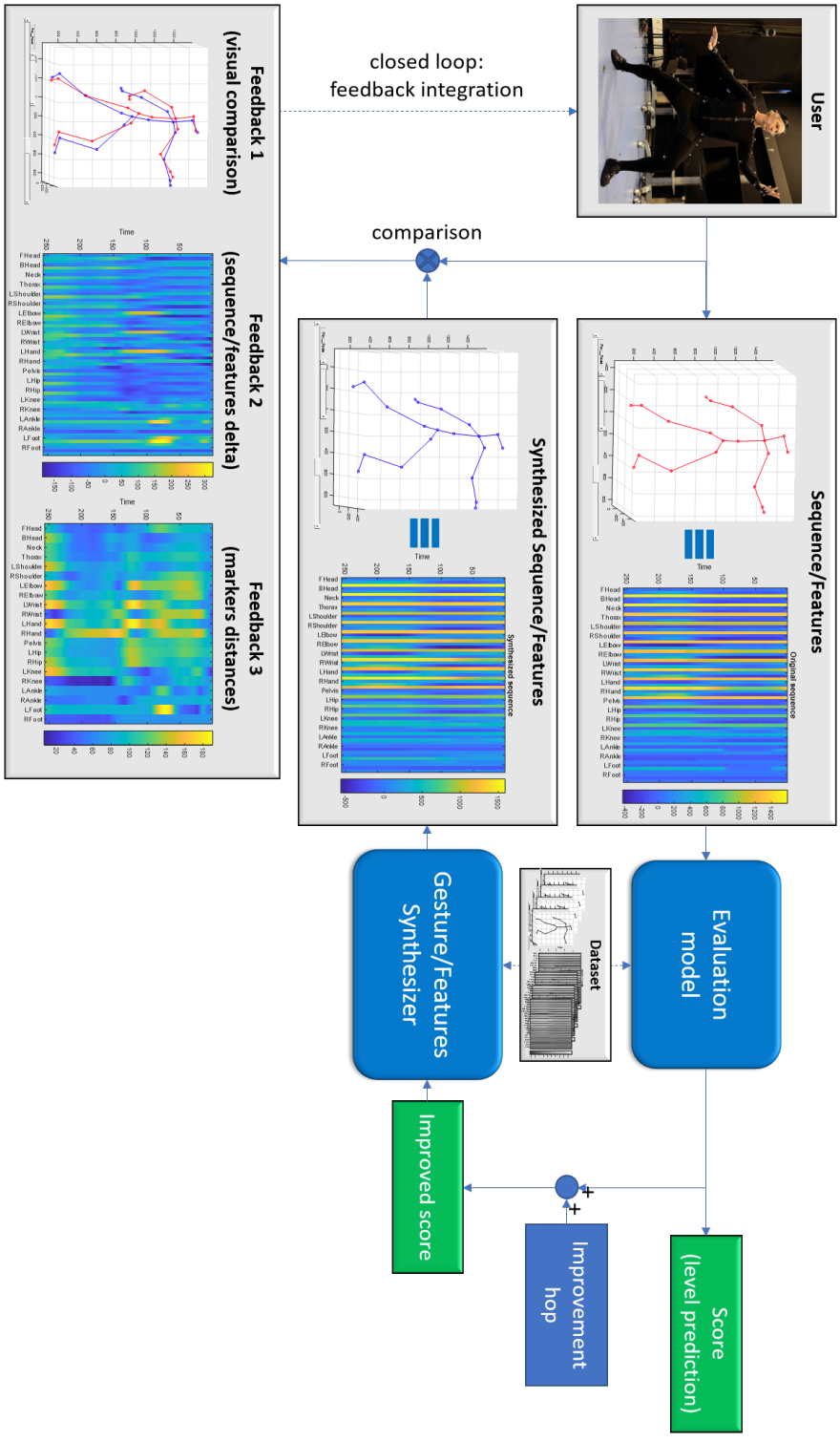

Figure 10.1: Synthesis-based feedback loop process. The integration of the feedback by the performer can be viewed as a conceptual closed-loop process for the user's progression.

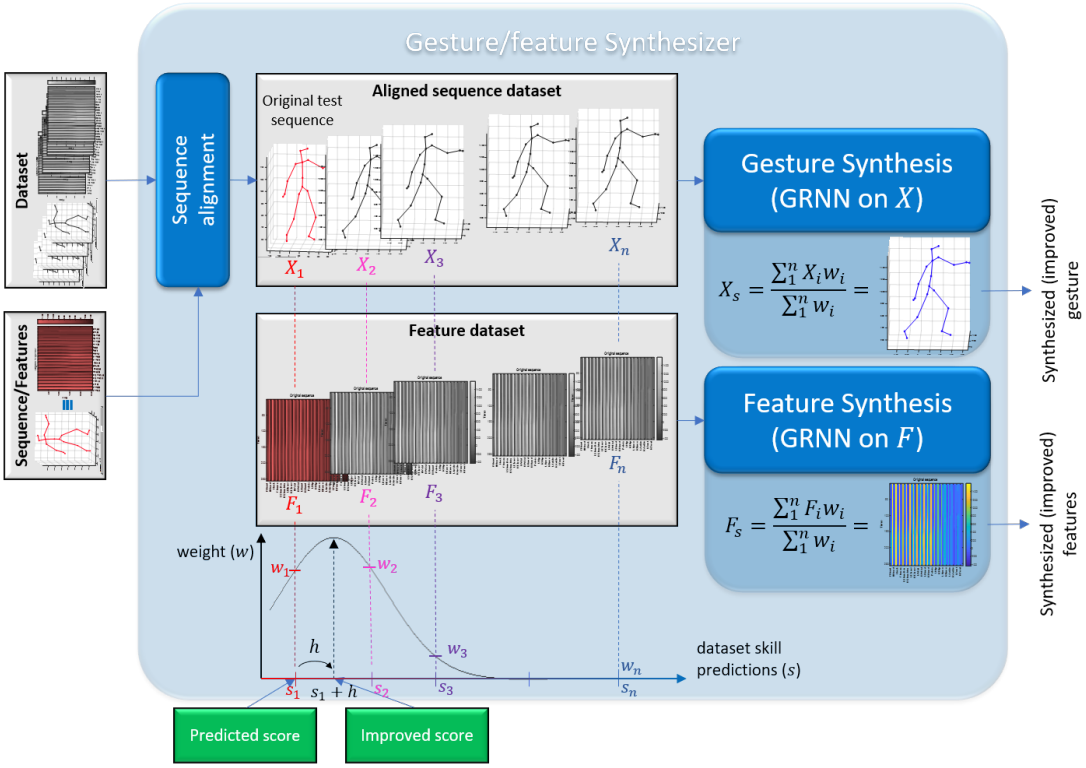

Figure 10.2: Skilled gesture synthesis workflow.

2. For each motion sequence of the dataset, a rigid transformation of the marker horizontal coordinates (x and y-axes) is applied to minimize the sum of the distances of the markers with the corresponding markers of the test sequence. This transformation is performed by using the Kabsch algorithm (Kabsch, 1976), allowing the best fitting of two paired sets of points. The algorithm is performed only on the horizontal coordinates to avoid a translation or rotation on the vertical axis. This would lead to a counterintuitive visualization of the generated motion sequence, with a virtually modified ground elevation and inclination for the synthesized sequence.

3. Each motion sequence of the dataset is then aligned temporally with the test sequence using DTW. This step allows a more relevant synchronized visualization of the final feedback.

As illustrated in Fig 10.2, the test sequence itself is then added to the dataset (denoted by X_1), so that the synthesis can produce a more similar gesture if a close skill level is provided as input of the synthesizer (i.e. for a small h). This step allows a more relevant comparison of the feedback sequences with the test sequence.

Finally, new motion data X_s and/or F_s are synthesized from the improved score $s_1 + h$ using the following equations:

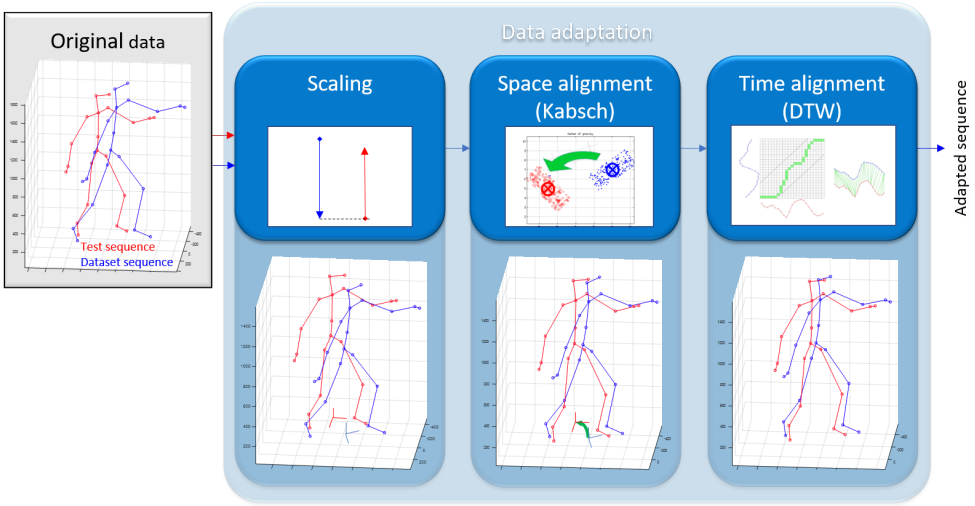

Figure 10.3: Data adaptation for better comparison with the user's motion. The red skeleton represents the test sequence performed by the user of the feedback system. The blue skeleton represents a sample of the dataset. Firstly, the sample is scaled to the size of the user. Then, horizontal coordinates are fitted to the test sequence using the Kabsch algorithm. Finally, data are aligned temporally using DTW.

$$w_i = exp(-\frac{(s_i - (s_1 + h))^2}{2\sigma^2_{smooth}}) \tag{10.1}$$

$$w'_i = \frac{w_i}{\sum_{i=1}^{n} w_i} \tag{10.2}$$

$$X_s = \sum_{i=1}^{n} X_i w'_i \tag{10.3}$$

$$F_s = \sum_{i=1}^{n} F_i w'_i \tag{10.4}$$

where σ_{smooth} is the smoothing parameter of the GRNN. This parameter was manually set to 0.3 in the experiment. The choice of this parameter is discussed in Section 10.4.

The synthesized motion sequence X_s can be visualized in synchronization with X_1 (Feedback 1, see Fig 10.1), the distances between their corresponding markers can be visualized as a striped image (Feedback 3, see Fig 10.1), or the difference between F_s and F_1 can also be visualized as a striped image (Feedback 2, see Fig 10.1).

10.2.3 Experiments

To validate the proposed synthesis-based feedback method, the skill-level of the synthesized data can be evaluated with the gesture evaluation model. For this experi-

ment, the best gesture evaluation model proposed in this research is used (feature statistics PCs EN-regression, see Chapter 8). The resulting score can then simply be compared to the input score of the synthesizer. As an experiment, feedback sequences are generated for each sample of the 8-Bafa dataset, for five improved scores from the initial predicted score ($h = 0$) to 12 ($h = 12 - s_1$). On these synthesized sequences, the level is predicted using the evaluation model, allowing to verify the assumption that the generated sequence is actually better than the test sequence. The results are presented in Section 10.3.1.

Besides quantitative validation, some examples are qualitatively interpreted, to show the interest of the proposed feedback method. These examples are presented in Section 10.3.2.

10.3 Results

10.3.1 Quantitative validation

Fig 10.4 shows the means of the level predictions of the synthesized sequences for each participant, each Bafa technique, and for different improved scores (referred to as *target score* in the graph). For each graph, corresponding to the results for one participant, the blue star shows the reference level of the participant (as annotated by the teachers). Each curve, referred to below as *feedback curve*, corresponds to the mean prediction of the feedback sequences for all renditions of one gesture by one participant, according to the improved score. The origin of each feedback curve corresponds to a feedback without any improvement ($h = 0$). The x-axis thus corresponds to the predicted score of the original sequence (s_1), and the y-axis to the prediction of a synthesized sequence corresponding to this level. Therefore, each curve should ideally start from the blue star. Then, as the improved score is increased ($h > 0$), the synthesized sequence should be of a greater quality, leading to a higher prediction, ideally equal to the improved score. The curves should hence follow the dashed line. It can be observed that for most of the participants, the curves seem to follow this line at the beginning. As the improved score reaches the maximum score from the annotation scale (10), the curves flatten and generally tend towards this maximum. This result is due to the fact that the feedback sequences are generated through an averaging of the available sequences in the dataset. As the best sequences of the dataset are rated about 10 by the gesture evaluation model, corresponding to the skill level of an expert, the synthesizer cannot generate sequences with a higher score. Note that for P1, some curves exceed the dotted line. This is due to the fact that some predicted scores from the available dataset (P2 to P12, following the LOPO procedure) are higher than 10.

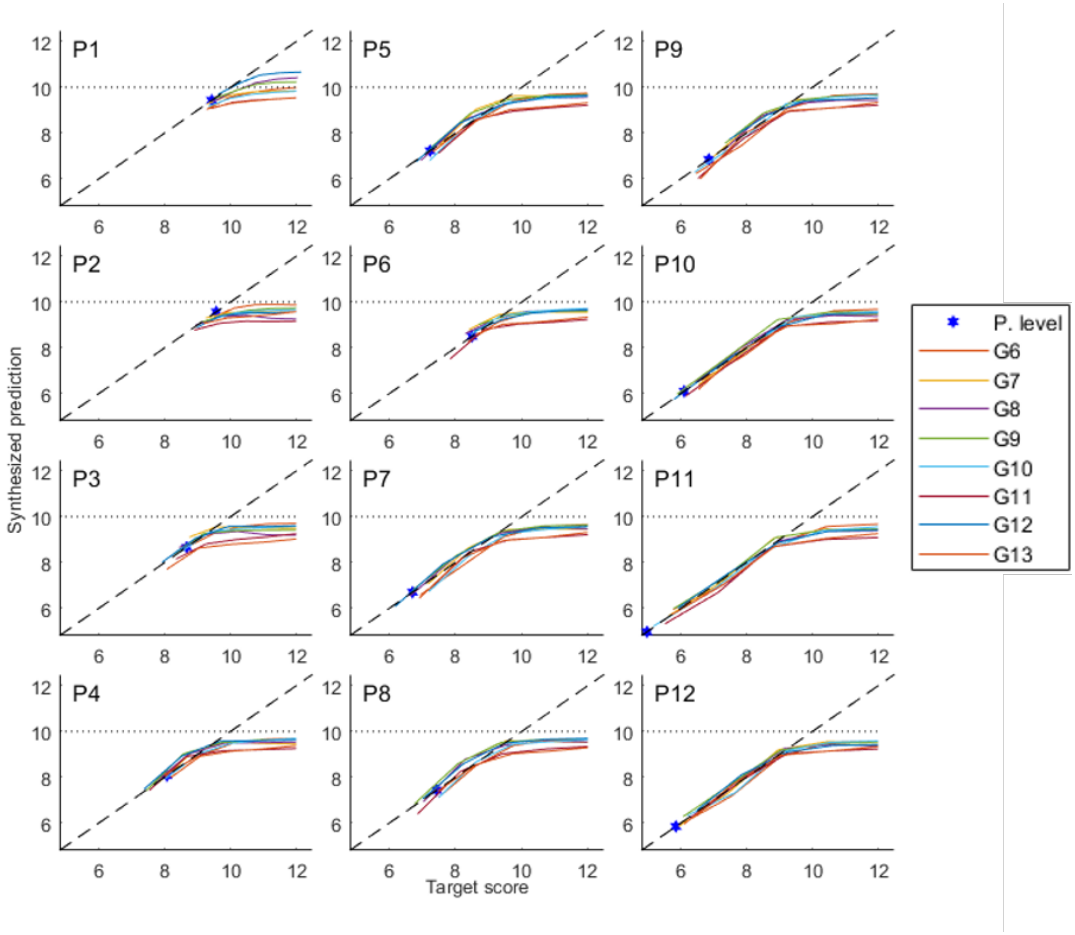

Figure 10.4: Validation of the feedback method. The level of the feedback sequences is predicted back by the gesture evaluation model (y-axis) and confronted to the improved score (the target score, x-axis)). Each curve corresponds to the mean of the predictions for all renditions of one Bafa technique by one participant.

10.3.2 Qualitative validation

In this section, a few examples of the use of the feedback system will be presented, in order to show its practical interest as an improvement tool for the user. Fig 10.5 shows the result of the feedback system for a rendition of G8 (Part the wild horse's mane) performed by P11, the lowest-skilled participant. The prediction of the original sequence was $s_1 = 5.55$. A hop of $h = 4.45$ was applied to generate a sequence corresponding to an improved score of 10. The gesture evaluation system was then used on the synthesized sequence and predicted an actual skill level of 8.58. The left graph shows a frame at the beginning of the gesture. It can be observed that the original skeleton (in red) is tilted back. On the opposite, the feedback skeleton seems more stable, with a more vertical trunk (cfr stability features, see Section 6.2). The right graph shows a frame at the end of the gesture. At this moment of the gesture, the feet of the feedback skeleton seem to be further away, both in frontal plane (sideways) and in sagittal plane (forward-backward direction), offering a larger support base. The attacking arm (left arm) is aligned vertically with the attacking foot (left foot). The head and the torso are also directed towards the attacking limbs (cfr alignment features, see Section 6.3). Both arms and both legs are more widely opened, and are more similar to the shape of a sphere. This spherical shape of the body is an important concept of Taijiquan (Caulier, 2010). Although the gesture evaluation model is not trained with any particular feature involving this spherical shape, the visual feedback seems to allow this observation, showing its interest. Even though the algorithm is based on rather low-level features interpretation (global positions and relational features), high-level interpretations can still be made by the human observer from the synthesis-based feedback system. It is important to note that these interpretations are made from an interactive 3D visualization of the sequences, allowing a better observation than allowed in the 2D snapshots shown in Fig 10.5.

The system also allows a visualization of the feedback sequence evolution, for different values of h. Fig 10.6 shows five feedback sequences corresponding to a linear ramp of h from 0 to 4.45. The colors of the feedback skeletons follow a mapping from red (for $h = 0$) to blue (for $h = 4.45$). It can be observed that for each increment of h the feet move away from each other, until they reach their optimal position in blue. It seems that for the first increments of h, the displacement of the feedback sequence is larger. The differences between the feedback sequences seem more subtle after a few increments. The difference between the last two increments seems to essentially lie in the head direction, and the wider opening of the left arm.

Fig 10.7 shows the results of the feedback system for a rendition of G11 (Kick with the heel) performed by P11 (larger images are provided in Appendix B with axes, for better readability). Graph (a) shows 3D visualization of the original and feedback sequences at the top of the kick (at about 45% of the sequence duration). It can be observed that the feedback sequence, in blue, has a higher kicking foot. Moreover, its left hand is vertically aligned with its left foot, and the head has the same direction.

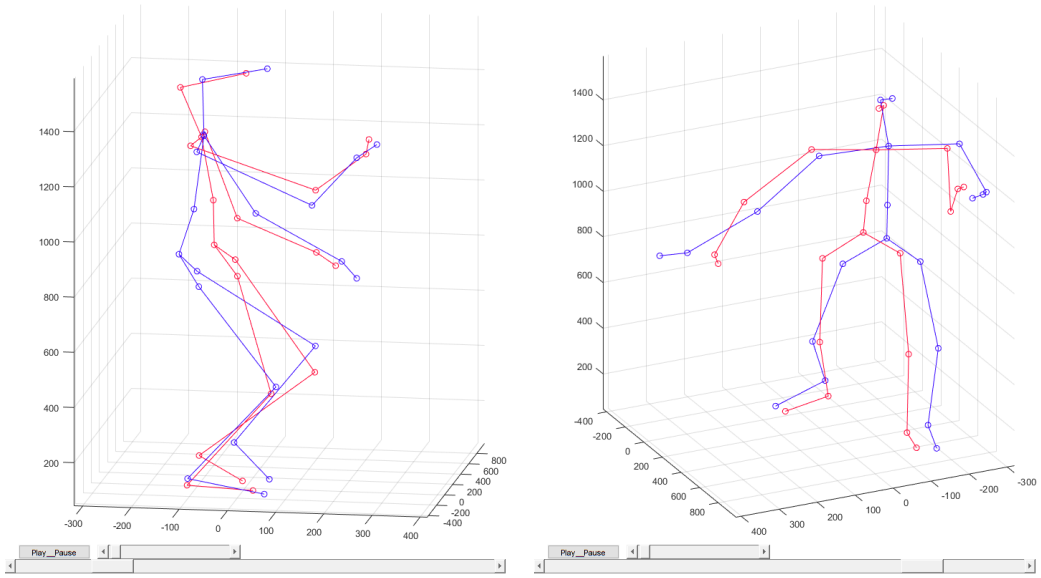

Figure 10.5: Visual feedback for a rendition of G8 (Part the wild horse's mane) by P11 (the lowest-skilled participant). Red: original sequence. Blue: feedback sequence with an improved score of 10 ($h = 4.45$). Left graph: gesture beginning, side-view. Right graph: gesture ending, front-view.

Both arms are opened more widely, and the position seems more stable, as the trunk is vertical.

Graph (b) displays the distances between the markers of the original sequence (red skeleton from Graph (a)) with the markers of the feedback sequence (blue skeleton in Graph (a)) across time. A large distance can be observed for the left foot, left ankle and left knee joints throughout the sequence. A peak for these distances occurs after about 60% of the sequence duration. A 3D visualization of the corresponding frames showed that this peak is due to the fact that the user laid down the left foot sooner than the feedback sequence. In other words, the left foot of the feedback sequence stays high for a longer time than the left foot of the original sequence. Two other points of interest can be raised from this graph, one for the left arm (especially the left hand), and one the right arm. A look at Graph (a) confirms this, as both arms of both skeletons are far apart.

Graph (c) displays the differences between joints global coordinates. It offers thus a few more information than the previous graph, emphasizing the important axes. It can be observed more clearly on this graph that the errors on the left leg joints seem synchronized with the errors on the left arm. Moreover, it can be seen that major differences stand for the z-axis (vertical) for the left leg joints, and the y-axis (lateral) for both arms.

Finally, Graph (d) displays the differences between some of the Taijiquan ergonomic principles. Though they were not the best feature subset for use with the proposed

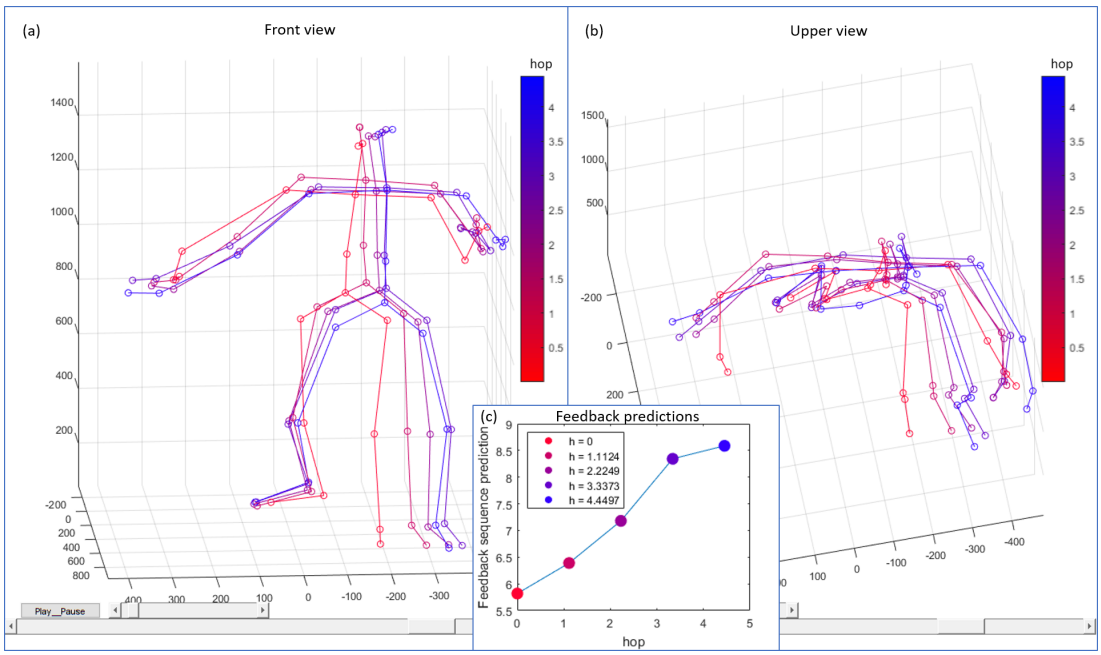

Figure 10.6: Visual feedback for a rendition of G8 (Part the wild horse's mane) by P11 (the lowest-skilled participant), for different values of h (linear ramp of five values from 0 to 4.45). The colors of the skeleton follow a color map from red to blue corresponding to h (0 = red, 4.45 = blue). (a): front view. (b): upper view. (c): feedback predictions.

gesture evaluation models, they still allow an interesting feedback that can be interpreted by a user or a teacher. For instance, a first deep blue stripe concerning verticality (F_3 in Table 6.1) appears throughout the entire sequence. This stripe indicates that the trunk should be more vertical, to provide more stability to the position. A second blue stripe indicates a wrong vertical alignment between the right shoulder and the right hip (F_6 in Table 6.1). A third deep blue stripe indicates a wrong frontal alignment of the left shoulder with the left wrist (F_{11} in Table 6.1). Finally, the fourth deep blue stripe indicates that at the end of the gesture, the left elbow is too far behind the body (F_{18} in Table 6.1). All these indications can generally be confirmed by a visualization of the 3D sequences.

10.4 Discussion

In this chapter, an original method for gesture evaluation feedback is presented. The method is validated quantitatively and qualitatively.

For the quantitative validation, the synthesized feedback sequences are evaluated back using the gesture evaluation model, in order to verify if the predicted skill level corresponds to the desired skill level provided at the input of the synthesizer. The results are consistent, as the prediction generally follows the desired skill level, and are bounded to the maximum level available in the dataset. Nonetheless, the smoothing parameter of the synthesizer (σ_{smooth}, see eq. 10.1) could be tuned according to the desired behavior of the synthesizer. A lower σ_{smooth} would theoretically allow the synthesizer to generate sequences better corresponding to the input score. As σ_{smooth} is increased, w_i in eq. 10.1 tends towards 1, giving the same weight to all the sequences of the dataset (flat weighting). This allows a more generalized feedback sequence, as more sequences from the dataset are taken into account. However, the quality of this sequence is also more limited. For an infinite σ_{smooth}, the synthesizer would always produce the same sequence resulting from the average of all sequences of the dataset, corresponding to the average skill level of the dataset. This would result in a flat feedback curve. On the opposite, as σ_{smooth} tends towards zero, the weighting becomes sharper and only the sequence with the closest prediction to the targeted score is considered in the synthesis. The synthesized sequence is therefore simply a copy of this sequence, and the evaluation will inevitably lead to the same score. In this case, the feedback curve will follow a step function following the dashed line and limited to the predictions available in the dataset. In this case, the synthesized sequence will not be generalized at all, as it will copy a single rendition of a gesture by a single participant.

As qualitative validation, a few examples are presented, showing the practical interest of the method. However, to prove the effectiveness of the method, it should still be tested in a real situation, by comparing its impact on actual training, with

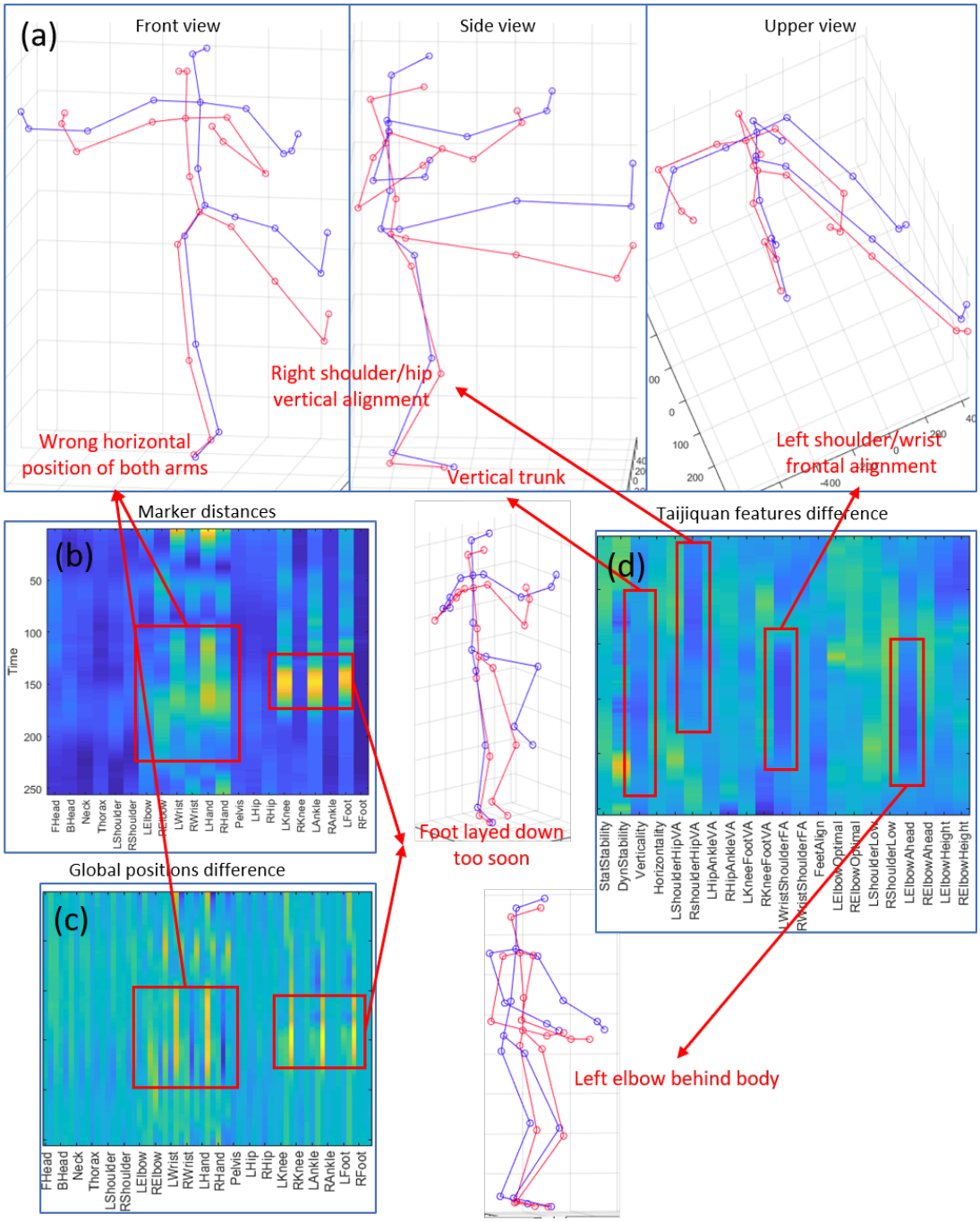

Figure 10.7: Feedback for a rendition of G11 (Kick with the heel) by P11 (the lowest-skilled participant). (a): Visual feedback. (Red: original sequence. Blue: feedback sequence with an improved score of 10). (b): Marker distances. (c): global positions difference. (d): Taijiquan features difference.

self-training and with teacher-supervised training. Nonetheless, the examples provided in Section 10.3.2 give first indications on the possibilities offered by the proposed method. The synchronized visualization is an intuitive feedback system that can be easily interpreted by a novice user. Even without a priori knowledge on the discipline, the user could learn a technique by successive imitation of the feedback sequences, iteratively adapting her/his own gesture. A particular interest of the proposed method is that the synthesized feedback sequence is comparable to the user's, as it was generated from adapted data, and from the user's data themselves. As the feedback sequence can correspond to any small improvement step (small h), a gradual feedback can be provided to the user, allowing a step-by-step progression. Besides the synchronized visualization, the striped images serve as a more precise feedback. They allow to highlight particular features that require modification, giving precise indications to the user. However, these images are harder to interpret than the synchronized visualization. This type of feedback could have more interest for advanced users or teachers who already know which kind of feature should be relevant in a particular technique, allowing them to focus on the relevant parts of the stripe image. Another solution could be a state-machine (similar to Patrona et al. 2018) that would provide a semantic feedback from pre-defined zones of the image. For instance, for the 'Kick with the heel' gesture, there is not much interest in features indicating the motion of the non-kicking foot as it is supposed to stay still. For standardized features, a relational feature such as 'the right foot is raised' (F_{17} in Table 2.5) could lead to a large feature difference (i.e. a stripe with a strong color) even if the synthesized foot is a few millimeters higher than the original one. Moreover, the differences near the beginning or the ending of the sequence may be due to the inaccuracies of the motion sequence segmentation procedure, and not to a wrong position of the user.

To improve the feedback method, it could be interesting to derive from the model the importance of each feature for the prediction. This importance could be obtained analytically for some models. For instance, in the model proposed in Chapter 8, the importance of the feature could be derived from the regression coefficients and the weight of features in the PCs. Otherwise, the importance of each feature could be extracted experimentally, by analyzing the impact of a modification of a feature in the prediction. From the importance of each feature, a weighting of the colors in the stripe images could be applied, in order to propose sharper images, easier to interpret. The synchronized visualization could also be adapted by highlighting the most important joints (e.g. by modifying their color, size or transparency).

10.5 Conclusion

In this chapter, a novel feedback method is proposed. The method is based on the synthesis of new data, either a motion sequence or a feature set, corresponding to a

desired skill level. To synthesize these data, a weighted average of samples of a pre-processed dataset is performed. The motion sequences of the dataset are adapted with three steps, to be more comparable to the user's motion (the test sequence). First, each motion sequence of the dataset is scaled to the size of the user. Secondly, each motion sequence is spatially matched to the test sequence using a rigid transformation minimizing the distances between the horizontal coordinates of the corresponding markers. Thirdly, each motion sequence is temporally aligned to the test sequence using DTW. Different types of feedback are then extracted by comparing the synthesized data to the user's ones: (i) a synchronized visualization of the test sequence with the feedback sequence synthesized with any hop of skill improvement, (ii) a striped image representing the differences between the synthesized features and the user's ones, and (iii) a striped image displaying the distances between the corresponding markers of both sequences.

The proposed model has different advantages compared to the related work. First, it can be used either with motion sequences, or with any type of motion features. Secondly, it is independent of the gesture evaluation model. It could be used with any gesture evaluation model providing a continuous score. It could even be used for direct interpretation of the scores annotated by teachers, without the need of any evaluation model.

As quantitative validation, various feedback sequences were generated for different desired skill levels, and their effective skill level was then estimated by the evaluation model. The comparison of the prediction with the desired skill level showed the effectiveness of the synthesis method.

Examples of the use of the feedback method were then proposed, showing its practical interest for training. The proposed visual feedback is intuitive and can be used by a novice user. However, the striped-image representations allow a specific feedback on various features, and allow a more precise interpretation by advanced users or by teachers. The proposed method could hence be used as an automated supervisor, or as a tool for a teacher, allowing a more objective supervision.

Finally, the method could be improved in the future, by weighting the feedback for each feature according to their importance in the used gesture evaluation model. This would allow a better highlighting of the most relevant features in the striped images, or the most relevant joints in the motion sequences.

Conclusions

Original contributions of the thesis

In the present thesis, a framework for the evaluation of the expertise in gestures has been proposed. This framework follows six main sequential steps.

First, a dataset of Taijiquan expert gestures has been collected and is presented in Chapter 4. Taijiquan is a general discipline focusing on various aspects of motion. This dataset is an important contribution to the field, as no public dataset known to the author was available for the study of gesture evaluation. The proposed dataset contains a total of 2200 manually corrected and segmented gestures, divided into 13 classes and performed by 12 participants of different expertise levels. The participants were ranked by three highly experienced Taijiquan teachers, allowing the training and validation of new evaluation models.

To deal with the issue of missing markers and improve the data quality, a method for automatic MoCap data recovery has been proposed, based on the probabilistic averaging of various methods (PMA) and simple constraints on trajectory continuity and marker distances (see Chapter 5). Results show that PMA used with the constraints outperforms methods used individually in various conditions, including various gap lengths, motion sequence durations and number of simultaneous gaps, showing the robustness of the method.

The next step was the choice and the development of new motion features, adapted for the representation of expertise. In this context, a large set of features from the literature was used, including various low-level features as well as relational features and ergonomic features. Moreover, a new feature set was developed, inspired by Taijiquan ergonomic principles (see Chapter 6). This high-level feature set allows a representation of motion in terms of ergonomy, and is assumed to be related with expertise. However, the best performance in gesture evaluation was not achieved with this type of feature, showing that global positions and relational features were more appropriate with the models tested. Further tests should be conducted with a larger dataset to investigate the use of these high-level features. It is possible that their relations are complex and that non-linear regression methods would perform better with these features if a larger dataset is available.

Nevertheless, these features are relevant for feedback interpretation, at least in the case of Taijiquan gesture evaluation, as illustrated in Chapter 10. They are of particular interest to the teacher used to this type of motion representation in the teaching of

Taijiquan. In order to validate their practical interest, a prototype of infield feedback should be developed, and be tested with various types of features. Moreover, infield tests should also be conducted in other disciplines than Taijiquan.

The most relevant representation of expertise must then be drawn from the available features. To that end, a novel method for morphology-independent feature extraction has been developed (MIRFE, see Chapter 7). The method has been shown to increase the correlation of features with expertise. Furthermore, MIRFE was validated with various evaluation models, and generally improved their performance. More specifically, it was shown that evaluation models could be used efficiently with a larger number of latent variables (PCs or eigenmovement weights), leading to significantly better predictions. It seems therefore that MIRFE allows the extraction of features that are easier to interpret by the models. This may be due to the fact that the processed features are more comparable between participants, or due to the reduction of useless and redundant information contained in the features about the participants' morphology.

From these processed features, a new statistical-based model for the automatic evaluation of expertise has been proposed (see Chapter 8). Basic statistics are computed for each sequence, resulting in a mean and a standard deviation per sequence. PCA is then applied on these statistics, providing a smaller set of variables from which a regression model can be trained to estimate a performer's expertise level. The proposed model has been designed to be generic, and can therefore be used with any type of features, and with various regression models. Both linear and non-linear regression models were tested and validated with the Taijiquan dataset. The best prediction accuracy ($R = 0.909$) was obtained for an EN-regularized linear regression, with 60 PCs extracted from global positions and relational feature statistics.

A first exploration of the use of deep learning algorithms was proposed for gesture evaluation (see Chapter 9). To that end, motion sequences were represented as RGB images, allowing their use with pre-trained image classification models. AlexNet was used as proof-of-concept, and a two-step transfer learning procedure allowed its adaptation for the classification of Bafa techniques first, and then for the prediction of the level of expertise. The results seem to show that the proposed two-step transfer learning procedure yields better accuracy than a simple transfer learning. Although the results were significantly lower than those of the statistical-based model (see Chapter 8), it could be observed that middle-level performers were evaluated more accurately than the lowest- and highest-skilled ones. This finding suggests that a larger dataset including more beginners and experts could allow for a better generalization of the model to new instances, showing the potential of deep learning for gesture evaluation.

Finally, as a practical application of the gesture evaluation models, a novel feedback method has been proposed (see Chapter 10). From a test sequence performed by a user of the system, a feedback sequence or feature set is synthesized, corresponding

to a desired skill level, and comparable to the user's performance. Three types of visual feedback are then proposed, allowing: (i) a synchronized visualization of the test sequence with the feedback sequence, (ii) a striped image representing the differences between the synthesized features and the user's ones, and (iii) a striped image displaying the distances between the corresponding markers of both sequences. The proposed method has as advantage that it can be used with any gesture evaluation model providing a continuous score as input for the synthesizer. Moreover, it can be used with any type of motion features.

The method was validated quantitatively by verifying that the synthesized sequences corresponded to the skill level predicted by the evaluation model. Moreover, as qualitative validation, examples of the use of the feedback method were presented, showing its practical interest for training. The proposed visual feedback is intuitive and could be used either by a novice user for automated supervision, or by a teacher as a tool allowing a more objective supervision.

General limitations and improvement prospects

The present thesis provides original elements to the recent research field of gesture evaluation. Moreover, the first three steps of the proposed framework, including the Taijiquan MoCap dataset, the MoCap recovery method and the morphology-independent feature extraction process (MIRFE) may have interests in a wider area of research with MoCap data.

A complete pipeline of methods has been successfully developed, from the recording of expert gestures to expertise evaluation and feedback. The evaluation results are promising, as well as the feedback examples, showing a possible integration in practical applications. However, various limitations must be raised, and numerous steps are still needed for an effective use in hospitals, sports fields, serious games, etc.

First, the proposed methods should be tested on other types of gestures. Although Taijiquan is assumed to be a relevant use case, a test on other gestural disciplines would allow to verify if the methods are effective with them. It would also allow to test if the same configurations in the evaluation models yield the best results (feature type, regression model used, MIRFE), and if the feedback method allows a relevant interpretation. According to the disciplines, other methods from the literature might also be more relevant. For instance, the two methods used for comparison in Section 8.3.4 rely on frame-by-frame relations, and might be more relevant for disciplines mainly focusing on timing.

Secondly, all results are limited to the size of the Taijiquan MoCap dataset. This dataset comprises about 208 sequences for each of the eight Bafa techniques, performed by 12 participants. Moreover, the annotation is limited to a global score for

each participant, and a score per sequence would be more relevant. For the modeling of complex relations with multiple non-linear regression, a larger dataset would be better. More complex relations systematically require more data to be efficiently modeled. In the present case, the best results obtained for non-linear regression models in Chapter 8 were based on a single PC (see Table 8.1). It is possible that with a larger dataset, a larger number of PCs could be used efficiently in these regression models. In that case, different types of features, describing abstract aspects of motion (such as ergonomics) and non-linearly related to expertise could yield better results. Furthermore, deep learning models, which are the state of the art in many other applications, could also be investigated more thoroughly.

Finally, no validation was proposed for practical infield use of the gesture evaluation and feedback methods. Their interest, either for automated supervision or as tools for teachers, must still be demonstrated. Practitioners of a discipline may all have their own perception of a gesture and their own learning strategy, based either on feeling, observation and intense repetition, and personal mental images. These strategies may also depend on the level of expertise already reached by the practitioner. According to their level, they may focus either on general aspects of motion or on specific details. Moreover, as suggested by Caulier (2010) and as discussed in Chapter 1, the mastering of a discipline can be divided in various steps, including the body external mechanics, the internal feeling and mental images, as well as a spiritual aspect related to concentration and awareness. The proposed methods, focused exclusively on external aspects of motion, could be more relevant in some cases, and for some users, than others.

A practical validation of the proposed methods based on an infield use by an end user would be complicated at the present stage of development. The accurate recording of the gestures with a MoCap system like Qualisys is expensive, complicated to set up, and too intrusive for various applications. Moreover, an automatic and user-friendly pipeline from the recording to gesture evaluation and feedback must still be developed.

Nonetheless, recent technologies, including real-time human 3D mesh recovery in 2D images (Güler et al., 2018; Kanazawa et al., 2018) show promising prospects for the development of low-cost and user-friendly MoCap. This type of technique could be used in the future with a single camera, either for the fast recording of a dataset, the recording of multiple performers at the same time, or for easy infield use. Other recent MoCap technologies such as the Vive Trackers™ allow a friendly use and are calibrated with virtual-reality headsets, such as the HTC Vive™ (HTC, 2018). Coupled with virtual reality, a new and more intuitive type of feedback could be investigated.

References

Abdullah, M.; Maliki, A.; Musa, R.; Kosni, N.; Juahir, H.; and Mohamed, S., 2017. Identification and comparative analysis of essential performance indicators in two levels of soccer expertise. *International Journal on Advanced Science, Engineering and Information Technology*, 7, 1 (2017), 305–314.

Ahmed, N.; Natarajan, T.; and Rao, K. R., 1974. Discrete cosine transform. *IEEE transactions on Computers*, 100, 1 (1974), 90–93.

Aizerman, M. A., 1964. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control*, 25 (1964), 821–837.

Alborno, P.; De Giorgis, N.; Camurri, A.; and Puppo, E., 2017. Limbs synchronisation as a measure of movement quality in karate. In *Proceedings of the 4th International Conference on Movement Computing*, 29. ACM.

Alexiadis, D. S. and Daras, P., 2014. Quaternionic signal processing techniques for automatic evaluation of dance performances from mocap data. *IEEE Transactions on Multimedia*, 16, 5 (2014), 1391–1406.

Andreoni, G.; Mazzola, M.; Ciani, O.; Zambetti, M.; Romero, M.; Costa, F.; and Preatoni, E., 2009. Method for movement and gesture assessment (mmga) in ergonomics. In *International Conference on Digital Human Modeling*, 591–598. Springer.

Aristidou, A.; Cameron, J.; and Lasenby, J., 2008. Predicting missing markers to drive real-time centre of rotation estimation. *Articulated Motion and Deformable Objects*, (2008), 238–247.

Aristidou, A. and Chrysanthou, Y., 2014. Feature extraction for human motion indexing of acted dance performances. In *Computer Graphics Theory and Applications (GRAPP), 2014 International Conference on*, 1–11. IEEE.

Aristidou, A.; Stavrakis, E.; Charalambous, P.; Chrysanthou, Y.; and Himona, S. L., 2015. Folk dance evaluation using laban movement analysis. *Journal on Computing and Cultural Heritage (JOCCH)*, 8, 4 (2015), 20.

Baak, A., 2013. Retrieval-based approaches for tracking and reconstructing human motions.

Bachynskyi, M.; Palmas, G.; Oulasvirta, A.; Steimle, J.; and Weinkauf, T., 2015. Performance and ergonomics of touch surfaces: A comparative study using biomechanical simulation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1817–1826. ACM.

Barliya, A.; Omlor, L.; Giese, M. A.; Berthoz, A.; and Flash, T., 2013. Expression of emotion in the kinematics of locomotion. *Experimental brain research*, 225, 2 (2013), 159–176.

Bartenieff, I. and Lewis, D., 1980. *Body movement: Coping with the environment*. Routledge.

Bates, D. M. and Watts, D. a. G., 1988. *Nonlinear regression analysis and its applications*, vol. 2. Wiley Online Library.

Baum, L. E. and Petrie, T., 1966. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37, 6 (1966), 1554–1563.

Bianco, S. and Tisato, F., 2013. Karate moves recognition from skeletal motion. In *Three-Dimensional Image Processing (3DIP) and Applications 2013*, vol. 8650, 86500K. International Society for Optics and Photonics.

Blanz, V. and Vetter, T., 1999. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 187–194. ACM Press/Addison-Wesley Publishing Co.

Boone, D. C. and Azen, S. P., 1979. Normal range of motion of joints in male subjects. *The Journal of Bone & Joint Surgery*, 61, 5 (1979), 756–759.

Boser, B. E.; Guyon, I. M.; and Vapnik, V. N., 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, 144–152. ACM.

Bouchard, D. and Badler, N., 2007. Semantic segmentation of motion capture using laban movement analysis. *Intelligent Virtual Agents*, (2007), 37–44.

Bradski, G. R. and Davis, J. W., 2002. Motion segmentation and pose recognition with motion history gradients. *Machine Vision and Applications*, 13, 3 (2002), 174–184.

Brooks, A. L. and Czarowicz, A., 2012. Markerless motion tracking: Ms kinect & organic motion openstage®. In *International Conference Disability, Virtual Reality & Associated Technologies*, 435–437.

Bruderlin, A. and Williams, L., 1995. Motion signal processing. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, 97–104. ACM.

Bruijn, S.; Meijer, O.; Beek, P.; and Van Dieën, J., 2013. Assessing the stability of human locomotion: a review of current measures. *Journal of the Royal Society Interface*, 10, 83 (2013), 20120999.

Burger, B. and Toiviainen, P., 2013a. MoCap Toolbox – A Matlab toolbox for computational analysis of movement data. In *Proceedings of the 10th Sound and Music Computing Conference*, 172–178. KTH Royal Institute of Technology, Stockholm, Sweden.

Burger, B. and Toiviainen, P., 2013b. Mocap toolbox-a matlab toolbox for computational analysis of movement data. In *10th Sound and Music Computing Conference, SMC 2013, Stockholm, Sweden*. Logos Verlag Berlin.

Cadopi, M., 2005. La motricité du danseur: approche cognitive. *Bulletin de psychologie*, 1 (2005), 29–37.

Camomilla, V.; Bergamini, E.; Fantozzi, S.; and Vannozzi, G., 2018. Trends supporting the in-field use of wearable inertial sensors for sport performance evaluation: A systematic review. *Sensors*, 18, 3 (2018), 873.

Caruana, R., 1997. Multitask learning. *Machine learning*, 28, 1 (1997), 41–75.

Caulier, E., 2010. *Comprendre le taijiquan*, vol. 1. Editions Modulaires Européennes InterCommunication SPRL.

Caulier, E., 2014. Le taijiquan : une voie d'incorporation et de comprehension des nouveaux paradigmes. *Plastir*, 37 (12 2014), 86 – 107.

Caulier, E., 2015. Du geste formel a la gestuelle habitee : la voie du taijiquan. *Recherches & Educations*, 13, 2 (06 2015), 59 – 71.

Chai, J. and Hodgins, J. K., 2005. Performance animation from low-dimensional control signals. In *ACM Transactions on Graphics (TOG)*, vol. 24, 686–696. ACM.

Chen, L.; Gibet, S.; Marteau, P.-F.; Marandola, F.; and Wanderley, M. M., 2016. Quantitative evaluation of percussive gestures by ranking trainees versus teacher. In *Proceedings of the 3rd International Symposium on Movement and Computing*, 13. ACM.

Chi, D.; Costa, M.; Zhao, L.; and Badler, N., 2000. The emote model for effort and shape. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 173–182. ACM Press/Addison-Wesley Publishing Co.

Chiari, L.; Della Croce, U.; Leardini, A.; and Cappozzo, A., 2005. Human movement analysis using stereophotogrammetry: Part 2: Instrumental errors. *Gait & posture*, 21, 2 (2005), 197–211.

CMU, 2003. Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu/.

Cohen, I.; Sebe, N.; Garg, A.; Chen, L. S.; and Huang, T. S., 2003. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and image understanding*, 91, 1-2 (2003), 160–187.

Cools, W.; De Martelaer, K.; Samaey, C.; and Andries, C., 2009. Movement skill assessment of typically developing preschool children: A review of seven movement skill assessment tools. *Journal of sports science & medicine*, 8, 2 (2009), 154.

Dadashi, F.; Millet, G.; and Aminian, K., 2015. Front-crawl stroke descriptors variability assessment for skill characterisation. *Journal of sports sciences*, 34, 15 (2015), 1405–1412.

Dael, N.; Mortillaro, M.; and Scherer, K. R., 2012. Emotion expression in body action and posture. *Emotion*, 12, 5 (2012), 1085.

Daffertshofer, A.; Lamoth, C. J.; Meijer, O. G.; and Beek, P. J., 2004. Pca in studying coordination and variability: a tutorial. *Clinical biomechanics*, 19, 4 (2004), 415–428.

De Aguiar, E.; Theobalt, C.; and Seidel, H.-P., 2006. Automatic learning of articulated skeletons from 3d marker trajectories. In *International Symposium on Visual Computing*, 485–494. Springer.

De Leva, P., 1996. Adjustments to zatsiorsky-seluyanov's segment inertia parameters. *Journal of biomechanics*, 29, 9 (1996), 1223–1230.

Deitz, J. C.; Kartin, D.; and Kopp, K., 2007. Review of the bruininks-oseretsky test of motor proficiency, (bot-2). *Physical & occupational therapy in pediatrics*, 27, 4 (2007), 87–102.

Dejnabadi, H.; Jolles, B. M.; and Aminian, K., 2008. A new approach for quantitative analysis of inter-joint coordination during gait. *IEEE Transactions on Biomedical Engineering*, 55, 2 (2008), 755–764.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 248–255. IEEE.

Devineau, G.; Xi, W.; Moutarde, F.; and Yang, J., 2018. Deep learning for hand gesture recognition on skeletal data. In *13th IEEE Conference on Automatic Face and Gesture Recognition (FG'2018)*.

Dietterich, T. G., 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, 1–15. Springer.

Dietterich, T. G., 2002. Machine learning for sequential data: A review. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, 15–30. Springer.

Ding, Z.; Wang, P.; Ogunbona, P. O.; and Li, W., 2017. Investigation of different skeleton features for cnn-based 3d action recognition. In *Multimedia & Expo Workshops (ICMEW), 2017 IEEE International Conference on*, 617–622. IEEE.

Dipietro, L.; Sabatini, A. M.; and Dario, P., 2003. Evaluation of an instrumented glove for hand-movement acquisition. *Journal of rehabilitation research and development*, 40, 2 (2003), 179–190.

Drucker, H.; Burges, C. J.; Kaufman, L.; Smola, A. J.; and Vapnik, V., 1997. Support vector regression machines. In *Advances in neural information processing systems*, 155–161.

Duclos, C.; Desjardins, P.; Nadeau, S.; Delisle, A.; Gravel, D.; Brouwer, B.; and Corriveau, H., 2009. Destabilizing and stabilizing forces to assess equilibrium during everyday activities. *Journal of biomechanics*, 42, 3 (2009), 379–382.

Dutoit, T., 1997. *An introduction to text-to-speech synthesis*, vol. 3. Springer Science & Business Media.

Ericsson, K. A. and Lehmann, A. C., 1996. Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual review of psychology*, 47, 1 (1996), 273–305.

Federolf, P.; Reid, R.; Gilgien, M.; Haugen, P.; and Smith, G., 2012. The application of principal component analysis to quantify technique in sports. *Scandinavian Journal of Medicine and Science in Sports*, 24, 3 (2012), 491–499. doi: 10.1111/j.1600-0838.2012.01455.x.

Federolf, P. A., 2013. A novel approach to solve the "missing marker problem" in marker-based motion analysis that exploits the segment coordination patterns in multi-limb motion data. *PloS one*, 8, 10 (2013), e78689.

Feng, Y.; Xiao, J.; Zhuang, Y.; Yang, X.; Zhang, J. J.; and Song, R., 2014. Exploiting temporal stability and low-rank structure for motion capture data refinement. *Information Sciences*, 277 (2014), 777–793.

Firat, M. and Gungor, M., 2009. Generalized regression neural networks and feed forward neural networks for prediction of scour depth around bridge piers. *Advances in Engineering Software*, 40, 8 (2009), 731–737.

Fischer, R., 2018. The history and current state of motion capture. http://www.motioncapturesociety.com/resources/industry-history. Retrieved: 2018/31/07.

FLEURANCE, P., 2009. Sport de haute performance et cognition. introduction. "je vois la balle avec les mains". *Intellectica*, 52 (2009), 7–27.

Friedman, J. H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, (2001), 1189–1232.

Fritsch, F. N. and Carlson, R. E., 1980. Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17, 2 (1980), 238–246.

Furuya, S.; Goda, T.; Katayose, H.; Miwa, H.; and Nagata, N., 2011. Distinct inter-joint coordination during fast alternate keystrokes in pianists with superior skill. *Frontiers in human neuroscience*, 5 (2011), 50.

Furuya, S.; Tominaga, K.; Miyazaki, F.; and Altenmüller, E., 2015. Losing dexterity: patterns of impaired coordination of finger movements in musician's dystonia. *Scientific reports*, 5 (2015), 13360.

Gallahue, D. L.; Ozmun, J. C.; and Goodway, J., 2006. *Understanding motor development: Infants, children, adolescents, adults*. Boston.

Gløersen, Ø. and Federolf, P., 2016. Predicting missing marker trajectories in human motion data using marker intercorrelations. *PloS one*, 11, 3 (2016), e0152616.

Gløersen, Ø.; Myklebust, H.; Hallén, J.; and Federolf, P., 2017. Technique analysis in elite athletes using principal component analysis. *Journal of sports sciences*, 36, 2 (2017), 229–237.

Glorot, X.; Bordes, A.; and Bengio, Y., 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315–323.

Gordon, A.; Huxley, A. F.; and Julian, F., 1966. The variation in isometric tension with sarcomere length in vertebrate muscle fibres. *The Journal of physiology*, 184, 1 (1966), 170–192.

Güler, R. A.; Neverova, N.; and Kokkinos, I., 2018. Densepose: Dense human pose estimation in the wild. *arXiv preprint arXiv:1802.00434*, (2018).

Haering, D.; Raison, M.; and Begon, M., 2014. Measurement and description of three-dimensional shoulder range of motion with degrees of freedom interactions. *Journal of biomechanical engineering*, 136, 8 (2014).

Hamilton, W. R., 1866. *Elements of quaternions*. Longmans, Green, & Company.

Han, X.; Zhong, Y.; Cao, L.; and Zhang, L., 2017. Pre-trained alexnet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification. *Remote Sensing*, 9, 8 (2017), 848.

Hannaford, B. and Lee, P., 1991. Hidden markov model analysis of force/torque information in telemanipulation. *The International journal of robotics research*, 10, 5 (1991), 528–539.

Harding, W. J. and James, A. D., 2010. Analysis of snowboarding performance at the burton open australian half-pipe championships. *International Journal of Performance Analysis in Sport*, 10, 1 (2010), 66–81.

Harrison, A.; Ryan, W.; and Hayes, K., 2007. Functional data analysis of joint coordination in the development of vertical jump performance. *Sports Biomechanics*, 6, 2 (2007), 199–214.

HAUW, D., 2009. Activité et performances acrobatiques de haut niveau. *Intellectica*, 52 (2009), 55–69.

Helm, F.; Munzert, J.; and Troje, N. F., 2017. Kinematic patterns underlying disguised movements: spatial and temporal dissimilarity compared to genuine movement patterns. *Human movement science*, 54 (2017), 308–319.

Herda, L.; Fua, P.; Plankers, R.; Boulic, R.; and Thalmann, D., 2000. Skeleton-based motion capture for robust reconstruction of human motion. In *Computer Animation 2000. Proceedings*, 77–83. IEEE.

Hodgins, J. Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu/. Accessed: 2017-09-20.

Hoeting, J. A.; Madigan, D.; Raftery, A. E.; and Volinsky, C. T., 1999. Bayesian model averaging: a tutorial. *Statistical science*, (1999), 382–401.

Hof, A.; Gazendam, M.; and Sinke, W., 2005. The condition for dynamic stability. *Journal of biomechanics*, 38, 1 (2005), 1–8.

Holden, D.; Saito, J.; and Komura, T., 2016. A deep learning framework for character motion synthesis and editing. *ACM Trans. Graph.*, 35, 4 (Jul. 2016), 138:1–138:11.

Holden, D.; Saito, J.; Komura, T.; and Joyce, T., 2015. Learning motion manifolds with convolutional autoencoders. *SIGGRAPH Asia 2015 Technical Briefs*, (2015), 18:1—-18:4.

Howarth, S. J. and Callaghan, J. P., 2010. Quantitative assessment of the accuracy for three interpolation techniques in kinematic analysis of human movement. *Computer methods in biomechanics and biomedical engineering*, 13, 6 (2010), 847–855.

HTC, 2018. Vive tracker for developers. https://developer.vive.com/fr/vive-tracker-for-developer/.

Iandola, F. N.; Han, S.; Moskewicz, M. W.; Ashraf, K.; Dally, W. J.; and Keutzer, K., 2016. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, (2016).

Jensenius, A. R. and Wanderley, M. M., 2010. Musical gestures: Concepts and methods in research. In *Musical Gestures*, 24–47. Routledge.

Kabsch, W., 1976. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32, 5 (1976), 922–923.

Kadous, W. et al., 1995. Grasp: Recognition of australian sign language using instrumented gloves. (1995).

Kakadiaris, L. and Metaxas, D., 2000. Model-based estimation of 3d human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 12 (2000), 1453–1459.

Kalman, R. E. et al., 1960. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82, 1 (1960), 35–45.

Kami, A.; Meyer, G.; Jezzard, P.; Adams, M. M.; Turner, R.; and Ungerleider, L. G., 1995. Functional mri evidence for adult motor cortex plasticity during motor skill learning. *Nature*, 377, 6545 (1995), 155.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J., 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7122–7131.

Kapadia, M.; Chiang, I.-k.; Thomas, T.; Badler, N. I.; Kider Jr, J. T.; et al., 2013. Efficient motion retrieval in large motion databases. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, 19–28. ACM.

Kay, B. A.; Turvey, M. T.; and Meijer, O. G., 2003. An early oscillator model: studies on the biodynamics of the piano strike (bernstein & popova, 1930). (2003).

Kee, D. and Karwowski, W., 2001. Luba: an assessment technique for postural loading on the upper body based on joint motion discomfort and maximum holding time. *Applied Ergonomics*, 32, 4 (2001), 357–366.

Kee, D. and Karwowski, W., 2003. Ranking systems for evaluation of joint and joint motion stressfulness based on perceived discomforts. *Applied ergonomics*, 34, 2 (2003), 167–176.

Kendall, M. G., 1938. A new measure of rank correlation. *Biometrika*, 30, 1/2 (1938), 81–93.

Kennedy, J. and Eberhart, R., 1995. Particle swarm optimization. In *Neural Networks, 1995. Proceedings., IEEE International Conference on*, vol. 4, 1942–1948 vol.4. doi:10.1109/ICNN.1995.488968.

Kim, Y. K.; Kim, Y. H.; and Im, S. J., 2011. Inter-joint coordination in producing kicking velocity of taekwondo kicks. *Journal of sports science & medicine*, 10, 1 (2011), 31.

King, M. A. and Yeadon, M. R., 2003. Coping with perturbations to a layout somersault in tumbling. *Journal of Biomechanics*, 36, 7 (2003), 921–927.

Kirk, A. G.; O'Brien, J. F.; and Forsyth, D. A., 2005. Skeletal parameter estimation from optical motion capture data. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2, 782–788. IEEE.

Kotsiantis, S. B.; Zaharakis, I.; and Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160 (2007), 3–24.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, 1097–1105. Curran Associates, Inc.

Kulpa, R.; Multon, F.; and Arnaldi, B., 2005. Morphology-independent representation of motions for interactive human-like animation. In *Computer Graphics Forum*, vol. 24, 343–351. Wiley Online Library.

Lai, R. Y.; Yuen, P. C.; and Lee, K. K., 2011. Motion capture data completion and denoising by singular value thresholding. (2011).

Lalys, F. and Jannin, P., 2014. Surgical process modelling: a review. *International journal of computer assisted radiology and surgery*, 9, 3 (2014), 495–511.

Laraba, S.; Brahimi, M.; Tilmanne, J.; and Dutoit, T., 2017. 3d skeleton-based action recognition by representing motion capture sequences as 2d-rgb images. *Computer Animation and Virtual Worlds*, 28, 3-4 (2017), 1–11. doi:10.1002/cav.1782.

Laraba, S. and Tilmanne, J., 2016. Dance performance evaluation using hidden markov models. *Computer Animation and Virtual Worlds*, 27, 3-4 (2016), 321–329.

Laraba, S.; Tilmanne, J.; and Dutoit, T., 2015. Adaptation procedure for hmm-based sensor-dependent gesture recognition. In *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*, 17–22. ACM.

Larboulette, C. and Gibet, S., 2015. A review of computable expressive descriptors of human motion. In *Proceedings of the 2nd International Workshop on Movement and Computing*, 21–28. ACM.

LaViola, J., 1999. A survey of hand posture and gesture recognition techniques and technology. *Brown University, Providence, RI*, 29 (1999).

LeCun, Y.; Bengio, Y.; and Hinton, G., 2015. Deep learning. *nature*, 521, 7553 (2015), 436.

Leroy, D.; Thouvarecq, R.; and Gautier, G., 2008. Postural organisation during cascade juggling: Influence of expertise. *Gait & posture*, 28, 2 (2008), 265–270.

Li, L.; McCann, J.; Pollard, N.; and Faloutsos, C., 2010. Bolero: a principled technique for including bone length constraints in motion capture occlusion filling. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 179–188. Eurographics Association.

Li, L.; McCann, J.; Pollard, N. S.; and Faloutsos, C., 2009. Dynammo: Mining and summarization of coevolving sequences with missing values. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 507–516. ACM.

Liu, G. and McMillan, L., 2006. Estimation of missing markers in human motion capture. *The Visual Computer*, 22, 9 (2006), 721–728.

Lord, S.; Galna, B.; Verghese, J.; Coleman, S.; Burn, D.; and Rochester, L., 2012. Independent domains of gait in older adults and associated motor and nonmotor attributes: Validation of a factor analysis approach. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, 68, 7 (2012), 820–827. doi:10.1093/gerona/gls255.

Ma, Y.; Paterson, H. M.; and Pollick, F. E., 2006. A motion capture library for the study of identity, gender, and emotion perception from biological motion. *Behavior research methods*, 38, 1 (2006), 134–141.

Mandery, C.; Terlemez, O.; Do, M.; Vahrenkamp, N.; and Asfour, T., 2015. The kit whole-body human motion database. In *International Conference on Advanced Robotics (ICAR)*, 329–336.

Mandery, C.; Terlemez, Ö.; Do, M.; Vahrenkamp, N.; and Asfour, T., 2016. Unifying representations and large-scale whole-body motion databases for studying human motion. *IEEE Transactions on Robotics*, 32, 4 (2016), 796–809.

Marey, E.-J., 1873. *La machine animale*, vol. 3. Germer Baillière.

Marques, J. M.; Yzerbyt, V. Y.; and Leyens, J.-P., 1988. The "black sheep effect": Extremity of judgments towards ingroup members as a function of group identification. *European Journal of Social Psychology*, 18, 1 (1988), 1–16.

McAtamney, L. and Corlett, E. N., 1993. Rula: a survey method for the investigation of work-related upper limb disorders. *Applied ergonomics*, 24, 2 (1993), 91–99.

Megali, G.; Sinigaglia, S.; Tonet, O.; and Dario, P., 2006. Modelling and evaluation of surgical performance using hidden markov models. *IEEE Transactions on Biomedical Engineering*, 53, 10 (2006), 1911–1919.

Metcalf, C. D.; Irvine, T. A.; Sims, J. L.; Wang, Y. L.; Su, A. W.; and Norris, D. O., 2014. Complex hand dexterity: a review of biomechanical methods for measuring musical performance. *Frontiers in psychology*, 5 (2014).

Morel, M., 2017. *Modélisation de séries temporelles multidimensionnelles. Application à l'évaluation générique et automatique du geste sportif.* Ph.D. thesis, Université Pierre & Marie Curie.

Morel, M.; Achard, C.; Kulpa, R.; and Dubuisson, S., 2017. Automatic evaluation of sports motion: A generic computation of spatial and temporal errors. *Image and Vision Computing*, 64 (2017), 67–78.

Morel, M.; Kulpa, R.; Sorel, A.; Achard, C.; and Dubuisson, S., 2016. Automatic and generic evaluation of spatial and temporal errors in sport motions. In *11th International Conference on Computer Vision Theory and Applications (VISAPP 2016)*, 542–551.

Motion-Analysis, 1982. Motion analysis corporation. https://www.motionanalysis.com/industry-firsts/.

Müller, M.; Baak, A.; and Seidel, H.-P., 2009. Efficient and robust annotation of motion capture data. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 17–26. ACM.

Müller, M. and Röder, T., 2006. Motion templates for automatic classification and retrieval of motion capture data. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, 137–146. Eurographics Association.

Müller, M. and Röder, T., 2008. A Relational Approach to Content-based Analysis of Motion Capture Data. *Human Motion*, (2008).

Müller, M.; Röder, T.; and Clausen, M., 2005. Efficient content-based retrieval of motion capture data. In *ACM Transactions on Graphics (ToG)*, vol. 24, 677–685. ACM.

Müller, M.; Röder, T.; Clausen, M.; Eberhardt, B.; Krüger, B.; and Weber, A., 2007. Documentation mocap database hdm05. Technical Report CG-2007-2, Universität Bonn.

Multon, F., 2013. Sensing human walking: Algorithms and techniques for extracting and modeling locomotion. In *Human Walking in Virtual Environments*, 177–197. Springer.

Multon, F. and Olivier, A.-H., 2013. Biomechanics of walking in real world: naturalness we wish to reach in virtual reality. In *Human Walking in Virtual Environments*, 55–77. Springer.

Mundermann, L.; Corazza, S.; Chaudhari, A. M.; Alexander, E. J.; and Andriacchi, T. P., 2005. Most favorable camera configuration for a shape-from-silhouette markerless motion capture system for biomechanical analysis. In *Electronic Imaging 2005*, 278–287. International Society for Optics and Photonics.

NASA, 1995. 3000. man systems integration standards. volume 1. section 3. anthropometry and biomechanics. *National Aeronautics andSpace Administration, Houston, USA*, (1995). https://msis.jsc.nasa.gov/sections/section03.htm.

Neverova, N., 2016. *Deep learning for human motion analysis*. Ph.D. thesis, Université de Lyon.

Newlove, J., 1993. *Laban for actors and dancers: putting laban's movement theory into practice: a step-by-step guide*.

Ofli, F.; Chaudhry, R.; Kurillo, G.; Vidal, R.; and Bajcsy, R., 2013. Berkeley mhad: A comprehensive multimodal human action database. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, 53–60. IEEE.

Olesh, E. V.; Yakovenko, S.; and Gritsenko, V., 2014. Automated assessment of upper extremity movement impairment due to stroke. *PLoS ONE*, 9, 8 (2014).

Parzen, E., 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33, 3 (1962), 1065–1076.

Patrona, F.; Chatzitofis, A.; Zarpalas, D.; and Daras, P., 2018. Motion analysis: Action detection, recognition and evaluation based on motion capture data. *Pattern Recognition*, 76 (2018), 612–622.

Pearson, K., 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2, 11 (1901), 559–572.

Peng, S.-J.; He, G.-F.; Liu, X.; and Wang, H.-Z., 2015. Hierarchical block-based incomplete human mocap data recovery using adaptive nonnegative matrix factorization. *Computers & Graphics*, 49 (2015), 10–23.

Peng, X. B.; Abbeel, P.; Levine, S.; and van de Panne, M., 2018. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *arXiv preprint arXiv:1804.02717*, (2018).

Pham, T.; Pathirana, P. N.; Won, Y.; and Li, S., 2016. A summative scoring system for evaluation of human kinematic performance. *Biomedical Signal Processing and Control*, 23 (2016), 85–92.

Pirsiavash, H.; Vondrick, C.; and Torralba, A., 2014. Assessing the quality of actions. In *European Conference on Computer Vision*, 556–571. Springer.

Plamondon, R. and Srihari, S. N., 2000. Online and off-line handwriting recognition: a comprehensive survey. *IEEE Transactions on pattern analysis and machine intelligence*, 22, 1 (2000), 63–84.

Polhemus, 1969. Polhemus. https://polhemus.com/company/history/.

Poppe, R., 2010. A survey on vision-based human action recognition. *Image and vision computing*, 28, 6 (2010), 976–990.

Raafat, R. M.; Chater, N.; and Frith, C., 2009. Herding in humans. *Trends in cognitive sciences*, 13, 10 (2009), 420–428.

Rabiner, L. R. and Juang, B.-H., 1993. *Fundamentals of speech recognition*, vol. 14. PTR Prentice Hall Englewood Cliffs.

Reiley, C. E.; Lin, H. C.; Yuh, D. D.; and Hager, G. D., 2011. Review of methods for objective surgical skill evaluation. *Surgical endoscopy*, 25, 2 (2011), 356–366.

Romero, V.; Amaral, J.; Fitzpatrick, P.; Schmidt, R.; Duncan, A. W.; and Richardson, M. J., 2017. Can low-cost motion-tracking systems substitute a polhemus system when researching social motor coordination in children? *Behavior research methods*, 49, 2 (2017), 588–601.

Rosen, J.; Hannaford, B.; Richards, C. G.; and Sinanan, M. N., 2001. Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/-torque signatures for evaluating surgical skills. *IEEE transactions on Biomedical Engineering*, 48, 5 (2001), 579–591.

Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J., 1985. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Samadani, A.-A.; Ghodsi, A.; and Kulić, D., 2013. Discriminative functional analysis of human movements. *Pattern Recognition Letters*, 34, 15 (2013), 1829–1839.

Schmitt, O. H., 1938. A thermionic trigger. *Journal of Scientific Instruments*, 15, 1 (1938), 24.

Sebastiani, F., 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34, 1 (2002), 1–47.

Seber, G. A. and Lee, A. J., 2003. *Linear regression analysis*. John Wiley & Sons.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G., 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. *arXiv preprint arXiv:1604.02808*, (2016).

Sie, M.-S.; Cheng, Y.-C.; and Chiang, C.-C., 2014. Key motion spotting in continuous motion sequences using motion sensing devices. In *Signal Processing, Communications and Computing (ICSPCC), 2014 IEEE International Conference on*, 326–331. IEEE.

Silverman, B. W., 1986. *Density estimation for statistics and data analysis*, vol. 26. CRC press.

Smola, A. J. and Schölkopf, B., 2004. A tutorial on support vector regression. *Statistics and computing*, 14, 3 (2004), 199–222.

Specht, D. F., 1991. A general regression neural network. *IEEE transactions on neural networks*, 2, 6 (1991), 568–576.

Starner, T.; Weaver, J.; and Pentland, A., 1998. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on pattern analysis and machine intelligence*, 20, 12 (1998), 1371–1375.

Sturman, D. J., 1994. A brief history of motion capture for computer character animation. *SIGGRAPH94, Course9*, (1994).

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A.; et al., 2015. Going deeper with convolutions. Cvpr.

Tachibana, H.; Uenoyama, K.; and Aihara, S., 2017. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. *arXiv preprint arXiv:1710.08969*, (2017).

Tahir, N. M. and Manap, H. H., 2012. Parkinson disease gait classification based on machine learning approach. *Journal of Applied Sciences*, 12, 2 (2012), 180–185.

Tan, C.-H.; Hou, J.; and Chau, L.-P., 2013. Human motion capture data recovery using trajectory-based matrix completion. *Electronics Letters*, 49, 12 (2013), 752–754.

Tan, C.-H.; Hou, J.; and Chau, L.-P., 2015. Motion capture data recovery using skeleton constrained singular value thresholding. *The Visual Computer*, 31, 11 (2015), 1521–1532.

Taylor, K. D.; Mottier, F. M.; Simmons, D. W.; Cohen, W.; Pavlak, R.; Cornell, D. P.; and Hankins, G. B., 1982. An automated motion measurement system for clinical gait analysis. *Journal of Biomechanics*, 15, 7 (1982), 505–516.

Tilmanne, J., 2013. *Data-driven Stylistic Humanlike Walk Synthesis*. Ph.D. thesis, PhD Dissertation, University of Mons, 2013.(Cited on pages 28 and 55.).

Tilmanne, J. and d'Alessandro, N., 2015. Motion machine: A new framework for motion capture signal feature prototyping. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, 2401–2405. IEEE.

Tilmanne, J.; d'Alessandro, N.; Barborka, P.; Bayansar, F.; Bernardo, F.; Fiebrink, R.; Heloir, A.; Hemery, E.; Laraba, S.; Moinet, A.; Nunnari, F.; Ravet, T.; Sarasua, A.; Tits, M.; Tits, N.; Zajega, F.; and Reboursiere, L., 2015. Prototyping a new audio-visual instrument based on extraction of high-level features on full-body motion. In *Proceedings of eNTERFACE 2015 Workshop on Intelligent Interfaces*.

Tilmanne, J. and Dutoit, T., 2010. Expressive gait synthesis using pca and gaussian modeling. In *International Conference on Motion in Games*, 363–374. Springer.

Tits, M.; Laraba, S.; Caulier, E.; Tilmanne, J.; and Dutoit, T., 2018a. Umons-taichi: A multimodal motion capture dataset of expertise in taijiquan gestures. *Data in Brief*, (2018).

Tits, M.; Tilmanne, J.; and d'Alessandro, N., 2016. A novel tool for motion capture database factor statistical exploration. In *Proceedings of the 3rd International Symposium on Movement and Computing*, 2. ACM.

Tits, M.; Tilmanne, J.; D'Alessandro, N.; and Wanderley, M. M., 2015. Feature extraction and expertise analysis of pianists' motion-captured finger gestures. *International Computer Music Conference*, (2015), 102–105.

Tits, M.; Tilmanne, J.; and Dutoit, T., 2017. Morphology independent feature engineering in motion capture database for gesture evaluation. In *Proceedings of the 4th International Conference on Movement Computing*, 26. ACM.

Tits, M.; Tilmanne, J.; and Dutoit, T., 2018b. Robust and automatic motion-capture data recovery using soft skeleton constraints and model averaging. *PLOS ONE*, 13, 7 (07 2018), 1–21.

Torresani, L.; Hackney, P.; and Bregler, C., 2007. Learning motion style synthesis from perceptual observations. In *Advances in Neural Information Processing Systems*, 1393–1400.

Troje, N. F., 2002. Decomposing biological motion: a framework for analysis and synthesis of human gait patterns. *Journal of vision*, 2, 5 (2002), 371–387.

Troje, N. F.; Westhoff, C.; and Lavrov, M., 2005. Person identification from biological motion: Effects of structural and kinematic cues. *Perception & Psychophysics*, 67, 4 (2005), 667–675.

Tzu, C., 1964. Basic writings (b. watson, trans.). *New York & London: Columbia University Press. Davies-Gibson, MR (1994, November). Storytelling in the multicultural classroom: A study in community*, (1964).

Unuma, M.; Anjyo, K.; and Takeuchi, R., 1995. Fourier principles for emotion-based human figure animation. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, 91–96. ACM.

Van Someren, E. J.; Vonk, B. F.; Thijssen, W. A.; Speelman, J. D.; Schuurman, P. R.; Mirmiran, M.; and Swaab, D. F., 1998. A new actigraph for long-term registration of the duration and intensity of tremor and movement. *IEEE Transactions on Biomedical Engineering*, 45, 3 (1998), 386–395.

Verlinden, V.; van der Geest, J.; Hoogendam, Y.; Hofman, A.; Breteler, M.; and Ikram, M., 2013. Gait patterns in a community-dwelling population aged 50 years and older. *Gait & Posture*, 37, 4 (2013), 500–505.

VICON, 1984. Motion capture systems | vicon. https://www.vicon.com/. https://www.vicon.com/vicon/about. Accessed: 2017-09-20.

Vintsyuk, T. K., 1968. Speech discrimination by dynamic programming. *Cybernetics*, 4, 1 (1968), 52–57.

Wang, J. M.; Fleet, D. J.; and Hertzmann, A., 2008. Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence*, 30, 2 (2008), 283–298.

Wei, D.; Zhou, B.; Torralba, A.; and Freeman, W. T., 2017. mneuron: A matlab plugin to visualize neurons from deep models. http://vision03.csail.mit.edu/cnn_art/index.html.

Welch, P., 1967. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15, 2 (1967), 70–73.

Whiting, H. T. A., 1983. *Human motor actions: Bernstein reassessed*. Elsevier.

Xia, S.; Gao, L.; Lai, Y.-K.; Yuan, M.-Z.; and Chai, J., 2017. A survey on human performance capture and animation. *Journal of Computer Science and Technology*, 32, 3 (May 2017), 536–554.

You, Y.; Zhang, Z.; Hsieh, C.; Demmel, J.; and Keutzer, K., 2017. Imagenet training in minutes. *CoRR, abs/1709.05011*, (2017).

Young, C. and Reinkensmeyer, D. J., 2014. Judging complex movement performances for excellence: A principal components analysis-based technique applied to competitive diving. *Human Movement Science*, 36 (2014), 107–122.

Zago, M.; Codari, M.; Iaia, F. M.; and Sforza, C., 2016. Multi-segmental movements as a function of experience in karate. *Journal of Sports Sciences*, (2016), 1–8.

Zago, M.; Pacifici, I.; Lovecchio, N.; Galli, M.; Federolf, P. A.; and Sforza, C., 2017. Multi-segmental movement patterns reflect juggling complexity and skill level. *Human Movement Science*, 54, February (2017), 144–153.

Zatsiorsky, V., 1990. Methods of determing mass-inertial characteristics of human body segments. *Contemporasy Problems of Biomechnics*, (1990).

Zhang, J.; Li, W.; Ogunbona, P. O.; Wang, P.; and Tang, C., 2016. Rgb-d-based action recognition datasets: A survey. *Pattern Recognition*, 60 (2016), 86–105.

Zhang, Y.; Lin, W. C.; and Chin, Y.-K. S., 2010. A pattern-recognition approach for driving skill characterization. *IEEE transactions on intelligent transportation systems*, 11, 4 (2010), 905–916.

Zhao, L. and Badler, N. I., 2005. Acquiring and validating motion qualities from live limb gestures. *Graphical Models*, 67, 1 (2005), 1–16.

Zhao, W.; Chellappa, R.; Phillips, P. J.; and Rosenfeld, A., 2003. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35, 4 (2003), 399–458.

Zou, H. and Hastie, T., 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 2 (2005), 301–320.

Appendix A

'Kick with the heel' feedback images

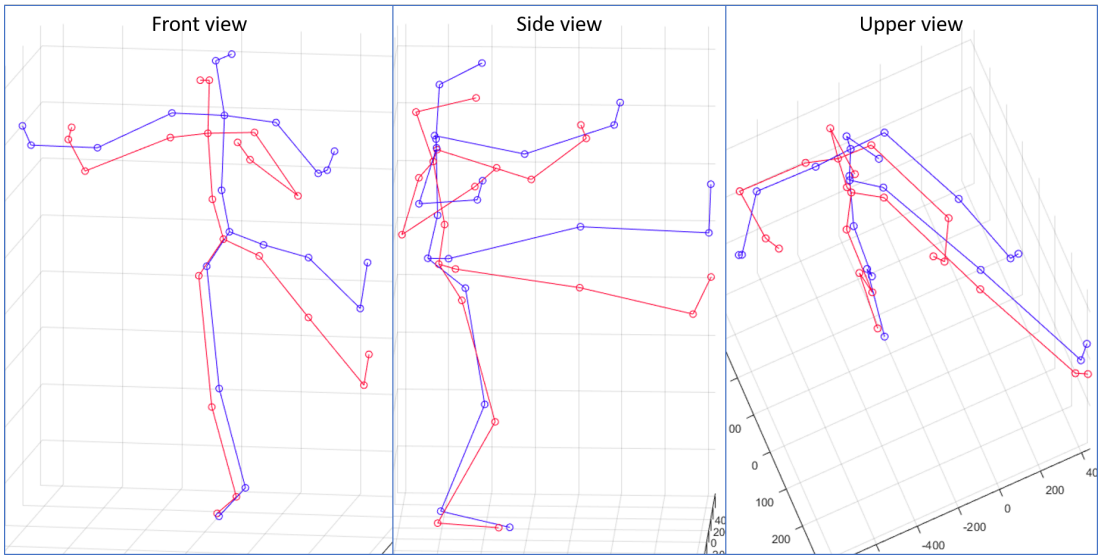

Figure A.1: Visual feedback for a rendition of G11 (Kick with the heel) by P11 (the lowest-skilled participant). Red: original sequence. Blue: feedback sequence with an improved score of 10.

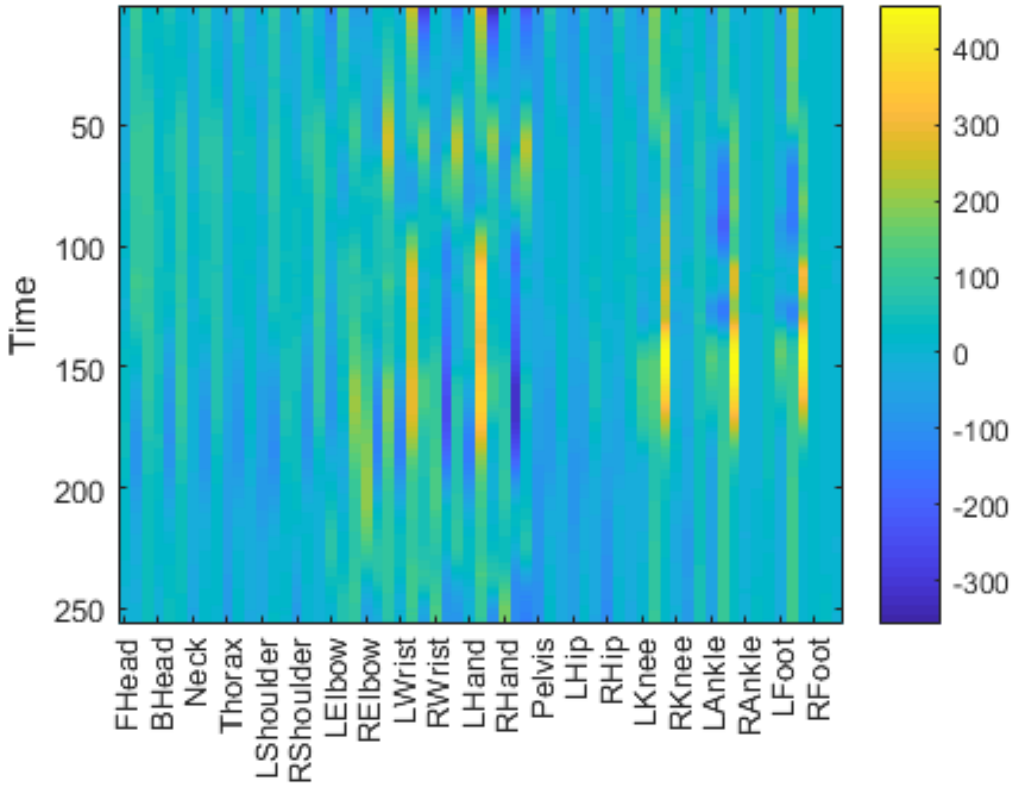

Figure A.2: Differences between global positions, for a rendition of G11 (Kick with the heel) by P11 (lowest-skilled participant). (scale in mm)

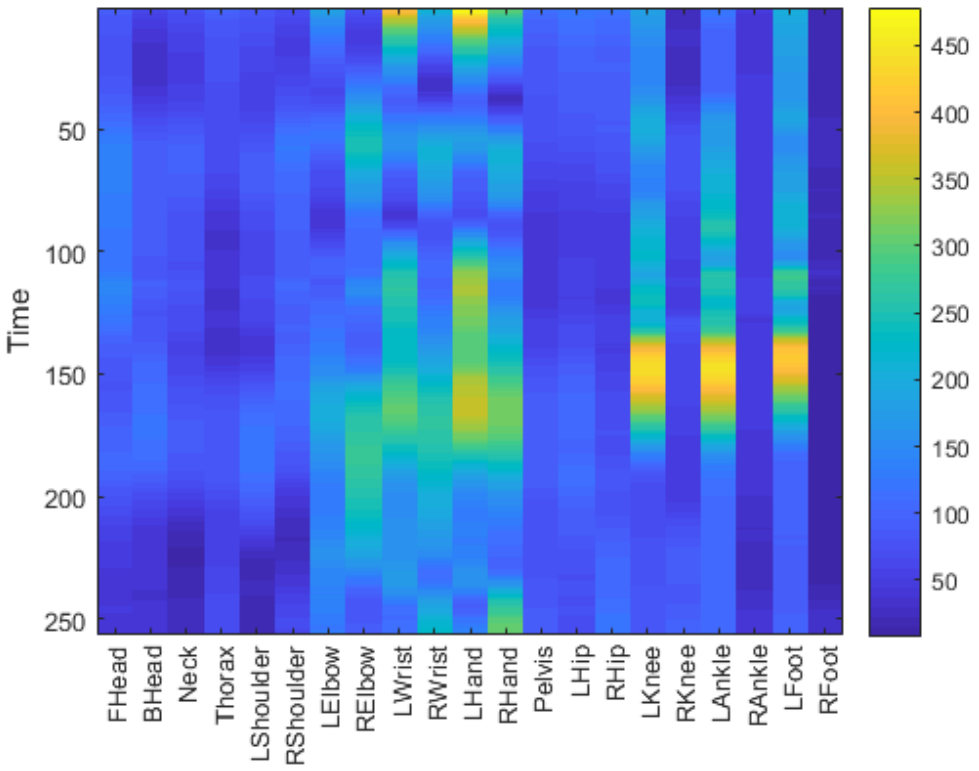

Figure A.3: Distances between the markers of the original sequence with the markers of the feedback sequence,for a rendition of G11 (Kick with the heel) by P11 (lowest-skilled participant). (scale in mm)

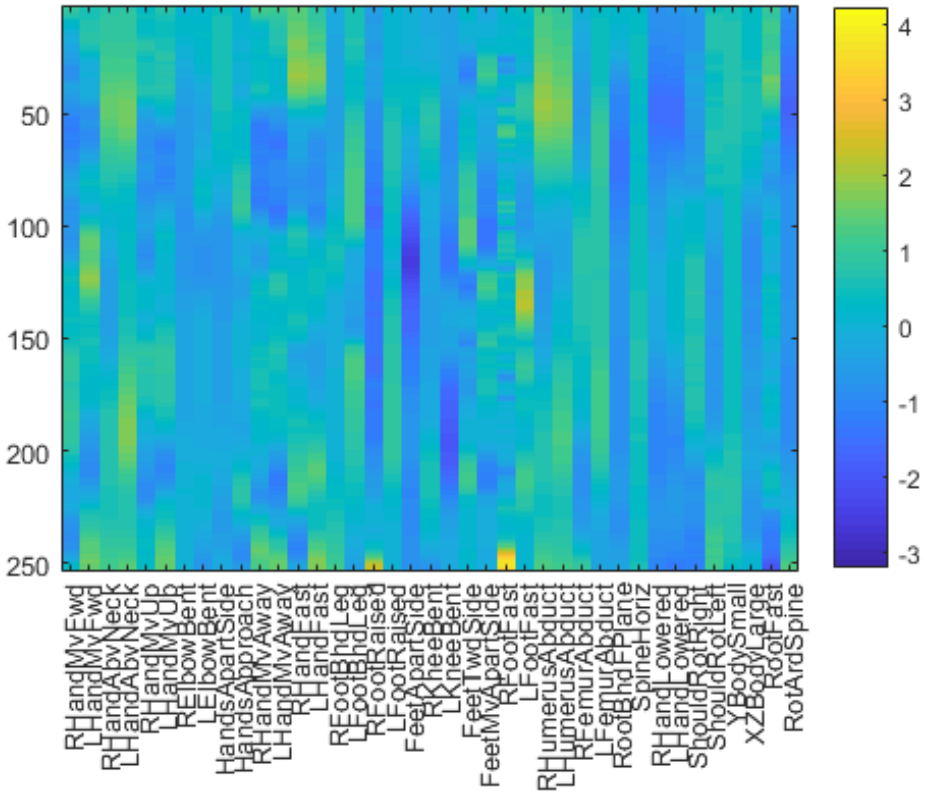

Figure A.4: Differences between relational features, for a rendition of G11 (Kick with the heel) by P11 (lowest-skilled participant). The features used are morphology-independent (processed with MIRFE), and on a standardized scale.

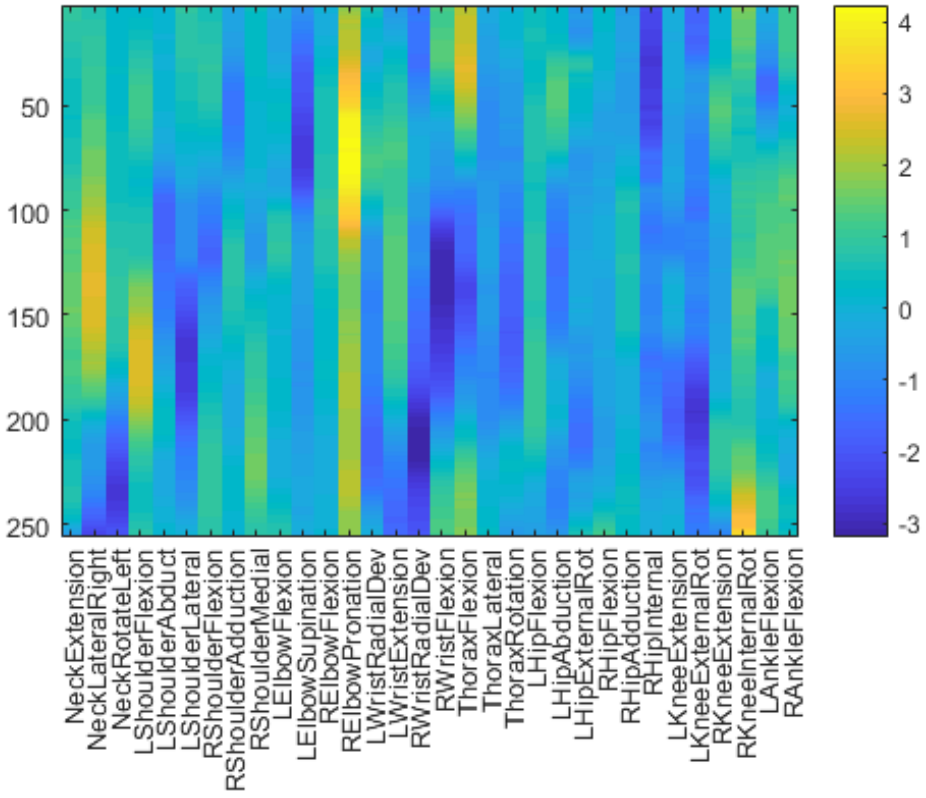

Figure A.5: Differences between ROMs, for a rendition of G11 (Kick with the heel) by P11 (lowest-skilled participant). The features used are morphology-independent (processed with MIRFE), and on a standardized scale.

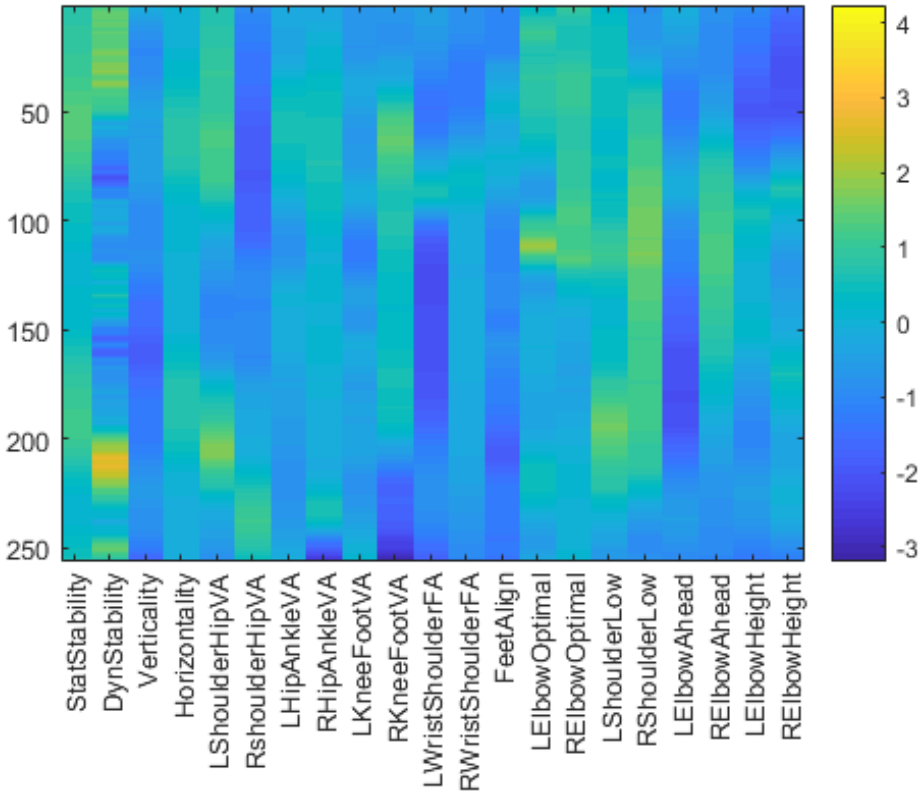

Figure A.6: Differences between some Taijiquan features, for a rendition of G11 (Kick with the heel) by P11 (lowest-skilled participant). The features used are morphology-independent (processed with MIRFE), and on a standardized scale.

Appendix B

Publications

B.1 Journals

- Tits, Mickaël and Tilmanne, Joëlle and Dutoit, Thierry, "Robust and automatic motion-capture data recovery using soft skeleton constraints and model averaging", PLOS ONE 13, 7 (2018), pp. 1-21.
- Tits, Mickaël and Laraba, Sohaïb and Caulier, Eric and Tilmanne, Joëlle and Dutoit, Thierry, "UMONS-TAICHI: A Multimodal Motion Capture Dataset of Expertise in Taijiquan Gestures", Data in Brief (2018).

B.2 Conferences

- Tits Mickaël, Tilmanne Joëlle, Dutoit Thierry, "Morphology Independent Feature Engineering in Motion Capture Database for Gesture Evaluation" in "4th International Conference on Movement Computing", London, United Kingdom (2017), ACM.
- Tits Mickaël, Tilmanne Joëlle, D'alessandro Nicolas, "A Novel Tool for Motion Capture Database Factor Statistical Exploration" in "3rt International Symposium on Movement and Computing", Thessaloniki, Greece (2016), ACM.
- Grammalidis Nikos, Dimitropoulos Kosmas, Tsalakanidou Filareti, Kitsikidis Alexandros, Roussel Pierre, Denby Bruce, Chawah Patrick, Buchman Lise, Dupont Stephane, Laraba Sohaib, Picart Benjamin, Tits Mickaël, Tilmanne Joëlle, Hadjidimitriou Stelios, Hadjileontiadis Leontios, Charisis Vasileios, Volioti Christina, Stergiaki Athanasia, Manitsaris Athanasios, bouzos Odysseas, Manitsaris Sotiris, "The i-Treasures Intangible Cultural Heritage dataset" in "IEEE Workshop on Movement and Computing" , Thessaloniki, Greece (2016), ACM.
- Tilmanne Joëlle, D'alessandro Nicolas, Barborka Petr, Bayansar Furkan, Bernardo Francisco, Fiebrink Rebecca, Heloir Alexis, Hemery Edgar, Laraba Sohaib,

Moinet Alexis, Nunnar Fabrizio, Ravet Thierry, Reboursiere Loic, Sarasua Alvaro, Tits Mickaël, Tits Noe, Zajega Francois, "Prototyping a New Audio-Visual Instrument Based on Extraction of High-Level Features on Full-Body Motion" in "Proceedings of the 10th International Summer Workshop on Multimodal Interfaces - eNTERFACE'15", Mons, Belgique, (2015).

- Tits Mickaël, Tilmanne Joëlle, D'alessandro Nicolas, Wanderley Marcelo, "Feature Extraction and Expertise Analysis of Pianists' Motion-Captured Finger Gestures" in "International Computer Music Conference (ICMC 2015)", 19, 102-105, Denton, Texas (2015).
- Lourenco Sofia, Tits Mickaël, Wanderley Marcelo, Castro Sergio, "European Piano Schools of Music Performance: Analysis towards a multimodal approach" in "International Conference on New Music Concepts (ICNMC 2015)", Treviso, Italy (2015).
- Lourenco Sofia, Martins Luis Gustavo, Wanderley Marcelo, Tits Mickaël, Megre Ricardo, "Towards a Multimodal Analysis of European Piano Schools of Music Performance" in "Conference on Interdisciplinary Musicology" , Berlin, Germany (2014).

B.3 Scientific reports

- Tits Mickaël, Laraba Sohaib, Tilmanne Joëlle, Ververidis Dimitrios, Nikolopoulos Spiros, Nikolaidis Stathis, Chalikias Anastasios-Papazoglou, "Intangible Cultural Heritage Indexing by Stylistic Factors and Locality Variations - FP7 i-Treasures Deliverable 4.5", 2016-03-13 (2016).