Analysis and Synthesis of Expressive Theatrical Movements

Pamela Carreno-Medrano

THÈSE / UNIVERSITÉ DE BRETAGNE SUD

UFR Sciences et Sciences de l'Ingénieur

sous le sceau de l'Université Européenne de Bretagne

Pour obtenir le grade de :

DOCTEUR DE L'UNIVERSITÉ DE BRETAGNE SUD

Mention : Informatique

École Doctorale SICMA

présentée par

Pamela Carreno Medrano

IRISA Institut de Recherche en Informatique et Systèmes Aléatoires

Analysis and Synthesis of Expressive Theatrical Movements

Analyse et synthèse de mouvements théâtraux expressifs

Thèse à soutenir le 25 Novembre 2016,
devant la commission d'examen composée de :

M. Frédéric Bevilacqua
Responsable d'équipe HDR, IRCAM, France / Rapporteur

M. Frank Multon
Professeur, Université Rennes 2, France / Rapporteur

Mme. Dana Kulić
Associate Professor, University of Waterloo, Canada / Examinateur

Mme. Catherine Pelachaud
Directrice de recherche CNRS, Telecom-ParisTech, France / Examinateur

M. Pierre De Loor
Professeur, ENIB, France / Invité

Mme. Sylvie Gibet
Professeur, Université Bretagne-Sud, France / Directeur de thèse

M. Pierre François Marteau
Professeur, Université Bretagne-Sud, France / Co-directeur de thèse

Contents

Chapter 1

Introduction

Contents

One of the main challenges in computer character animation is to design compelling characters capable of creating a more intuitive, engaging and entertaining interaction with a user. Whereas we play with these animated characters, observe them or control them, the quality of the interaction and our level of engagement strongly depend on how believable we consider these characters to be.

The notion of believability, and consequently of believable characters, comes from the traditional arts (film, theater, drama, literature, etc.) and describes the property of a synthetic character to engage in consistent, life-like and comprehensible behavior in such a manner as to provide the illusion of life [171]. Once we perceive an animated character as having a life of its own, we suspend all judgment about the implausibility of what we see and perceive in favor of our experience and interaction with this character. By doing so, natural responses – inherent to human-to-human social behavior and interaction – as engagement and empathy will emerge [22].

Although the idea of believability is not new – it was brought into the computer animation and artificial intelligence fields by [17] at the beginning of the nineties –, it still remains a highly subjective concept for which there is no generally agreed or precise definition. Nevertheless, among the existing definitions and characterizations of believability [163, 171, 195, 168], effective emotional communication is a recurrent element and it is often defined as one of the most important qualities of believable characters. An animated character that is able to perceive and interpret the emotions of others and to express emotions in an appropriate and comprehensible manner, it is most likely perceived as believable since the expression of

emotions relates to the perception of most complex internal processes as personality, mind, intentions, desires, etc.

Humans receive and send emotional information through a variety of channels: prosody, eye gaze, facial expressions, posture, and body motion. For animated characters, movement is the principle means of communication and interaction within the virtual world they inhabit [25]; hence they portray emotional information through movement and the way we perceive and understand them has everything to do with how they move. Based on these ideas, the quest for believability has sent researches into two different paths: one pragmatic approach, hereinafter referred as computer animation, whose main goal is to generate physically realistic animated characters (i.e., characters that look and move like humans); and other, hereinafter referred as affective computing, that seeks to understand and exploit the power behind human emotional communication to facilitate and improve human-machine interaction. However, visually realistic characters are not necessarily perceived as believable; and building a system that successfully discriminates emotion from one or several channels does not imply that the same system can convey comprehensible emotion-related information to a user.

1.1 Motivation and Main Challenges

The work I present here lies in the intersection of these two research fields. Founded on recent studies that show the importance of body motion in the perception of emotion; and in the successful traditionally animated characters that illustrated how accurate emotional cues can be incorporated into body movement, I aim to generate physically plausible and expressive body motions for virtual characters through data-driven methods. This goal poses significant research challenges:

1. **Define a significant motion corpora:** The human body has hundreds of degrees of freedom and can perform an extensive variety of movements of diverse complexity. From simple tasks many of us do not consciously think about as walking, pointing and grasping to motions that require high precision and self-control as artistic gymnastics and acrobatics. All of this while implicitly communicating emotions, intentions, desires, etc. Data-driven generation of expressive body motions requires the construction of a motion corpora that accounts for the motion patterns we wish to replicate on the synthesized motions. More precisely, it must comprise body motions on which emotions can be easily recognized, and thus measured and modeled.

2. **Build a model that accounts for the cues that are indicative of the expression of affect and emotions:** Despite its high-dimensionality and complexity, human motion is extremely redundant, correlated and coordinated. Thus, it is possible to extract a simpler and low-dimensional motion representation that abstracts out the unneeded complexity while preserving all the dynamic and kinematic motion cues that characterize the effective communication of affect. Additionally, as many of these cues are not directly observable but are rather encoded in the execution of the action itself, the resulting motion model should remove as much action-dependent information as possible while preserving all affect-related cues.

3. **Synthesize new expressive bodily motions based on a motion model**: Once a low-dimensional motion model has been define, new expressive bodily motions can be produced. Under this type of motion model, the synthesis of whole-body movements generally involves two stages: *i.)* generation of a sequence of observations in low-dimensional space; *ii.)* a mapping that relates those observations to high-dimensional full-body space, while retaining all emotion-related features. The resulting sequence of observations usually depends on a set of constraints and control signals provided to the synthesis algorithm.

4. **Evaluate quantitative and qualitatively the resulting animations:** The quality of the synthesized expressive bodily motions needs to be assessed through objective and subjective measures. The former determines whether the resulting motions exhibit the same kinematic and statistical patterns than the examples comprised in the motion capture database; the latter assess the perception of the generated expressive motions in comparison with those produced by a real human.

1.2 Contributions of Thesis

The overall objective of the work presented herein is to contribute to the creation of more compelling and expressive animated characters. To achieve this goal, we first define a new motion capture corpora inspired by the performing arts (theater, pantomime, mime, etc.). Our aim is to capture bodily motions that are especially crafted to communicate meaning and emotion to an audience. As findings reported by the psychology research community [205, 234, 32] underline the importance of body extremities in the discrimination of biological motion and emotional states, we then propose to characterize and synthesize expressive bodily motions through the spatio-temporal trajectories of six main joints in the human body: head, wrists, feet and pelvis. This claim is further supported by previous applications of such representation to motion compression [239], motion retrieval [145] and performance animation [51]. The suitability of this representation as well as the quality of the synthesized motions are assessed through a perceptual study and two quantitative methods: automatic classification of affect and divergence measures computed from empirically estimated probability distributions.

Our contribution are as follow:

- A new motion capture dataset consisting of a theatrical scenario in which a magician performs three different magic tricks. This database also includes examples of locomotion and short improvisation sketches. Five actors performed these five motion sequences under four emotional states (happiness, stress, sadness, relaxedness) and the neutral condition. A total of two-hundred and seventy-five motion examples, with durations ranging from 7 to 78 s., were recorded.

- A simple and intuitive, yet powerful, low-dimensional parameterization of expressive bodily motions. We validate and evaluate the relevance and suitability of this parameterization within two studies, which quantify the loss of affective information we may induce when only a limited subset of spatio-temporal trajectories are considered instead of whole body information.

- A motion synthesis approach in which new expressive bodily motions are generated by randomly sampling from the low-dimensional space spanned from the end-effectors and pelvis trajectories. Furthermore, we show that once a set of expressive trajectories has been generated, a simple multi-chain Inverse Kinematics controller suffices to synthesize whole-body expressive motions. This model generates motions that preserve the expressiveness of the examples while overcoming the semantic content of the movement (all the information related to the functional behavior depicted in the motion, e.g., a gait cycle, grasping an object,...). Thus, it is possible to isolate and measure the expressive content only.
- A qualitative and quantitative evaluation of the synthesized and reconstructed motions. We evaluate the expressive content of the resulting motions within three studies: a perceptual study in which both the synthesized end-effector trajectories and the reconstructed whole-body motion are compared with those displayed by a real human; a classification experiment in which the performance of a classifier trained on real human data is tested on the synthesized motions; and a distance-based measure in which the statistical and stylistic similarity between the ground-truth data and the synthesized motions is assessed.

1.3 List of Relevant Publications

The following publications were produced based on the research presented in this thesis:

1. Pamela Carreno-Medrano, Sylvie Gibet, and Pierre-François Marteau. "Synthèse de mouvements humains par des méthodes basées apprentisage : un état de l'art". In: *Revue Électronique Francophone d'Informatique Graphique* 8.1 (2014)
2. Pamela Carreno-Medrano, Sylvie Gibet, Caroline Larboulette, and Pierre-François Marteau. "Corpus Creation and Perceptual Evaluation of Expressive Theatrical Gestures". In: *Intelligent Virtual Agents - 14th International Conference, IVA 2014, Boston, MA, USA, August 27-29, 2014. Proceedings*. 2014, pp. 109–119
3. Virginie Demulier, Elisabetta Bevacqua, Florian Focone, Tom Giraud, Pamela Carreno, Brice Isableu, Sylvie Gibet, Pierre De Loor, and Jean-Claude Martin. "A Database of Full Body Virtual Interactions Annotated with Expressivity Scores". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.* 2014, pp. 3505–3510
4. Pamela Carreno-Medrano, Sylvie Gibet, and Pierre-François Marteau. "End-effectors trajectories: An efficient low-dimensional characterization of affective-expressive body motions". In: *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015, Xi'an, China, September 21-24, 2015*. 2015, pp. 435–441
5. Sylvie Gibet, Pamela Carreno-Medrano, and Pierre-François Marteau. "Challenges for the Animation of Expressive Virtual Characters: The Standpoint of Sign Language and Theatrical Gestures". In: *Dance Notations and Robot Motion, 1st Workshop of the Anthropomorphic Motion Factory, at LAAS-CNRS, Toulouse, France, 13-14 November, 2014*. 2016, pp. 169–186

6. Pamela Carreno-Medrano, Sylvie Gibet, and Pierre-François Marteau. "From Expressive End-Effector Trajectories to Expressive Bodily Motions". In: *Proceedings of the 29th International Conference on Computer Animation and Social Agents*. CASA '16. Geneva, Switzerland: ACM, 2016, pp. 157–163

Chapter 2

A Motion Model Based on End-Effector Trajectories

Contents

Animating virtual articulated human characters whose movements are perceived as expressive is a difficult task for mainly three reasons. First, the human body is a complex mechanism composed of hundred of bones and muscles and capable of executing complex spatio-temporal motion patterns. Even though animated characters are represented as simplified abstractions of this mechanism, generating varied, compelling and humanly plausible motions for those simplified representations remains an open challenge [10, 57]. Second, despite the extensive research on the important role of human body motion in expressing emotions and other affective phenomena done during the past decade [8, 136], there is still a limited and incomplete understanding about what aspects of movement, also called motion features or motion qualities, have the most significant impact on the perception and expression of emotion and affect. Furthermore, since human motion is conditioned by factors such as individual characteristics, personality, cultural constructs, among others [138, 240], it is difficult to isolate, extract and analyze the motion features that are mainly related to the expression of emotion-related information. Last but not least, there is a one-to-many correspondence between emotions and movements. That is, the same emotional state can be communicated through a large choice of kinematically distinct movements [211]. Thus, determining the motion cues that are common to different bodily expressions of the same emotional state and hence should be considered and mimicked when animating an expressive virtual character is still an open challenge [78]. We find then that animating a virtual human

character with expressive movements supposes to: *i.)* identify the fundamental movement aspects and parameters that should be used to generate perceptually expressive motions and *ii.)* determine how these aspects are mapped to and control the simplified articulated mechanism used to represent a character's body. Furthermore, the resulting motions should preserve the naturalness and visual appeal human observers usually associate to other humans' movement [250].

Due to the complex, high-dimensional, and dynamic nature of human motion, several assumptions have to be made in order to facilitate the analysis and inference of the relationships between emotions and the characteristic changes they produce in body movement. We highlight below the two most common assumptions found in the literature.

The first assumption relates to where the knowledge of what makes human motion expressive can be found. Two main approaches can be identified:

- The first one has its foundations on psychology studies and supposes that movement aspects important for the generation of expressive motions can be determined by analyzing how humans perceive emotions via different types of visual stimuli [8, 56, 255]. The identified motion patterns are then used to build computational models that can be invoked by an user to generate novel expressive motions [192]. The intuition behind this approach is that by using perceptually validated motion aspects to guide the generation of the character's movements, the perceptual and expressive qualities of the new motions will be ensured. There are three main limitations associated to this kind of methods: *i.)* most of the motion patterns identified in this manner are of qualitative nature and an additional effort is needed in order to translate them into concrete numerical parameters that can be applied to particular body movements [131, 268], *ii.)* an additional motion model is required in order to constrain the synthesized motions to the space of humanly plausible and visually appealing motions, and *iii.)* movement qualities sufficient for an accurate emotion perception may not be sufficient for the generation of new expressive motions [211].

- The second approach applies data-driven techniques based on machine learning methods to examples of expressive motion capture data in order to automatically determine and learn what movement aspects characterize the expressiveness in the captured data [78, 211]. The relevant motion qualities are determined either by estimating the differences between the emotionally expressive and non-expressive realizations of the same kinematic action [79] or by separating what is being done (content) from how it is being done (expressiveness) [232]. The main advantages of this approach are that the relationship between the emotion-related motion qualities and numerical motion parameters is automatically determined by the machine learning methods and that the generated motions are guaranteed to be visually appealing and respect implicit human motion constraints such as joint limits. However, this kind of models highly depend on the dataset of expressive motions [78, 106] and are hard to interpret and understand.

The second assumption supposes that the knowledge and motion qualities obtained by analyzing a particular type of motor behavior such as human locomotion [16] or a determined kinematic action, e.g., knocking motion [197], generalize to all other possible body movements. However, motion aspects related to the expression of emotions can be confounded by the characteristics of the studied movement or motor behavior itself [58, 101].

Thus, in order to infer and establish more general principles and characteristics of emotional body expressions, it is of critical importance to study and analyze much broader movement classes [205].

To ask a human observer to rate hundreds of examples of expressive motions seems to be an inefficient and tedious way of determining which motion qualities relate to a particular emotional state across different motion classes. To do the same through machine learning methods will require a large amount of sufficient and consistent training examples, since the number of samples required to adequately cover and describe human motion space increases exponentially as a function of dimensionality. Furthermore, the complexity and computational cost associated to such methods will considerably increase with the dimensionality and heterogeneity of the dataset used during training [258]. Fortunately, despite the high-dimensionality and complexity of the observed human motion space, it has been shown that due to the bio-mechanical and functional constraints ruling human motion, e.g., legs ands arm operate in a coordinate way in most cases [210], most human movements have an intrinsic representation in a lower-dimensional space [73]. We believe that this low dimensional space can be used:

i.) to efficiently characterize and analyze diverse examples of kinematically distinct expressive human motions;

ii.) to determine the motion cues indicative of expressed emotions; and

iii.) to animate expressive virtual characters.

The definition, validation, and use of this low-dimensional space represent the main contributions of this thesis.

In the following sections we detail how by incorporating expert-knowledge and known results on different research areas (i.e., human perception of emotion and biological motion, motion synthesis and automatic recognition of affect), we define a low-dimensional motion space that we believe successfully characterizes expressive motions. This low-dimensional space is both intuitive and easy to analyze, and can be directly used to control a virtual character's body. We then briefly describe the validation, synthesis and evaluation methodologies adopted in the work presented in this thesis. We end this chapter with the outline of this dissertation.

2.1 Low-Dimensional Motion Space and Motion Model Based on End-Effector Trajectories

The task of recovering meaningful low-dimensional structures hidden in high-dimensional data is known as dimensionality reduction [225]. An extensive number of dimensionality reduction methods can be found in the literature. From the simple and widely used Principal Component Analysis (PCA) to more complex and powerful methods such as Gaussian Process Dynamical Models (GPDM) [258]. Each of them makes different assumptions about the relationship between the high-dimensional data and the low-dimensional representation, provides or not a function mapping points in low-dimensional space to their respective pre-images in the original high-dimensional representation, and produces low-dimensional

representations with different purposes [26]. For instance, a low-dimensional representation that takes into consideration the class membership of the high-dimensional data points will be considerably different from another low-dimensional representation which purpose is to maximize the variance of the projected data.

In this thesis, instead of determining a low-dimensional space through standard and ready-to-use dimensionality reduction methods as the ones we just briefly reviewed, we propose to define such space based on expert-knowledge about: *i.)* the expression of emotions via body movements and *ii.)* the available character's motion generation and control techniques. We believe that the resulting low-dimensional space will be suitable for both the analysis and generation of novel emotionally expressive motions. We have chosen this course of action because of the following reasons:

- A particularity of working with expressive bodily motions is that the motion features we seek to infer and generalize are usually embedded in an intricate manner in the execution of the motion itself and in their own have small spatio-temporal significance [78, 114]. Thus, it is likely that the minimal number of free variables needed to represent the original high-dimensional data within the resulting low-dimensional space will be determined and associated to the motor behavior within the motion example rather than to the emotional stated conveyed by the expresser. Therefore, if we apply whatsoever standard dimensionality reduction method to a highly heterogeneous dataset of expressive human motions, it might happen that:

  i.) the resulting low-dimensional space will be determined by the motor behaviors in the dataset rather than by their expressive content, and

  ii.) the number of free variables necessary to define a low-dimensional space that accounts, with minimal information loss, for the motion features and qualities in which we are interested will considerable increase with the size and heterogeneity of the high-dimensional data.

  For these reasons, we believe that such low-dimensional space should be better defined using the motion dimensions and features that have been consistently reported as critical for the expression and perception of emotions. By doing so, we ensure that the dimensionality of this space will be independent of the size and heterogeneity of the original high-dimensional data and that the new representation likely contains the information we are interested to study and analyze.

- Recent studies on the perception of emotion from bodily motions have pointed out that although human observers seem to process expressive movements as a whole [205], different body regions convey different amounts of emotion-related information and motion features [187]. Thus, it might not be necessary to process and analyze the entire body in order to determine the motions aspects the most salient to the perception and generation of expressive bodily motions [78]. A low-dimensional space that accounts for the spatio-temporal patterns of the most meaningful and informative body parts can be used instead.

- Independently of the manner in which the motion cues that are relevant for the perception and expression of emotions and affect are inferred, we still need to define how

those qualities might be used and/or applied during the generation of new expressive bodily motions. By choosing a low-dimensional representation that contains most, if not all, of these relevant motions cues and that at the same time can be directly used to control the motion synthesis process, the additional effort of defining the relation between motion qualities and motion synthesis parameters [131] is not longer necessary. The choice of such representation can be also guided by our knowledge of the different motion synthesis and control techniques commonly used in character animation applications.

The low-dimensional space that we propose consists of the spatio-temporal trajectories of both the extremities of the limbs of a human-like character, i.e., *head*, *hands* and *feet*, and the equivalent of the human pelvis in the character's body, i.e., the root joint. The extremities motion, hereinafter called end-effector trajectories, simultaneously account for the kinematic and motion-mediated structural information (i.e., how body limbs move with respect to each other) of a movement [240]. Furthermore, as we will discuss it later in this thesis, the position of a limb's extremity determines in large part the orientation of other joints within the chain [145, 236], e.g., hand position determines the orientation of elbow and shoulder. The root joint trajectory provides information about the motion path described by the animated character while traveling and moving in world space. Additionally, both the root and feet trajectories describe the character's lower-body motion behavior.

Below we detail and discuss the expert-knowledge on which we based our choice of low-dimensional representation. We survey results from three different research areas: perception of biological motion and emotions, automatic recognition of affect and motion synthesis.

2.1.1 The Importance of End-Effector Trajectories in Perception of Biological Motion and Emotions

Regardless of the high-dimensional and redundant nature of human body, studies about the perception of biological motion have shown that human observers seem to effortlessly recognize and extract information about human motion from sparse body representations [27]. Point-light (PL) displays [124] are one of the most common sparse representations used in perceptual studies. They consist of only a handful of markers attached to the head and main joints of the body. Since their introduction by Johansson [124] in 1973, PL displays have been proven to contain enough information to make possible for humans the determination of gender [240], affective states [7], and the identity of individuals [241]. These results indicate, from a perceptual standpoint, that: *i.)* there exists a simple, yet rich, low-dimensional representation that accounts for all relevant motion information [240] and *ii.)* it is possible to parameterize human motion and all the nuanced affect information conveyed by it through a limited set of motion trajectories, i.e, a spatial subset of body joints [80, 187].

Several studies have aimed to determine which are the main parts of the human body from which cues, also referred to as features, that are critical for the perception of biological motion and emotions are extracted. End-effectors are often among this group. Ekman and Friesen advocated that head motion provides strong cues about the nature of an emotion (whether the moving subject conveys anger, fear, sadness, etc.) [72, 131]. This claim was further confirmed by the work done in [18]. Beck *et al.* [18] conducted a user study in which

the effect of head position in the interpretation of emotional body language was studied. They changed head positions across bodily depictions of six different emotions in a Nao robot. They found that changing head position does indeed affect the expressiveness of the analyzed emotion displays. For instance, moving the head down decreased arousal and valence ratings, while moving the head up produced the opposite effect. Furthermore, they confirmed that head position has a strong effect on the correct interpretation of emotion displays. Anger displays were interpreted as happiness or excitement when the head was up and pride displays with the head down were instead labeled as anger.

There is evidence suggesting to a great extent that hand-arm movements are most significant when distinguishing between affective expressions [255]. This claim has been validated by diverse studies. Pollick *et al.* [197] successfully conveyed different emotional states through arm and hand drinking and knocking movements only. Observers recognized all emotions above chance level. Hietanen and collegues [116] reported that affective expressions can be decoded from hand movements used during Finish sign language communication by human observers having no knowledge of the language itself. Sawada *et al.* [216] observed that hand-arm trajectories considerably differ between motions expressing different emotional states. For instance, happy motions are characterized by indirect arm-hand trajectories, while sad movements include slow and weak hand-arm trajectories.

Rather than studying the importance on individual body parts, other studies have focused on determining which particular joints and/or body parts receive more attention from human observers. There is consistent evidence that head and arm-hand movements are of significant importance for human observers when distinguishing between emotional states. Montepare and colleagues [178] were one of the first ones to suggest the influence of head and arm motion on the perception of emotion from gait. This observations was later confirmed by the work presented in [205]. Roether *et al.* [205] analyzed and identified emotion-specific posture and movement features critical for the perception of emotion from gait examples. They found that those critical features depend only on a small number of joints rather than on the whole body. Specifically, they identified head and arms trajectories as the most important sources of information for perception of emotion and automatic analysis of expressive bodily motions, even when movements of the entire body are presented.

Finally, although feet trajectories have not yet been identified as important sources of information for the expression and perception of affect and emotional states, several studies have advocated for their influence in the perception of biological motion. Thurman and colleagues [234] aimed to reveal the key features human observers use to discriminate biological motion stimuli. They found that human observers strongly relied in features extracted from the lower-body kinematics (feet trajectories). Similarly, van Boxtel *et al.* [32] identified feet and hand trajectories as the most consistently critical motion sources used by observers for action discrimination.

2.1.2 The Importance of End-Effector Trajectories in Automatic Recognition of Affect and Emotions

Automatic recognition of affect and emotional states is another of the research areas interested in identifying a low-dimensional representation that is the most salient for the inference of affective and emotional states from bodily movements [211]. The identification of this space will not only make easier the discrimination between distinct affective and emotional

states, but it will also shed some light on our understanding of how humans perceive and express emotions via body movements. Results provided by several previous works on automatic recognition of affect from body motion support our hypothesis of using end-effector trajectories to define such low-dimensional space.

A survey of the most recent literature on automatic recognition of affect reveals that head and hand trajectories are among the bodily cues the most frequently and successfully used [105]. Bernhardt *et al.* reported 50% recognition rates on upper-body functional movements (knocking motion) depicting neutral, happy, angry and sad emotional states. Their classifier was trained using four kinematic features (maximum distance, average speed, acceleration and jerk) computed from hand trajectory only. Bouënard *et al.* [29] analyzed expressive percussion gestures and found that a reduced dimensional representation consisting of motion features computed from hand trajectories was sufficient for accurately classifying new expressive percussion gestures. Glowinski and colleagues [95] found that it is possible to define a minimal representation of expressive upper-body movements by analyzing the kinematic qualities of head and hand trajectories. The resulting representation was later used to determine meaningful groups of emotions. Similarly, several studies [103, 131, 218] indicated that through the analysis of simple head movements and the way they are performed it is possible to obtain above chance recognition rates in a person-dependent context.

Rather than investigating automatic recognition of affect and emotions from isolated body parts (e.g., hand and head) movements, Sadamani *et al.* [213] focused on whole body expressions of affect. Through a combination of hidden Markov models (HMM), Fisher score representation and supervised principal component analysis (SPCA), they obtained a minimal discriminative representation of expressive bodily motions. A thorough analysis of this minimal representation revealed that hands and head trajectories are the most informative joints for affective full-body movement recognition.

2.1.3 The Importance of End-Effector Trajectories in Computer Character Animation

In computer character animation, task-based motion editing and synthesis is one of the most commonly used approaches [250] for the generation of novel body movements. In this approach, the animation system aims to automatically generate or modify a body motion such that a set of constraints provided by a user are respected. Due to their easy specification and intuitive use, end-effector trajectories are one of the most commonly employed types of constraints. They can simultaneously specify the spatio-temporal as well as expressive properties that should be present in the resulting motion [198]. For instance, [28] used hands trajectories as the input of a hybrid motion control system that generated upper-body percussion motions. Aubry and colleagues [9] also used hands trajectories as the control signal of their a synthesis system.

Motion reconstruction, also known as performance animation, is one the most successful examples of motion generation via end-effector trajectories. In this particular application, devices such as inertial sensors [161], accelerometers [230], 3D motion sensors [132] and retro-reflective markers [51] are used to record and specify the trajectories of head, hands, feet and pelvis that define the desired motion. From this information, it is possible to generate full body motions that smoothly follow the specified trajectories and exhibits the correct temporal variations [50].

End-effector trajectories have also been successfully employed in other character animation related areas. In motion indexation and retrieval, Krüger and colleagues [145] showed how a feature set consisting of end-effector trajectories is sufficient to characterize, index and query full body motions. The authors compared end-effector trajectories against richer motion representations (e.g, end-effector trajectories plus the trajectories of elbows, knees, and one chest joint) using two of the largest motion capture databases publicly available: CMU [42] and HDM05 [180]. They found that the use of richer representations give little or not advantage over the use of end-effector trajectories. In motion compression, Tournier *et al.* [239] observed that it is possible to recover, with minimal information loss, whole-body poses and motions by specifying only the end-effectors and root positions of a skeleton structure as input.

2.1.4 Discussion

The perception of emotion and the automatic affect recognition literatures consistently reported the importance of head and hand trajectories. We observe that in isolation or combined, both body joints provide enough cues to make the distinction between different emotional states possible for human observers and automatic classifiers. Furthermore, we found also that even though full body information has been used to define a low-dimensional space through a combination of either automatic dimensionality reduction and stochastic modeling [213] or nonlinear mixture model and sparse regression [205], it is the features extracted from hands and head trajectories that later emerged as the most salient [205, 213].

It is important to notice that although lower-body motion (i.e., feet and pelvis trajectories) have not received as much attention as head and hands trajectories in both emotion perception and automatic emotion recognition literatures, they remain key elements in the specification of full-body biological motion [32, 234]. This claim is supported by previous work on motion synthesis, compression and retrieval [132, 145, 239], in which feet and pelvis trajectories are critical for indexing, reconstructing and synthesizing full body motions. Furthermore, results reported by [145] indicate that richer representations that include more body joints do not entail higher accuracy in the specification of highly diverse human poses and motions.

Based on this evidence, we believe that a low-dimensional representation consisting on end-effector trajectories will simultaneously provide:

- *i.)* most of the motion qualities necessaries for ensuring and enhancing the perception, recognition and generation of emotional content;
- *ii.)* all the spatio-temporal specifications needed for generating and reconstructing expressive whole body motions; and
- *iii.)* an intuitive, comprehensible, and useful motion representation that can be employed for the analysis and control of diverse expressive motion classes (non-periodic movements, goal-oriented actions, performing arts motions, among others).

2.2 Methodology and Outline of This Thesis

We hypothesize that a low-dimensional space defined by the spatio-temporal trajectories of five end-effectors (i.e., hands, head and feet) and the pelvis, also referred to as root joint, is both sufficient and suitable for the analysis and generation of expressive bodily motions. In order to assess the validity of our hypothesis, we propose the following methodology:

1. **Qualitative and quantitative measures of how informative are end-effector trajectories with respect to whole-body motions**: To say that the end-effector trajectories are suitable for the analysis of expressive bodily motions suppose that the selected representation accounts for most of the motion features and qualities that are indicative of the expressed emotional states. Furthermore, this implies that when compared to the original motion representation (whole-body motions), our low-dimensional parameterization should provide results closer to the ones obtained from whole body representations. Inspired by this idea, we proposed and conducted:

   i.) A user study in which we measured how observers' perception of emotions changed according to the type of visual stimuli (end-effector trajectories or whole-body motions) presented to them. The study design, the motion capture database we employed and the results we obtained are all discussed in Chapter 3.

   ii.) A quantitative evaluation in which the performance of an automatic affect classifier trained on different subsets of motion features was analyzed and compared. Within this evaluation we considered the fact that the motion features computed from other joint combinations might be as informative as those computed from end-effector trajectories. Hence, we analyzed two cases: *a.)* features from the proposed low-dimensional representation against features from whole body representation and *b.)* features from the proposed low-dimensional representation against features subsets automatically determined via feature selection methods. A detailed description and discussion of the motion features employed for classification, the feature selection method we used and the classification tasks on which all representations were evaluated is presented in Chapter 4.

2. **Generation of novel expressive bodily motions and assessment of the quality of the generated motions**: The use of a low-dimensional motion representation implies that a mapping function between high-dimensional full-body space and low-dimensional trajectory space needs to be defined. This mapping should preserve all motion cues indicative of expressed emotions that are present in the low-dimensional representation. Furthermore, in order to assess whether the proposed low-dimensional representation generalizes beyond the movements in which it has been primarily tested, it is necessary to evaluate full body motions generated from novel end-effector and pelvis trajectories. To do so, we have proposed and implemented:

   i.) An inverse-kinematic based mapping model. Since the geometry and configuration of an anthropomorphic limb is quite dependent on the limb's extremity position [145], herein referred to as end-effector, we generate full-body motions by defining an inverse kinematic controller for each limb within the character's

body. Each controller determines the best sequence of poses such that the end-effector trajectory used as control signal is followed as smoothly and closely as possible by the corresponding limb.

ii.) A re-sampling scheme that can generate novel end-effector and pelvis trajectories while preserving the underlying emotional content. The main principle behind the proposed re-sampling scheme is the generation of random, semantically empty, trajectories that considerable differ from the original low-dimensional observations. That is, trajectories that are independent of the motor behaviors observed in the MoCap database.5 By evaluating the full body motions from those trajectories, we can have an initial assessment about the generalization capabilities of the proposed low-dimensional representation.

Once we have defined a function that maps end-effector and pelvis trajectories to full body motions and a re-sampling scheme that generates novel low-dimensional trajectories, the next step is the evaluation of the generated and reconstructed motions. To do so, we proposed and conducted one qualitative and two quantitative evaluations. In each evaluation we considered three movement generation sources: original MoCap database, motions reconstructed from end-effector trajectories issued from the MoCap database and motions generated using re-sampled trajectories. The three evaluations consisted on:

i.) A user study in which we measure the impact of the movement sources on the perception of emotion.

ii.) A first quantitative evaluation in which we compared the recognition rates of an automatic affect classifier tested on motions generated from the three sources we listed above.

iii.) A second quantitative evaluation in which we measured the statistical similarity between: *a.)* MoCap database and motions reconstructed from end-effector trajectories issued from the MoCap database and *b.)* MoCap database and motions generated using re-sampled trajectories.

The inverse-kinematic based mapping model, the re-sampling scheme, and the evaluations designs and their results are presented and discussed in Chapter 5.

Chapter 3

A Theater Motion Capture Database

Contents

As it was stated in the introduction of this thesis, the purpose of working with expressive bodily motions is to make virtual characters look more alive, believable and engaging in the eyes of the humans interacting with them. In this context, emotions and any other affective phenomena (e.g., affect, mood, feelings, attitudes, etc. [217]) are the ideal means to indicate that characters have an inner state and, to some extent, a life of their own.

In a very broad sense, emotions can be defined as a set of mechanisms that facilitates an individual's adaptation to his constantly changing surroundings [217]. This implies that the individual makes internal judgments on the environment, his self and the other individuals, and responses through changes in his state and/or his behaviors. All of this is done accordingly to the individual's goals, beliefs and well-being [101, 120]. These internal changes and responses often have external manifestations visible and recognizable to others. Such manifestations can be intentional or not, and constitute important signals for social interactions and communication [68]. Hence, both emotions and affective expressions are important elements in the generation of compelling and successful human-character interactions.

Evidence from many diverse tasks [7, 56, 197, 205, 216, 255] has shown that humans are capable to identify the external manifestations of emotions and affective phenomena from bodily expressions in the absence of other channels such as facial and vocal cues [8]. As it was previously stated, movement is the main means of communication and interaction for animated characters, which explains our interest in the expression of emotions through body motions. However, working with expressive motions suppose an understanding of how body movements are quantitatively and qualitatively modified when an individual is experiencing an emotion.

Emotions as well as the way in which they are expressed or signaled through body motions remain highly subjective and complex open problems. In order to gather an understanding about bodily expressions of emotions, the recording of the movements and emotions of interest through different devices and technologies is the most common approach. In this chapter, we briefly review the existing and available expressive motion corpora and describe the corpus on which all the work presented in this thesis was done.

3.1 Emotions, Elicitation and Categories of Expressive Bodily Movements

Recording motion corpora for the analysis and synthesis of body motions expressing emotions and other affective phenomena requires a careful design and raises important questions: how to describe and represent the emotions or affective states of interest?, what kind of movements or motor behaviors to record?, how many subjects need to be recorded?, should emotions be elicited, felt or portrayed? All these important elements should be taken into consideration when building a new expressive motion corpora. Similarly, these elements can be also used to analyze and determine the suitability of existing corpora. In this section, we review the emotion models and theories most frequently employed for the study of expressive bodily motions. Furthermore, we discuss the variety of body movements found in the existing motion corpora as well as the procedures used to gather this type of data.

3.1.1 Emotion Theories

Several theories on the definition and categorization of emotions have been proposed by psychologists and other emotion researchers. Here we review the models most commonly used in the construction of expressive motion corpora and computational modeling of emotions. A larger overview of the different emotion models and taxonomies can be found in [217, 214].

Discrete or categorical model

First postulated on Darwin's book *The expression of emotion in man and the animals*, this model argues that there is a limited number of basic or fundamental emotions. Each of this basic emotions has its own elicitation conditions and its own physiological, expressive, and behavioral patterns [217]. Most complex, nonbasic emotions are made up of blends of these fundamental emotions. The most popular set of basic emotions was proposed by Ekman

[71] and consists of anger, happiness, sadness, surprise, disgust and fear. Many of the current research on emotion perception and recognition has predominantly focused on this categorization.

Dimensional or continuous model

Emotions are described in terms of a continuum spanned by a few independent dimensions. One of the most popular examples of these dimensional models is the *circumplex model of affect* proposed by Russel [207]. This is a two-dimensional model in which emotions can be easily defined and differentiated by their position along the pleasantness-unpleasantness (valence) and active-passive (arousal or activation) dimensions. This allow researchers to easily differentiate positive and negative emotions of different levels of intensity and activation [217]. Models with additional dimensions such as dominance [173], attention, action tendency [20], among other have been also proposed. Since dimensional models focus on the subjective component of emotions, verbal and categorical labels can be easily mapped to the continuous space.

3.1.2 Categories of Expressive Bodily Movements

Both Atkinson [8] and Gross *et al.* [101] suggested that there are two ways in which emotions can be expressed through bodily movements. First, we have what [8] referred as to *expressive actions*, namely movements that are direct manifestations of one's internal emotional state. Expressive actions can be sometimes motivated by the deliberate need of communicating emotions and affective content [20]. Some of these expressive actions have been adopted as conventional gestures with an emblematic meaning. For example shaking a fist or raising one's arms are motions people commonly associate to anger and joy respectively. Second, we find what [101] denoted as *non-emblematic movements*. That is, everyday movements or actions that are performed in an emotional way. In other words, motions motivated by goals other than to communicate affective content are modulated and modified in such a manner that an emotional state is expressed and recognized. An example of non-emblematic movements is a slumped walk that might suggest a sad state.

Existing expressive motion corpora consider one of these two approaches. They consist of either movements associated with the expression of an emotion (expressive actions) or movement behaviors in which we can study what kind of modifications are characteristic of a particular emotional state (non-emblematic movements). A more thorough classification is possible if we take into consideration the spontaneity and naturalism of the expressive motions. Both Bazinger *et al.* [15, 14] and Cowie *et al.* [58] defined three major categories of emotional expressive corpora:

- **Natural emotional expressions** occurring in everyday life settings. The researcher does not directly control or influence the recorded emotion expressions. Motion corpora belonging to this case can contain both expressive actions or non-emblematic movement behaviors.
- **Portrayed emotional gestures** where expressions are produced upon instructions. Expressers are asked to produce motions in which emotions can be easily recognized

by external observers. Movements can be performed freely or the expresser is told precisely what to do and how to do it. We find in this category the aforementioned conventional gestures often referred as to emotion archetypes.

- **Induced emotional expressions**, also called felt experience enacting, are a combination of induction methods and movements deliberatively communicating affect. Expressers – often trained laypersons or professional actors – are asked to produce plausible and believable expressive motions occurring in a controlled setting. The researcher uses elicitation techniques in order to facilitate the enacting of the expressive motions [14]. Most of the motion corpora belonging to this category contain non-emblematic movement behaviors.

3.1.3 Induction Methods

Induction methods, also known as mood induction procedures (MIP) can be used to elicit specific emotional responses and changes that can entail the generation of expressive bodily motions. These changes are usually of long duration [93], thus the interest of employing these techniques during motion recordings sessions in which subjects are asked to perform different movements several times for distinct emotional states. However, since the elicited states are of medium intensity [93], MIP alone not always produce distinguishable realizations of expressive motions. Hence, when used in the construction of expressive motion corpora, they are usually coupled with the researcher's explicit instruction. In this manner, expressers are more likely to produce spontaneous yet recognizable expressive movements. Here we present the techniques the most widely known and used.

- **Velten MIP:** one of the first and most used techniques. Subjects are asked to read a number of statements (approx. 60) describing either positive or negative self-evaluations, somatic states or situations in which the intensity of the emotional state is constantly increasing [93, 262]. Subjects are then asked to try to feel the mood described by these statements.
- **Music MIP:** subjects listen to mood-suggestive musical pieces after being instructed to try to get into the emotional state called to their mind by the music. There exit two variants of this method: (a) the experimenter chooses, from some standardized options, which kind of music to use for the intended emotions and/or moods; (b) the subject chooses the piece of music that he or she finds to be the most suitable for eliciting the intended mood [262].
- **Films MIP:** it is one of the most efficient and simplest techniques. Subjects are presented with extracts from films to stimulate their imagination and facilitate their immersion on the intended mood. As for the music method, standard lists are employed to choose the best film extract for each intended mood. This technique is analogous to story-based MIP [262], in which films are replaced with some narrative or short story. In both techniques subject are asked to get involved and try to identify them-selves with the situation and the mood being suggested by the film extract or short story .
- **Imagination MIP:** also called autobiographical memory [93]. Subjects are asked to vividly recall and imagine situations from their lives that had evoked the desired emotion. In addition, they are usually asked to write down the imagined event and to

elaborate on the original perceptions, thoughts, feelings, sensations and affective reactions [93, 262].

- **Combined techniques**: with the aim of increasing the effectiveness of the mode induction, several researches have combined two or more of the aforementioned procedures. The general idea is to associate two complementary methods. The first method will induce the intended mood, while the second will create a convenient atmosphere for maintaining the intended mood for as long as possible [93]. Velten-Imagination MIP and Velten-Music MIP are some examples of these combinations.

One of the main concerns when using mood induction procedures is the nature of the changes observed in the expresser after the elicitation has taken place. Researchers frequently debate whether these changes are indeed produced by the induced mood or are instead motivated by the subconscious desire of the expresser to comply with the experiment and researcher demands [262]. Other important issues are the duration of the induced mood and the specificity of the intended mood. There is evidence suggesting that the use of mood induction procedures can give rise to emotional states other than the intended emotion. Hence, is it possible that the changes we observe in the expresser, in particular the expressive motions, have been amplified, attenuated or inhibited by other emotional states [93, 262]. Similarly, it was found that the duration of the induced mood is proportional to the type of induction procedure and to the intended emotion. In particular, it seems that the changes due to negative emotions last longer than those of positive states. Furthermore, several studies found that the effect due to the Velten technique tends to be less stable and lasting than other techniques such as films and imagination MIP [93].

In order to address these issues, most of the expressive motion corpora used for the analysis, recognition and synthesis of expressive motions have been recorded using trained lay or professional actors who knew the purpose of the recordings. Thus, the changes and expressive motions showed by an expresser are mostly due to the explicit instruction of the experimenter and their nature is not longer a problem. Induction mood procedures are used to facilitate the enacting of an emotion and to make the changes and motions associated to it more authentic and believable [14]. In the same manner, since the intended emotions have been precisely defined and framed, the experimenter can be almost sure that the changes observed in the expresser movements are certainly due to the target emotional state. Hence, the specificity of the intended emotion is not longer an issue. Finally, since actors are enacting emotional states rather than truly 'feeling' them, the duration of the induced mood is not a critical issue anymore. It is important to remark however, that if the experimenter aims to analyze expressive motions produced in a natural setting, all the concerns listed above will still apply.

3.2 Existing Databases

With gestures and movements being increasingly exploited in advanced interactive systems, the number of existing motion corpora, hereinafter called *Motion Capture (MoCap)* databases, has considerably grown in the last few years. Different MoCap databases have been designed and recorded with the purpose of studying plausible and believable human motor behaviors. In the context of this thesis, we have identified two main categories of MoCap

corpora: general purpose databases and expressive motion databases. Since this thesis is interested in the analysis and synthesis of expressive whole-body motions, we briefly survey the MoCap databases that belong to the first category before presenting a detailed discussion about those considered among the second category.

3.2.1 General Purpose Databases

Most of the MoCap databases in this category have been designed to analyze, classify or synthesize believable and high-quality human motions. They have been used in many different domains such as sport sciences, biometrics (action or person identification), and data-driven character animation. Among all the databases belonging to this category we can identify those publicly available and that are largely used by the academic research community: HDM05 [180] provided by the Max Planck Institute, CMU MoCap from Carnegie-Mellon University [42], Human Motion Database (HMD) created by the University of Texas at Arlington [102, 246], KUG [121] created by Korea University, KIT whole-body motion database [134] provided by Karlsruhe Institute of Technology, NUS MOCAP a data-driven character animation database [188] from National University of Singapore, among others.

These databases contain a wide range of human movements within different categories of themes, including locomotion, sport activities, and everyday life motions. Some of them include few style variations directly instructed to the actors, or induced from different emotional states. However, they were not built from a profound reflection on how to represent both the expressiveness in the movements induced by various factors (personality, emotional state, gender), and the meaningful expressions conveyed by body movements [58]. Moreover, they do not usually contain multiple repetitions of various movements with different affects for several subjects. These repetitions are necessary for a robust understanding of the patterns and expressive cues common to several expressers and different motions, and that should be reproduced when animating emotionally expressive characters.

3.2.2 Expressive Motion Databases

Recently an increased interest for expressive variations of body movements has led to the design and construction of several emotionally expressive movement databases. Many of them have been designed to study human perception of emotions in different contexts such as human communication [14, 104, 38], music and dance performances [41], narrative scenarios [254], daily actions such as knocking [197], etc. Others have been created to train automatic affect and emotion classifiers. These databases are usually characterized by a fixed number of individual actions, multiple repetitions by action and emotional state, and a large number of expressers [165, 130]. Similarly, the character animation domain has also created expressive databases to study the effect of emotionally expressive motion on game characters [75] and how the emotional content present in body motions is affected by the type of embodiment [172] and by different motion re-use techniques such as motion editing and motion blending [187].

Since all these databases have been designed with a particular functionality and modality (e.g., body, face, or both) in mind [58], many of them might not be suitable for the synthesis of expressive whole-body motions. For instance, FABO [104] and GEMEP [14] databases contain only upper-body motions recorded with 2D cameras, which make them not suitable

for animating 3D virtual characters and synthesizing whole-body motions. Other databases such as the Glagow corpus [165] and the dataset proposed by Volkova and colleagues [254] provide the 3-dimensional data necessary for animating virtual characters, but focus mainly on movements necessitating only the upper-body.

Seeing that our aim is to animate virtual characters, we need full-body movement recordings of high-quality such as the one obtained from maker-based motion capture systems and which involve full-body movements. Hence, we focus our review on expressive motion databases containing whole-body motions and that might be suitable for both the validation of the motion model proposed in this thesis and the generation of new expressive bodily motions. Table 3.1 lists all the databases fulfilling the aforementioned criteria. Each one of them is catalogued according to: number of expressers and samples, category of movements (see Section 3.1.2) and specific body actions, emotional states, number of samples, function, modality (e.g., body, face, speech, or multimodal), availability and labeling of emotional content.

Most of the MoCap databases listed in Table 3.1 have been developed to answer specific research questions. Hence, their design is based on methodological choices oriented and constrained by these questions [14]. For example, the UCLIC acted database [138] needed body motions conventionally associated to specific emotional states, i.e., *emotion portrayals*, in order to ensure recognition across different cultures. In the same manner, the expressive corpora designed by Aristidou *et al.* [4, 5] and later analyzed through Laban notation is mainly composed of dance motions, since the Laban notation system was specially developed for writing and analyzing the structure and expressivity of this kind of movements [147].

In this thesis, we aim to study, from a quantitative perspective, how body motions change and are modulated by the internal emotional state of the expresser. In this manner, we will be able to generate character body animations that convey the idea of more believable and emotionally rich virtual characters. This implies that MoCap databases in which emotions are expressed through specific gestures and/or movements, such as raising the arms as an indication of joy, are not suitable for our work (e.g., [75, 129, 187]. Similarly, we want to generate whole-body motions in which all body limbs are equally involved during movement, hence we need a database containing such type of body motions. Among the databases listed in Table 3.1, five of the them fulfill this requirement: UCLIC games [137], USC CreativeIT [174], Emilya [87], Karg *et al.*, Roether *et al.* [205], and Aristidou *et al.* [5]. However, only the two first are publicly available. UCLIC games [137] is not suitable because the body motions it contains are strongly influenced and constrained by the context in which they were recorded, i.e., Wii games. In the case of USC CreativeIT [174], further requirements need to be considered.

An additional, yet crucial, aspect when working with body expressions of emotional states it that the quality and intensity of those expressions may be influenced by numerous distinct factors such as the expresser's personality, gender, culture, age, and idiosyncrasy; the context, the mood induction procedure (if used), the body movement itself, etc. USC CreativeIT [174] database design is based on dyadic interactions. Hence, the expressive motions, and consequently emotional states, contained in this database might differ to those observed in a single expresser scenario. Namely, since USC CreativeIT [174] consists mostly of interactive communication scenarios, it is possible that many of the expressive bodily motions in this database were mainly generated to accompany speech and thus are strongly

Database	Expressers	Category	Function	Emotions/Affective States	Samples	Available	Data	Labeling
MPI [253]	8 amateur actors	Natural body expressions in narrative scenarios	Study expressive motions in a close to natural context	Amusement, joy, pride, relief, surprise, anger, disgust, fear, sadness, shame, neutral	1400	Yes	Body	Intended emotion
UCLIC Games [137]	11 non-actors	Natural body expressions. Actors played Wii games	Automatic recognition of emotional body expressions	Concentrating, frustrated, defeated, triumphant	36	Yes	Body	8 observers
UCLIC acted [138]	13 non-actors	Portrayed emotional expressions freely chosen	Study cross-cultural differences on emotion perception	Anger, sadness, fear, happiness	183	Yes	Body	Intended emotion
Emilya [87]	11 lay actors	Induced emotional expressions for daily actions: walking, sitting down, knocking, lifting, throwing, moving objects	Study emotion expression on non-emblematic movements (daily actions)	Joy, anger, panic, fear, anxiety, sadness, shame, pride, neutral	9031	No	Body	Intended emotion
USC CreativeIT [174]	19 professional actors	Induced emotional expressions. Improvisation performances based on two-sentences exercises and para-phrases	Study of human expressive behavior in dyadic interactions	PAD dimensional model. Performance ratings: interest, naturalness	59	Yes	Body Speech	15 annotators

Continued on next page

Database	Expressers	Category	Function	Emotions/Affective States	Samples	Available	Data	Labeling
Aristidou *et al.* [4, 5]	6 professional dancers	Induced emotional expressions of non-emblematic movements	Characterization and automatic recognition of affect using Laban notation	Afraid, angry, annoyed, excited, happy, pleased, satisfied, relaxed, tired, miserable, sad, bored	74	No	Body	Intended emotion
Roether *et al.* [205]	25 non-actors	Induced emotional expressions of non-emblematic movements (gait)	Identification of emotion-specific cues in human gait	Neutral, anger, fear, happiness, sadness	380	No	Body	Intended emotion
Kapur *et al.* [129]	5 non-actors	Portrayed emotional motion freely chosen	Automatic classification of emotionally expressive body motions	Sadness, joy, anger, fear	500	No	Body	Intended emotion
Karg *et al.* [130]	13 male non-actors	Portrayed emotional motions. Non-emblematic movements (gait)	Recognition of affect through gait patterns	Sadness, anger, happiness, neutral. Affective states along the PAD dimensions.	1300	No	Body	Intended emotion
Normoyle *et al.* [187]	1 professional actor	Portrayed emotional motions freely chosen	Study the impact of motion edition on emotionally expressive motions	Sadness, surprise, anger, disgust, fear, happiness	60	No	Body	Intended emotion
Ennis *et al.* [75]	8 professional actors	Portrayed emotional motions	Study which cues (body, face or both) are better indicators of emotion	Sadness, anger, fear, happiness	96	No	Body Face	Intended emotion

Table 3.1: Affective body movement databases. PAD stands for Pleasantness-Arousal-Dominance dimensional model of affect.

influenced by factors inherent to human-to-human interaction and interlocution. This kind of gestures and body expressions of emotions are out of the scope of this thesis. Similarly, MPI database [253] consists of body motion recorded in the context of narration scenarios. Hence, the body movements in this database are likely of co-verbal nature, i.e., body actions performed mostly by the hands and the arms and that are complementary to speech. Furthermore, since all expressers were seated on a stool during the recording sessions, the resulting body motions involve only the upper-body.

In order to ensure that the movement patterns we analyze and synthesize are indeed associated to the emotional state of the expresser, we need to standardize and control the expressive motions we consider [14, 58]. Furthermore, from a technical point of view, we need high-quality motion data with a well defined and known bio-mechanical model. This model will facilitate the purely kinematic approach we have adopted for the synthesis of whole-body motions. Finally, a big part of the work done in this thesis aims to prove and validate the suitability of end-effector trajectories for the characterization and generation of emotionally expressive whole-body motions. Hence, we need a MoCap database whose design and content facilitates the evaluation of this model. This new database should be large enough and include a significant number of different expressers – also referred to as subjects or actors –, emotional states and motor behaviors (movement classes). In this way, we can measure whether the proposed model is invariant to subjective (actors) and objective (movement class) factors. For all these reasons, we decided to develop our own expressive corpora. In the remainder of this chapter we introduce the design, methodology and evaluation of the MoCap database used in our work.

3.3 Design of a Theater-Based Motion Capture Database

As it was pointed out, none of the expressive MoCap database publicly available seems to be suitable for the type of motion model and application defined in this thesis; mostly because they were designed with a different purpose. Nonetheless, they are the product of a profound reflection and understanding of the issues that arise when studying how emotional states affect bodily expression. For this reason, several of the choices made during the design and development of our corpus are based on the work done during the definition and construction of the existing expressive corpora as well as on lessons provided by the computer animation community. Similarly, many other choices are underpinned by the methodological validation approach proposed in this thesis.

We propose a new motion capture database of expressive body motions in which all emotion-related content and any additional meaning are solely conveyed through body motion. The requirements considered in the design of our database are:

- several expressers (different ages, experience, gender),
- different motor behaviors, i.e., actions and motion sequences in which all body limbs are used,
- several repetitions for each expressive motion and emotional state,
- high-quality and high-dimensional data with few, close to none, artifacts,

- expressive variations rather than emotion portrayals,
- plausible yet recognizable expressive motions for each target emotional state,
- and a standardized and controlled capture protocol such that the perceived variations are due to emotional states rather than to other non-controlled factors.

Building a new motion capture database of expressive bodily motions requires a careful selection of the movements to be performed, the emotional states to be expressed, and the subjects to be recorded. Furthermore, it is crucial to define both the category of expressive motions we wish to record and the procedure we will use for gathering them, i.e., natural expressions, induction techniques or acted motions. Below we present and explain the design criteria we employed as well as the reasons and motivations behind it.

3.3.1 Why Create a Motion Corpora Inspired by Theater?

Identifying and producing *expressive bodily motions* can be a highly subjective and context-dependent process. Movements that are meaningful for an observer in a particular situation can also be considered as plain and with no specific emotion by a different observer. When looking for possible sources of *expressive motions*, the performing arts (theater, pantomime, dance, magic, mime, etc.) might be a good starting point, since their prime goal is to create visually believable characters capable of communicating meaning and emotion to an audience.

In his work on character animation, Neff [183] suggests also that "the performing arts literature offer arguably the most comprehensive analysis of expressive character movement available". He then shows how elements from specific fields of performing arts such as traditional animation, actor training (theater) and movement theory can be used to animate convincing virtual characters. Among these fields, we think that theater, and hence theatrical body movements, can be of interest and employed as a source of inspiration for both the synthesis of virtual character movements and the understanding of emotionally expressive human movements. The reasons behind this idea are threefold:

- First, in the creation of theater it is required to develop a deep understanding of "the language of gesture and movement" [151], because it is through movement that an actor transforms feelings, emotions, intentions, and passions into performance and meaning. By studying and analyzing theatrical body motions it is possible to gain a more practical insight into how meaning and emotions are encoded through movement.
- Second, in stage every movement is deliberately chosen and executed to arouse an emotions in the audience [183], and thus ensure that every character in scene to be perceived as believable. By using theatrical movements as the knowledge base of a character movement synthesis system, it is likely that a virtual character will also be thought of as believable.
- Finally, previous work on both character animation and the design of expressive motion corpora has shown that: (a) theater can be used as a model for the design and development of believable human characters [194, 174, 183], (b) valuable emotional

databases can be recorded from actors using theatrical techniques [76, 38], and (c) theater can provide some insight into how humans, specially actors, use their expressive behavior to convey the idea of believable and livable characters [174].

3.3.2 From Physical Theater to a Magician Scenario

Theater in its simplest definition is the branch of performing arts concerned with live, and most of the time, collaborative performances. Actors present reality-based or imaginary-based experiences and stories to an audience through a combination of gestures, movement, speech and other performing arts' elements such as dance, music and sound [191, 193, 233]. Among all the existing forms of theater, we mainly focus on *physical theater*.

Physical theater is a form of theater born from a tradition of mime and physical expression [152]. Contrary to the Stanislavsky system in which expressiveness is achieved once the actor has established the proper internal and psychological state [183], physical theater practitioners enact expressiveness through their body and its movement. Thus, the body and its movement are both the center of attention and the center of the theater making process [181]. Emotional states and any other affective phenomenon are channeled through the body rather than through any other modality like face or speech [39].

Physical theater accentuates the actor's body and its expressive capabilities. It is through these two elements that the audience will connect with the imaginative world developed in stage. Hence, physical theater seeks to understand movement, its laws and how the slightest nuance of physical movement affects meaning [39, 152]. As psychologist and emotion researchers have done during the last years, physical theater theorists (e.g., Decroux, Lecoq, Meyerhold, among others) have also addressed the relationship between body movement and emotions for years. Through their practical and empirical work, they have found that bodies reflect and project inner life and that emotions are linked to specific somatic patterns. Thus, they can be successfully conveyed through the right physical configuration [39, 183].

Our interest on this particular form of theater comes from the closeness between its principles and the objectives of this thesis. By exploiting some of the key elements and ideas behind the theory and practice of physical theater, we will be in measure of analyzing and generating good examples of expressive bodily motions. Furthermore, we will go a step forward towards the creation of more livable and believable virtual characters.

We have borrowed three key ideas from the training process on physical theater for the definition of the main content in our MoCap database. These three ideas are: *the expressive body*, *the corporal mime* and *the neutral mask*. In its original definition, the *neutral mask* allows to detach face and speech from the body. This latter emerges then as the only means of expression, i.e., the *expressive body*, and every movement becomes powerfully revealing [39]. Similarly, *corporal mime* makes reference to the art of bodily movements. The actor-mime is constrained by silence and can only make him-self understood through his body. The same body, *the expressive body*, creates the illusion of a living universe and allows the audience to see what is invisible to the eye: the hidden meaning and the inner-life of the character portrayed by the actor [152]. Furthermore, through the work of the *neutral mask* and *corporal mime*, physical theater explores how alterations of postures and motion can provide notions of an emotional state to the external observer. The definitions and applications we present here come from our understanding and interpretation as gathered from the reading of three major books on the theory and practice of physical theater: Dymphan Callery's book *Through*

the Body: a Practical Guide to Physical Theatre [39], *Theatre of Movement and Gesture* [152] and *The Moving Body: Teaching Creative Theatre* [151] by Jacques Lecoq.

Based on these principles, we developed a theatrical scenario in which each expresser, hereinafter called actor, will perform as a mime-magician. That is, based on a combination of the *neutral mask* and *corporal mime* principles, we will ask each actor to only use his/her body movements when conveying ideas, emotional states, meanings, etc., to an imaginary audience. The magician context was inspired by the *Cirque du Soleil* spectacle's Kooza and comes from the idea that a magician is an artist of misdirection. During a magician performance, the entire body is used to both engage, captivate and at the same time mislead the senses of the spectator [125], [167]. Therefore, we consider it an interesting trial case for the kind of expressive movements we are interested in and for applying physical theater ideas.

The mime-magician scenario consists of three common and known magic tricks: *the disappearing box*, *pulling a rabbit from a hat*, and *taking scarves from an empty jacket*. Each magic trick involves three stages:

- **Introduction,** the magician makes his appearance and introduces himself to the public.
- **Preparation,** the magician shows to the public each object he is going to use. A box and a scarf for the first trick, a top hat for the second, and a jacket for the last one. This stage ends when the magician *invokes his magical powers* with his wand.
- **Results and bowing,** the magician shows the result of his trick and makes a bow to the audience.

Since the whole scenario was conceived to start in any random order, the body motions of the first and last stages are common to all magic tricks. In the second stage, the mime-magician uses his body to convey the physical properties of the objects needed for his magic trick. These movements, also referred as to actions, are different among all three magic tricks. The proposed scenario has in total three sequences that can be decomposed into 17 individual semantic actions.

The interest of working with three different sequences which have few common motions between them is two-fold. Since all actors will perform the same scenario for all target emotional states, we can use them to measure how well the end-effector trajectories motion model registers the patterns of expressive cues that are common to several actors when enacting certain emotions. Cowie *et al.* [58] argued that expressive cues used to characterize the motion patterns related with the expression of emotions are also affected by the type of motion executed by an expresser. Thus, by considering movements not common to all sequences, we can get some insight into the suitability of the proposed motion model for the generation of distinct expressive whole-body motions.

3.3.3 Expressed Emotions and Actors

In general, as showed in Table 3.1, the set of emotions included in the existing expressive motion corpora corresponds to the basic emotions proposed by Ekman [71]. However, from a practical point of view, some of these emotional states, e.g., surprise, disgust or fear, are not suitable for the type of content and scenario we have chosen. We decided to select a subset of emotional states based on the circumplex model of affect proposed by Russell [207]

instead. Namely, four emotional states: *happiness*, *sadness*, *stress*, and *relaxedness* were selected based on their activation and valence levels. The goals motivating this choice are: (a) a more detailed evaluation of the similarities and contrasts between the kinematic patterns exhibited by the chosen categories, and (b) a more precise measure of the suitability of the proposed motion model for different emotional states, e.g., we can identify if end-effector trajectories characterize emotional states opposed on activation and arousal equally well for example. We also included a fifth state, *neutral*, so as to define a baseline expression. This state is associated to the performances in which no emotion was intended and represents the intersection point between arousal and valence dimensions.

With a scenario and a set of emotional states already defined, the next step is to determine what kind of expressers should be recorded. There are two main categories: skilled actors and untrained expressers. The choice between them depends on both the procedure used for gathering the expressive bodily motions and the content to be recorded [14]. Since our scenario is based on physical theater and requires expressers to use only their bodies to convey emotional states and meaning, we require acting skills and a good awareness and control of the body. Hence, we have narrowed the set of possible expressers to skilled and experienced actors only.

One of the main concerns when working with skilled actors is that of the naturalness and spontaneity of the emotionally expressive motions. Researchers fear that acted motions will be far from natural and thus the observed patterns related to the expression of emotion will not be of use for real-life applications. However, working with acted motions brings numerous advantages such as the number of repetitions and variations that can be recorded for each emotional state, the quality of the recordings, the control of extraneous factors, and the systematic study of movement-related elements contributing to the expression of emotions [14, 87]. Previous studies on the design of expressive corpora [38, 15] found that the combination of acting skills and mood induction procedures will induce, at least to a certain extent, real affect in the actor and facilitate the enacting of emotions. Consequently, the resulting expressive motions are likely to show significant variations closer to naturalistic data rather than exaggerated prototypical patterns. Hence, by using induction mood procedures it is possible to achieve a compromise between data naturalness and the advantages of acted data [58]. For all these reasons, we decided to use both skilled actors and induction mood procedures during the recording of our MoCap database.

3.4 Building the Corpus of Expressive Theatrical Motions: First Capture

To produce a high-quality motion capture database that can be used for the analysis and synthesis of expressive body motions requires of a carefully designed capture protocol. In this section we review the controlled environment in which the data was gathered as well as how the recording sessions and elicitation procedure took place.

3.4.1 Technical Setup

The understandability and expressiveness of whole-body motions require accuracy and high definition in the recording of captured motion. A Qualisys [200] motion capture system com-

posed of 8 Oqus400 cameras was used. An additional video camera was placed such as to record the actor's upper-body. Both the MoCap data and the video recordings were automatically synchronized by the capture software. All full-body actions and hand movements inside a 2.5m×2m×2m volume were recorded. A total of 64 passive markers were placed on the body of the actor including 5 markers on each hand and 2 facial markers. The markers on the hands enabled capturing all the grasping movements involved in a magic performance, and the facial markers gave a more accurate idea of the direction of the head of the actor. We used a 200Hz capture frequency to correctly capture hand motion, since this kind of motion requires a higher accuracy.

3.4.2 Number of Actors and Repetitions

Since the suitability of end-effector trajectories is going to be statistically estimated through automatic classification, numerous repetitions of each sequence performed by several subjects are needed. Each magic trick was recorded three times per emotional state. In addition, the most representative actions (8 in total) were selected among the initial 17 (see second column in Table 3.3). For each selected action, 2 sequences of 5 repetitions per emotional state were recorded. Actors did not receive instructions on how to express the emotions so as to avoid prototypes. They were however asked to use only body movements, to always remain inside the capture volume, and to respect the order of actions within a magic trick sequence. In this first capture session, two skilled amateur actors, a woman with dance experience and a man trained in theatrical improvisation were recorded.

We believe that theatrical motions are not only interesting for conveying emotional states, but are also perceptually perceived as spatially and temporally richer than the human movements found in the existing motion capture databases. In order to validate this hypothesis we asked our two actors to perform eight additional bodily actions in a neutral emotional state. These motions were selected after surveying the most recurrent actions among all the databases surveyed in Section 3.2. Three repetitions for each selected action were recorded. As result we obtained 123 motion capture files for each actor. Specifically we have:

Emotion	**Sequence/Action Type**			
	Magician sequences	**Magician actions**	**Daily actions**	**Subtotal**
Happiness	18	160	0	178
Neutral	18	160	48	226
Relaxedness	18	160	0	178
Sadness	18	160	0	178
Stress	18	160	0	178
Subtotal	90	800	48	938

Table 3.2: Count of motion sequences and actions in first recording session across emotion categories (both actors included).

3.4.3 Emotion Elicitation and Recording Procedure

Each recording session took place as follows: first, a video of each magic trick was presented to the actors the day before the capture. This made possible for the actors to learn the actions, perform more fluently and show less hesitation between actions. Second, on the day of the capture, the actors were asked to perform each magic trick several times before we started to record. By doing so, we could correct all possible doubts about how each gesture should be performed. Third, an emotional state was randomly chosen and the emotion elicitation was done using an imagination mood induction procedure. During the elicitation process, each actor was instructed to remember an emotional event in their lives that corresponded to the selected emotion. Actors were then asked to perform the whole scenario, i.e., the 3 sequences plus the 8 individual actions for the elicited state. After all recordings for a given emotion were captured, a debriefing was done to re-establish the initial emotional state of the actor. Lastly, actors were asked to perform the common actions in a neutral state.

3.5 A First Evaluation

Three perceptual experiments were performed to validate the suitability of the chosen scenario, and the effectiveness and efficiency of the experimental motion capture protocol described above. This first perceptual evaluation gave us an initial idea about the usability of the produced MoCap data for the analysis and synthesis of expressive body motions. We were aiming to answer the following questions:

1. Do observers perceive theatrical actions as being more kinematically significant than more common actions? Do people perceive theatrical movements as motions conveying more information?

2. Can observers associate the spatio-temporal variations introduced through the elicitation of emotional states to one of the five selected emotions? If they can do so, how expressive do they find the theatrical motions?

3.5.1 Stimuli Creation

For the theatrical scenario, one realization of the eight most significant individual actions and of two magic tricks were chosen per emotion and per actor. We selected also one realization for each daily action. Table 3.3 lists the selected stimuli.

All stimuli were played on a stick-like character representation (see Figure 3.1). We chose this kind of representation as we did not want to convey any additional information about the avatar's gender and appearance that might influence the categorization of the selected emotions. Additionally, previous studies have shown that this type of representation does not stop observers from perceiving any emotional state at any intensity [7, 172].

For the theatrical and common actions, individual video clips of the same duration (10s) were created at 25Hz. For the magic trick sequences videos of 42s were produced. The character was displayed in the center of the screen, facing forward at the beginning of each clip. Video clips were presented at 1280x1024 resolution. A total of 116 videos were generated.

Daily actions	Theatrical actions	Sequences
Lifting	Show empty jacket	The disappearing box
Waving	Take scarves out of jacket	Taking scarves from an empty jacket
Kicking	Invoke magic with wand	
Hand shake	Show box disappeared	
Walking	Cover box	
Knocking	Invoke magic with hand	
Throwing	Introduction bow	
Punching	Final bow	

Table 3.3: Motion stimuli used for the first perceptual study

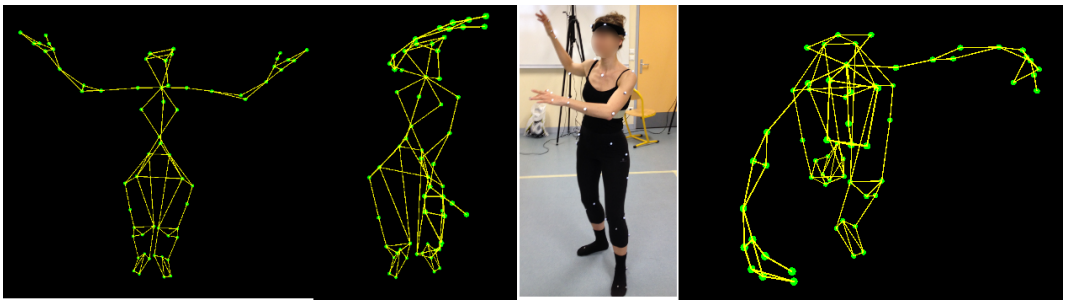


Figure 3.1: Stick-like character representation, marker set and posture examples

3.5.2 Participants and Duration of Each Study

Twenty participants took part in the studies we will detail in the remaining of this section, a total of 100 different individuals contributed to our experiments. Participants came from various educational backgrounds and were all naive to the purpose of the experiment. They only knew they would watch some avatar videos and answer a few questions about what they perceived from those videos. Detailed information about the gender and age distribution of each group of participants and the duration of each study are presented in Table 3.4.

Study	Gender	Mean age	Duration
Daily actions vs theatrical motions	11M, 9F	24.0+10.0	15
Individual theater actions (emotions male actor)	10M, 10F	23.5+6.0	40
Individual theater actions (emotions female actor)	15M, 5F	23+7.0	40
Magician sequences (emotions male actor)	13M, 7F	21.6+7.5	15
Magician sequences (emotions female actor)	13M, 7F	25.0+13.0	15

Table 3.4: Information about each study's participants and duration in minutes

3.5.3 First Experiment: Everyday Life Movements vs. Skilled Theatrical Movements

In our first experiment we wished to determine whether observers perceived theatrical movements as kinematically richer than more common bodily actions. Additionally, we wished to investigate whether participants regarded theatrical actions as motions conveying more information compared to common actions.

For this study, we presented participants with 32 video clips of 10s duration, depicting 8 daily actions and 8 theatrical movements for each actor. Participants viewed each video clip in a random order and could play it as many times as they wished. They were asked to rate on a scale of 1-7 whether the performed action was considered as current, spontaneous and habitual (1 on the scale) or as skilled, meaningful and elaborated (7 on the scale).

Since the answers of the participants were nominal variables, we did not think the data fits the assumptions of an ANOVA. Results for this study were analyzed using Kruskal-Wallis one-way of variance and paired T-Tests for all post-hoc analyses. We found that the gender of the participants and actors had no effect on the ratings of daily and theatrical actions. A significant difference ($H = 158.5377, 1d.f, p < 0.001$) between the mean rank scores of the two types of actions was found. As we confirmed a significant divergence between the two categories of gestures, we were then interested in identifying which particular motions were considered more kinematically significant and conveying more information. The results of the Kruskal-Wallis test ($H = 270.15, 15d.f, p < 0.001$) were significant; the mean ranks scores of 7 of our 8 theatrical gestures were significantly different among the 16 different movements presented to the participants. For common daily actions, we found that *kicking* and *punching* gestures were perceived as the most kinematically significant actions among the everyday motions. A possible reason for this could be that both actions are considered more sportive actions than everyday motions, thus a higher kinematic variance can be attributed to them. Mean rank scores for both categories and for the 16 gestures are shown in Figure 2.

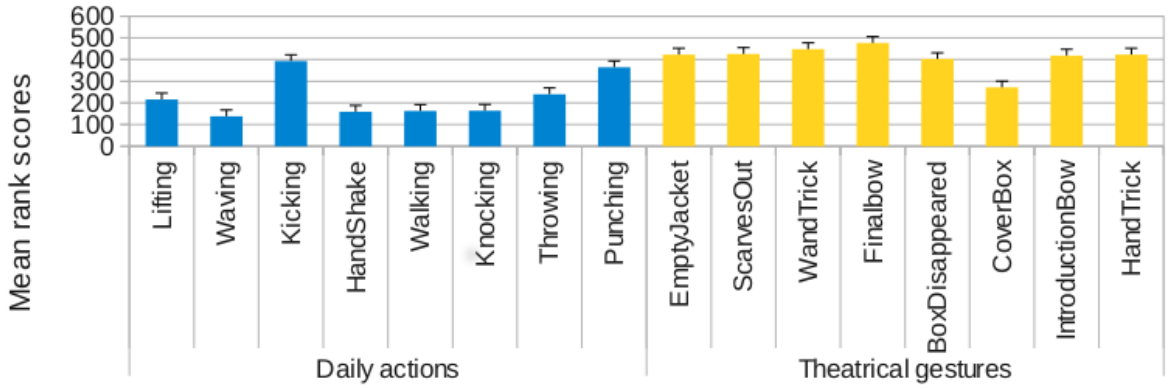


Figure 3.2: Mean ranks scores for each category and each one of the sixteen presented actions

3.5.4 Second and Third Experiment: Perception of Emotion in Isolated Actions and Sequences

In this study, we take into account the fact that emotional states might be expressed differently by each subject and that such states might be more easily recognized over longer stimuli. For this reason, we used 4 separated groups. Two groups rated the emotions of our male character and female character over the theater individual actions, and the 2 other groups did the same over the sequences.

We wished to determine whether the 5 emotions portrayed in our theatrical motions could be recognized among a 6 non-forced-choice list of emotions (the 5 emotions already listed plus the *other* option). Additionally, we wished to investigate the intensity with which each emotion was perceived.

For the individual actions, we presented participants with 8 video clips of 10s, representing our 8 theatrical actions (see. Table 3.3 for a detailed list) in each one of the 5 emotional states, where each actions was presented twice. Participants viewed each video clip in a random order as many times as they wished and were asked to choose an emotion among the 6 possible options (also randomly presented). They were also asked to rate the intensity of the selected emotion on a scale from 1 (not intense) to 7 (very intense)

For the magic trick sequences, we followed the same methodology applied in the evaluation of individual actions. However instead of using short videos of a unique action, we presented participants with a whole realization of a magic trick. For this study only the recordings of *the disappearing box* and *taking scarves from an empty jacket* were considered.

Results for these studies were analyzed using standard analysis of variance (ANOVA) and paired T-Tests for all post-hoc analysis. As done in [75], we calculated and analyzed the accuracy rate for emotion, i.e., how many stimuli were correctly recognized for each intended emotion by participant. We found no effect of participants and actors gender on the accuracy of emotion identification.

For the individual actions experiment we found a main effect of emotion ($F = 18.68, 4d.f, p < 0.001$). Post-hoc tests showed that the 5 emotional states were recognized with means ranging from 29% to 64%. The most accurately identified emotions were *stress* and *sadness*. No main effect of actor gender and type of action were found. However, an interaction between these 2 factors was shown as significant ($F = 2.81, 7d.f, p < 0.007$). This interaction might be due to both actors having different acting qualities for each type of emotion and action.

For the sequences experiment we also found a main effect of emotion ($F = 6.04, 4d.f, p < 0.001$). Post-hoc tests showed that the 5 emotional states were recognized with means ranging from 40% to 70%. Contrary to the individual actions experiment, in this study participants were more accurate in emotion categorization. This could be explained by the length of the stimuli presented to participants. We found that the most accurately identified emotional states were *stress* and *sadness*, followed by *relaxedness*, *happiness* and *neutral*. As for the individual actions, no main effect for actor gender and action type were found. However, an interaction between the emotion and actor factors was again observed ($F = 4.75, 4d.f, p < 0.001$). We believe this interaction might be due to the acting qualities of our two actors.

To have a better insight of where the miscategorizations happened, the confusion ma-

Isolated gestures						
Correct answer	**Relaxedness**	**Happiness**	**Neutral**	**Sadness**	**Stress**	**Other**
Relaxedness	**29.22%**	13.44%	28.13%	14.06%	5%	10.16%
Happiness	16.41%	**35.31%**	16.72%	2.03%	16.09%	13.44%
Neutral	17.03%	18.44%	**38.91%**	5%	10.16%	10.47%
Sadness	8.44%	1.56%	15.63%	**49.84%**	8.91%	15.63%
Stress	2.81%	6.72%	5.47%	2.34%	**64.69%**	17.97%
Sequences						
Correct answer	**Relaxedness**	**Happiness**	**Neutral**	**Sadness**	**Stress**	**Other**
Relaxedness	**50.63%**	10%	13.13%	11.88%	4.38%	10%
Happiness	11.88%	**52.50%**	13.75%	1.25%	13.75%	6.88%
Neutral	20%	19.38%	**40%**	3.13%	7.5%	10%
Sadness	11.88%	1.88%	13.75%	**53.13%**	1.88%	17.50%
Stress	8.13%	7.50%	6.88%	2.50%	**70.63%**	4.38%

Table 3.5: Confusion matrices for first perceptual study.

trices of both studies is shown in Table 3. For both studies, the accuracy rate of the participants is above chance (20%) and the results obtained for *neutral* and *sadness* emotional states are consistent with previous works [75, 271]. Additionally, *stress*, an emotional state that is not frequently used, has the highest recognition rate. However, the *happiness* and *relaxedness* emotional states were frequently misclassified between them or with the *neutral* state. We think possible reasons for this could be the proximity of these 3 emotional states in the circumplex model [207], the utilization of actions whose sole purpose is not to convey emotional cues, and confusions in the actors interpretation of the target emotional state. Furthermore, as we are using motions that are already spatially and temporally rich, we think it is also possible that the variations added by those 3 emotional states were perceived by the participants as indistinct.

To identify possible significant differences in the intensity of the emotional states, Kruskal-Willis tests were applied. We found no difference between the emotional intensities portrayed for both actors and for each action or sequence. For both studies, the emotions rated as the most intense are those the most accurately categorized ($H = 30.82, 4d.f, p < 0.001$ for the individual actions and $H = 14.09, 4d.f, p < 0.001$ for the sequences).

3.5.5 Discussion

We have proposed a new motion capture corpus composed of 17 theater-inspired actions, in the context of a mime-magician performance, into which emotional variations were added. Three perceptual studies were conducted in order to validate the suitability and usability of this new MoCap database as well as the relevance of the capture protocol. We found that theatrical bodily actions can be globally considered more skilled, meaningful and elaborated movements compared to more common daily actions. We also established that for the selected theatrical scenario, emotional states can be successfully elicited in the laboratory setting, and most importantly that our recognition results are significantly close to those of previous studies in which archetype emotion portrayals and more complete visual clues were used [172]. However, this first perceptual study also points out some shortcomings on the recording protocol and the definition and elicitation of the intended emotional states.

First, when comparing the recognition rates between individual actions and sequences we find that observers were in general less accurate when presented with short individual actions. They also showed much more uncertainty if we consider the percentage of stimuli rated as belonging to other emotional states. Discussions with the observers posterior to the study pointed us to the fact that isolating actions from a scenario defined in term of long sequences removes important context information. It seems that since the observers could not place the meaning of the individual actions, they found harder to perceive the emotional states conveyed through them. Additionally, since actors were asked to make pauses between each repetition of an individual action, it is possible that they were not able to sustain and enact the same emotional state with the same fluidity and intensity as they did for the magician sequences.

Second, a deeper analysis of the misclassification rates showed that, independently of the type of stimuli (e.g., actions or sequences), raters had more difficulties perceiving the nuances differentiating the *happiness, relaxedness* and *neutral* states between them. We observe also that *sadness*, an emotional state that is frequently reported as one of the best recognized from whole-body stimuli [7, 197], was often labeled as depicting other, non-listed, emotional state. Since actors were only asked to remember a past-experience in which they felt sad or happy for example, it is possible that some ambiguities on the interpretation of the intended emotional state appeared. It is likely that by better describing and contextualizing each intended emotion, through short scenarios for example, actors will be able to enact the kind of nuances we wish to study.

Finally, this first evaluation showed that theater-inspired motions seem to be perceived differently and as kinematically richer than other more common body motions. However, we still do not know if emotions are conveyed more clearly through this kind of motions. Depictions of emotional states thorough more common actions such as walking are needed in order to assess if there is any difference on the perception of emotions from theatrical motions. Additionally, since we have only recorded two actors, it is not possible to determine the influence in the perception of the expressiveness of theatrical motions of factors such as acting skills, gender and personal style. For all these reasons, we decided to carry a second set of recordings in which the number of actors was increased and more common bodily motions were added.

3.6 Extending the Corpus with New Actors and Sequences

To validate the motion model proposed in this thesis through automatic classification of affect supposes to have a sufficiently large learning dataset at disposition. This dataset should not only provide observations of the type of phenomenon we expect the classifier to discriminate, but also contain a wide range of non-emotion related conditions in which the generalization of the end-effector trajectories can be easily tested. For instance, we are interested in observing how this motion parameterization behaves with elements that might influence the expressive content of body motions such as different types of actions/motor behaviors and subject-dependent properties, including identity, gender and expressive capabilities. Taking this into consideration and in the light of the aforementioned shortcomings, we introduced two main changes in our recording protocol:

1. In addition to the three sequences already defined in the magician scenario, we de-

cided to include one walk motion at least one minute long and a short improvisation sketch freely chosen by each actor. The former will allow us to compare our results to previous work, since locomotion, in particular walking, has been extensively studied in both perception and automatic recognition of affect/emotions. The latter sequence will provide us with measurements about the end-effector trajectories behavior on considerably different movement behaviors. Combined, we can use them to evaluate the expressiveness of the actors and the benefits of using motions inspired by theatrical scenarios.

2. Five additional actors (two females and three males, ranging in age from 38 to 54) were asked to perform the five motion sequences already defined. Once again, actors were asked to convey both emotions and meaning solely through body motions. To facilitate the enacting of all emotions during the sequences, a combined induced mood procedure was defined. We used a story-based and imagination-based mood induction procedures. Five short stories related to the magician scenario were created (listed in Appendix A). The stories were used to better contextualize and reduce ambiguity on actors understanding of each intended emotion. The intent behind the imagination mood induction procedure was to facilitate the enacting of the target emotion all along the scenario, i.e., the magician sequences, the walk motion and the improvisation sketch.

Each actor came separately for a unique motion capture session. All actor's motions were recorded at a rate of 200 frames per second with a Qualysis motion capture system [200] consisting of nine cameras. Actors wore a manually defined marker-set containing a total of seventy-five passive markers. Each motion capture session took approximately four hours. For each actor, we recorded three repetitions of each magician sequence, one example of the walking motion, and one repetition of the improvisation sketch for each emotional state. We recorded a total of 275 motion sequences, with 55 motions for each actor and for each emotional state. For a more detailed description of the motion capture database content see Table 3.6. The duration of motion sequence is variable and depends on the type of sequence, the expressed motion and the actor. Motion sequences have an average duration of 35 ± 12 seconds. After the collection of the motion capture data, all gaps in the marker trajectories – caused either by occlusion or by going pass the capture volume – were filled using a combination of the reconstruction procedure described in [182] and classic spatio-temporal interpolation.

3.7 Second User Study

Previous studies confirm that an average human can recognize the affective content – more precisely the emotional states – conveyed by someone else's body motion with high accuracy, even when impoverished stimuli are used. However, most of these studies have focused on gait motion, dance and/or musical performances, portrayed representations, co-verbal gestures, and everyday actions.

There are noticeable differences between the nature of the body motions we present in this thesis and those that are commonly used in the computer animation and affective computing fields. Since expressive bodily motions are highly subjective, it is necessary to quali-

Emotion	Sequence Type			
	Magician scenario	Walk motion	Improvisation	Subtotal
Happy	45	5	5	55
Neutral	45	5	5	55
Relaxed	45	5	5	55
Sad	45	5	5	55
Stressed	45	5	5	55
Subtotal	225	25	25	275

Table 3.6: Count of motion sequences in database across emotion categories and type of sequence

tatively validate whether the extended version of our database: (a) contains actual examples of expressive bodily movements, and (b) can be used for investigating and building motion models capable of generating equally expressive bodily motions. For this purpose, we designed and conducted a second user study that not only determined the expressiveness of the sequences in the databases, but also provided us with a basis for accuracy comparison against automatic affect classification models. Furthermore, the aforementioned perceptual evaluation was also used to qualitatively measure the impact of displaying only the information carried by end-effectors and pelvis trajectories on the perception of the intended emotions.

3.7.1 Motion Sequence Representation and Stimuli

Since the movement of the body joints recurrently employed when recording whole-body actions can not be directly measured, each posture in our database is initially represented as a vector $\mathbf{V} \in \mathbb{R}^{3\times m}$ of 3D Cartesian positions, with $m = 72$ being the number of passive markers used during the MoCap recording sessions. Nonetheless, since at least two passive markers were placed near each one of the body joints of interest, it is possible to obtain a simpler and more concise representation from each posture $\mathbf{V}$.

Accordingly, we now describe a motion, $\mathbf{M} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_T)$, as a sequence of T postures, where $\mathbf{X}_t \in \mathbb{R}^{3\times n}$ with $1 \leq t \leq T$ corresponds to the simplified skeletal representation computed from $\mathbf{V}$. This skeletal representation consists of $n = 27$ body joints whose 3D Cartesian positions are approximated by the average position across all markers placed at immediate proximity of the target joints. For example, the position of the knee joint corresponds to the average of the two passive markers placed at each side of the Tibia's proximal end. Figure 3.3 shows both representations, the original and the resulting posture, $\mathbf{V}$ and $\mathbf{X}$.

One of the main contribution of this thesis is a motion model for the synthesis of expressive bodily motions that is based only on a subset of manually selected body joints, i.e., end-effectors and pelvis trajectories. In order to perceptually evaluate the proposed model, two possible body-representation stimuli were defined:

- A full-body stimuli based on skeletal representation $\mathbf{X}$ (see image on the right in Figure 3.3), hereinafter called whole body representation.

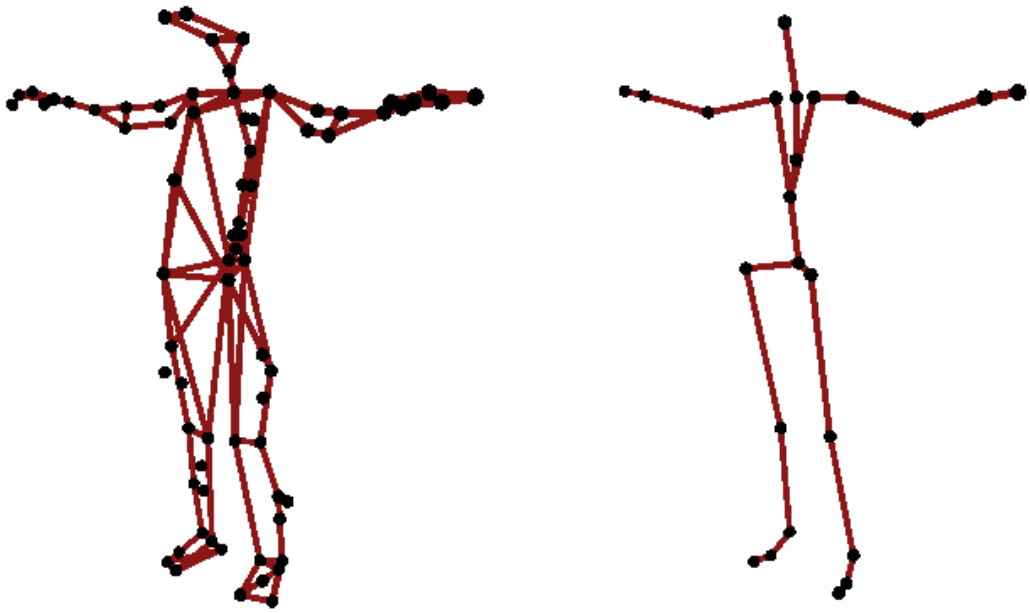


Figure 3.3: Body representations. Left image corresponds to marker-set used during the second motion capture sessions (called posture **V**). Right image corresponds to the skeletal representation (called posture **X**) computed from **V**.

- A partial-body stimuli $\mathbf{Z} \in \mathbb{R}^{3\times 6}$, with $\mathbf{Z} \subset \mathbf{X}$, also referred to as end-effectors representation, in which only the trajectories of a subset of six body joints consisting of head, hands, feet and pelvis trajectories were shown (see Figure 3.4).

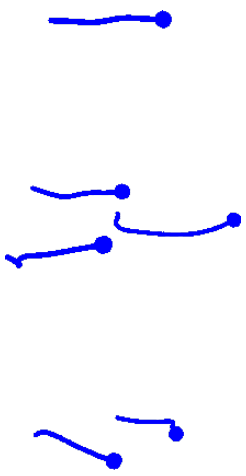


Figure 3.4: Partial body representation **Z**. Only the 3-dimensional trajectories of head, hands, feet, and pelvis joint are considered.

3.7.2 User Study Design

To correctly assess the relationship between body representation and participants' accuracy rates, both stimuli were evaluated separately. More precisely, participants were randomly assigned to either the whole-body or the partial-body condition. In this way, we discarded any possible carry-over effect between representations and guaranteed that participants would remain naive to one of the main purposes of the study.

Although participants would rate motion sequences for only one body-representation stimuli, watching and evaluating all motion sequences in the database would have required at minimum two and a half hours (275 videos with an average duration of 35 s.) non-stop. Hence, to avoid the effects of boredom and fatigue in participants, the number of sequences to evaluate was reduced to 125. Only one repetition for each one of the 5 sequence examples in the database was considered by actor and emotion (5 sequences $\times$ 5 emotions $\times$ 5 actors = 125). The selected sequences were then randomly assigned to one of five groups. The assignment was done so that each participant rated 25 sequences in total and all emotions were depicted in each group. Since participants' accuracy rates would be later compared to those reported by an affect automatic classifier, we decided to add two additional constraints to the random assignment of the selected sequences into groups. Each group would include at least two different sequences and two actors. In this way, the conditions under which the classifier and participants accuracies would be compared were, to some extent, analogous. With these additional constraints each participant rated 5 sequences for each emotional state. Table 3.7 shows a detailed description of the 5 groups.

Video clips at 30 fps were created for each selected sequences; one for each body representation. For each video clip, the character representation was placed at the center of a 3D virtual space and facing the virtual camera at approximately 45°. For the partial-body stimuli, we displayed the 3D position of the selected joints along with the trace of their respective trajectories as shown in Figure 3.4. Although technically we defined 5 groups, we had 10 groups in total. For each group listed in Table 3.7, we created two further subgroups: one containing the full-body stimuli only and another for the partial-body representation.

Group	Actors and Sequences	Emotions	Total trials
A	GR: 1, 3, 4 GA: 1, 3	All	25
B	GR: 2, 5 GA: 2, 4, 5	All	25
C	PP: 1, 2 SLM: 1, 2, 5	All	25
D	PP: 3, 4, 5 SLM: 3, 4	All	25
E	LC: 1-5	All	25

Table 3.7: Groups used during perceptual evaluation. Actors and sequences were randomly assigned to each group. Sequences and actors are identified as follow: magician sequences (1-3), walking motion (4), improvisation (5), females actors (GR, SLM), and male actors (GA, LC, PP).

A total of 200 participants, 98 women and 102 men, ranging in age from 21 and 75, were recruited through Amazon Mechanical Turk (MTuk) service [55]. They were randomly assigned to one of the ten groups. The same participant could no rate the sequences of more than one group. We had 20 participants by group.

Before starting the experiment, participants were presented with a general description of the task they would perform. They were informed that each video clip depicted the body movement performed by an actor under different emotional states. They were also told that for each video clip they would be asked to answer four questions about the expressive content and the motion qualities they perceived in the actor's movement. The questions they were asked to answer for each video clip are:

1. *"Which of the five (5) listed emotions do you think is conveyed through the motion?"* A forced-choice question was used since we were only interested in assessing whether the sequences in the database effectively conveyed the target emotions.
2. *"How deeply does the motion express the emotion?"* Participants were asked to rate the intensity of the emotion on a scale from 1 (very lightly) to 7 (very deeply).
3. *"How do you qualify the emotion conveyed in the video?"* Participants were asked to rate the arousal and valence components of the emotion perceived for each movement on a scale from 1 (lethargic/negative) to 7 (energetic/positive).
4. *"How difficult was to rate this motion?"* Using a 7-point scale, participants rated how difficult it had been to estimate both the emotion and motion qualities from the observed sequence. This question provided an additional measure of the expressive quality of the motions in the database as well as of the information that might have been lost with the partial-body stimuli.

The user study consisted of two stages. First, during the training stage, participants were presented with one video clip, different from those being later evaluated, for each emotional state. The aim of this stage was to allow the participants to familiarize with the type of representation they would rate as well as with the questions and the course of the exercise. Second, during the testing stage, participants were presented with the 25 video clips they would rate. For both stages, the video clips were presented in random order. Participants could watch each video clip as many times as they wished. However, once they moved forward to the next video clip, they could not longer go back and change their answers. The exercise took in average between 20 and 40 minutes.

3.7.3 Improving Data Quality Through Outliers Detection

One of the main concerns when using crowd-sourcing services such as Amazon Mechanical Turk is how to ensure the quality of the answers submitted by the participants. Since the participants are not longer in a controlled environment and within the reach of the experimenter, it becomes harder to ensure that the participants understand the task they are asked to accomplish and that they do it with the care, diligence and seriousness expected by the experimenter [265].

Several strategies have been adopted in order to evaluate the quality of the answers submitted by MTurk workers as well as to guarantee, to some extent, that all participants follow

and perform the study as the experimenter expect them to do. The most common and popular strategy are catch trails. That is, questions with obvious answers are presented at specific points during the perceptual study. Since they are simple questions, any participant should be able to answer them correctly. If the participants fail to answer them, we can assume they were not paying attention to the questions or answering them randomly, thus we can discard their answers from the study's results [190]. However, a recent study showed that catch trials questions can influence participants answers and change the manner in which they approach the task [111]. Hauser *et al.* [111] found that when exposed to catch trials, participants change their response behavior to a more systematic thinking approach in order to not be tricked again. Since the aim of this user study was to find how well the actors' motions conveyed the elicited emotional states, rather than how well participants could answer our questions, we decided to favor spontaneous answers. Hence, we adopted a different approach in order to evaluate and ensure the quality of the answers used later on our analysis.

Feng and colleagues [81] suggested to use inter-rater agreement measures to automatically detect outlier participants and improve the general quality of the collected answers. Once all agreement coefficients have been computed, the authors proposed to identify which participants are outliers through the analysis of the empirical distribution of the agreement scores. We adopted a modified version of this approach. More precisely, outliers are identified using the Tukey's method [243]. That is, all participants whose agreement score lie outside a determined interval are considered as outliers. The exact outlier detection algorithm is presented below:

Data: A_g: Answers of participants in group g, P_g: List of participants in group g
Result: O_g, set of outliers in group g
for $v_p \in A_g$ **do** // Answers given by participant p
    Define V_g as the majority vote answers of $A_g \backslash \{p\}$;
    Compute and store $kappa_p(v_p, V_g)$ // Cohen's kappa coefficient between v_p and V_g ;
end
Compute first, Q_1, and third, Q_3, quantiles for all kappa coefficients of group g;
/* Tukey's method */
Define accepted interval $I = [(Q_1) - k \times (Q_3 - Q1), (Q_3) + k \times (Q_3 - Q1)], k = 1.5$;
Define $O_g = \{\}$ as set of outliers in group g;
for $p \in P_g$ **do** // Each participant p in P_g
    if $kappa_p$ *not in* I **then**
        /* Add p to O */
        $O = O \cup \{p\}$;
    end
end

Algorithm 1: Algorithm used to detect outliers among all participants in user study.

The outlier detection algorithm was applied to each one of the ten groups (see Table 3.7) used during the perceptual evaluation. Among 200 participants who took part of the study, a total of 10 were estimated as outliers by Algorithm 1; one for each group. Hence, the results presented in the next section were computed using the answers of 190 participants.

3.8 Second Evaluation Results

In this section we review and analyze the results obtained for the second user study in which we evaluated the extension of our MoCap database. We also present and discuss how participants performed when presented with end-effector trajectories only.

3.8.1 Analysis of Participants' Emotion Predictions

We start by analyzing the effect of intended emotions, also referred as to true label, on participant predictions [1]. That is, given all the predictions of type $\hat{l}$, for $\hat{l} \in \{$neutral, sadness, happiness, stress, relaxedness$\}$, we compute and analyze the proportion of predictions associated to each one of the intended emotions for each participant. This first analysis shows us whether emotions were unambiguously recognized as well as where misjudgments might have happened. We analyzed separately each representation.

Bar charts of the predictions of neutral, sadness, happiness, stress and relaxedness for each intended emotion, averaged over the participants, are shown in Figure 3.5 for whole-body stimuli and Figure 3.6 for partial-body, i.e., end-effector trajectories, stimuli. Two sets of five one-way repeated measures ANOVA were performed, one set for type of stimuli. Each one of the ANOVA analysis tested the main effect of intended emotions on the participant's predictions. That is, for example, whether most of the *happiness* predictions were indeed associated to stimuli whose intended emotion was *happiness*.

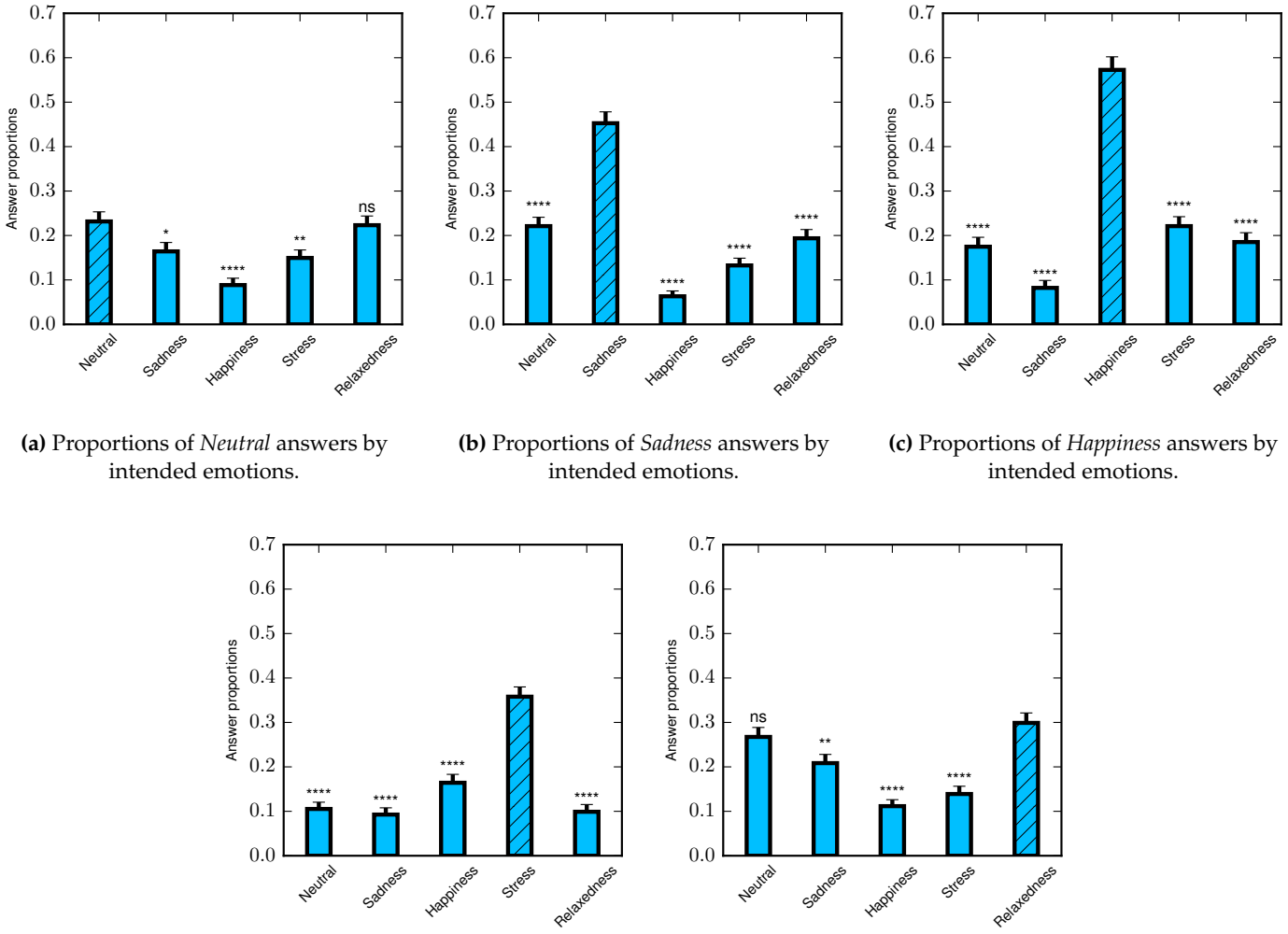
We evaluated the following null hypothesis:

> $H_0(\hat{l})$: *The means of the participants' predictions of label $\hat{l}$ for different intended emotions are equal. For $\hat{l} \in \{$neutral, sadness, happiness, stress, relaxedness$\}$.*

Table 3.8 show the resulting F-statistics, p-values and effect sizes (η^2). At first glance, we note that all ANOVAs analysis showed a significant difference between the proportion of predictions associated to each intended emotion since all p-values are below the level of significance we established, $\alpha = 0.05$. We also observe that according to the interpretation of effect sizes[2] suggested by [54] and summarized in [140], in most of the cases, the intended emotion had a large effect, i.e., $\eta^2 > 0.26$ on the participants predictions $\hat{l}$. Hence, we can reject the null hypothesis $H_0(\hat{l})$ for both type of stimuli.

[1] By predictions we make reference to the emotion label selected for each one of the videos rated by a participant.

[2] Small effect: $0.01 \leq \eta^2 < 0.06$, medium effect: $0.06 \leq \eta^2 < 0.14$ and large effect: $\eta^2 \geq 0.14$.



(a) Proportions of *Neutral* answers by intended emotions.

(b) Proportions of *Sadness* answers by intended emotions.

(c) Proportions of *Happiness* answers by intended emotions.

(d) Proportions of *Stress* answers by intended emotions.

(e) Proportions of *Relaxedness* answers by intended emotions.

Figure 3.5: Proportion of answers obtained for each emotional state for whole-body representation.

	Predictions($\hat{l}$):	*Neutral*	*Sadness*	*Happiness*	*Stress*	*Relaxedness*
$H_0(\hat{l})$ whole body	F(4, 376) =	14.569	71.033	95.959	43.523	22.043
	p-value =	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
	η^2 =	0.349	0.658	0.702	0.579	0.491
$H_0(\hat{l})$ partial body	F(4, 376) =	4.350	14.213	38.364	17.771	29.251
	p-value =	**0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
	η^2 =	0.184	0.429	0.594	0.405	0.603

Table 3.8: F-statistics[3], p-values and effect size (η^2) results from one-way repeated measures ANOVA's for main effect of intended emotion. Both representations are considered. Greenhouse-Geisser correction was used when sphericity assumption was violated. P-values indicating a significant difference at level of significance $\alpha = 0.05$ are highlighted.

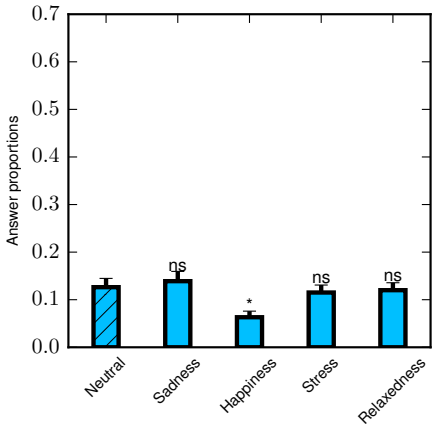
We performed Tukey-HSD post-hoc test with Bonferroni corrections in order to identify the significant differences from the ANOVA's tests. Results are presented in Figure 3.5 for whole-body stimuli and Figure 3.6 for partial-body, i.e., end-effector trajectories, stimuli. The labels above each bar indicate the levels of significance at which the differences among the average predictions for each intended emotion are statistically important. Labels should be interpreted as follows:

Label	Interpretation
****	p-value < 0.0001
***	p-value < 0.001
**	p-value < 0.01
*	p-value < 0.05
ns	p-value > 0.05 (No significant difference)

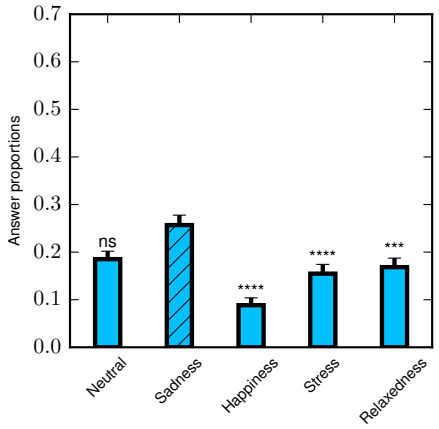
Table 3.9: Equivalence of labels used from Figure 3.5 to Figure 3.9 to indicate statistically significant differences.

For whole-body stimuli, we find that the highest percentage of *happiness*, *sadness* and *stress* predictions are significantly different at $p < 0.0001$ and indeed associated to visual stimuli conveying *happiness*, *sadness* and *stress* respectively. We also observe that most misjudgments happened between stimuli conveying emotional states that share the same arousal levels. More precisely, both Figures 3.5c and 3.5d show that among all intended emotions, stimuli conveying *stress* were more frequently labeled as *happiness* and conversely stimuli depicting *happiness* were more often judged as conveying *stress*. A similar tendency is observed between *sadness* and *relaxedness*. However, we also find that these two emotional states are often judged as *neutral*, that is, as conveying no emotion.

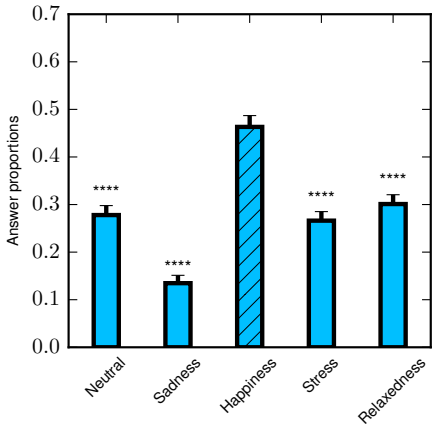
[3]Since we have 5 emotional states and 95 subjects for each representation (i.e., full-body or partial-body), the degrees of freedom of the F-statistics presented here correspond to: $df_{emotion} = 5 - 1 = 4$ and $df_{error} = (5 - 1)(95 - 1) = 376$.



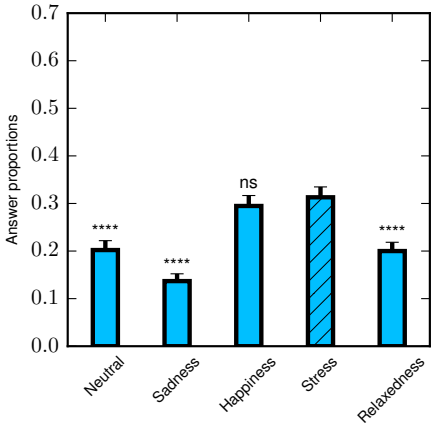
(a) Proportions of *Neutral* answers by intended emotions.



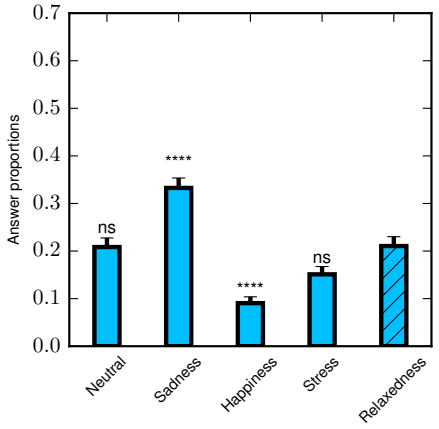
(b) Proportions of *Sadness* answers by intended emotions.



(c) Proportions of *Happiness* answers by intended emotions.



(d) Proportions of *Stress* answers by intended emotions.



(e) Proportions of *Relaxedness* answers by intended emotions.

Figure 3.6: Proportion of answers obtained for each emotional state for partial body (end-effectors and pelvis trajectories) representation.

In particular, both Figures 3.5a and 3.5e show that there is no statistically significant difference between the number of *relaxedness* predictions associated to stimuli conveying a *neutral* or *relaxedness* state, and vice-versa. Our hypothesis regarding this finding is that as suggested by [205], speed plays a significant role on emotion classification and stimuli conveying either *sadness*, *relaxedness* or a *neutral* emotional states were, from a visual perspective, often matched on speed, hence making their discrimination from body motion only much more difficult. Coupled with the similarities in speed profiles, it is also possible that our actors did not generated equally recognizable and emotionally expressive body movements for these three emotional states, which resulted in very different stimuli belonging to the same target emotion. Unfortunately, because of the manner in which our study was designed, we can not determine if actors played a substantial role in the observer's misjudgments, sometimes exceeding the effect of intended emotions.

The results obtained from the evaluations of partial-body stimuli, i.e., end-effector trajectories show, to some extent, a similar tendency. We observe that the highest percentage of *happiness*, *sadness* and *stress* predictions are associated with stimuli depicting these three emotional states. However, it seems that participants were much more prone to errors in their predictions when presented with end-effector trajectories alone. For example, we can see that although they accurately identified a higher activation on stimuli depicting *stress* with respect to low activation emotional states (*sadness* and relaxedness for *instances*), they could not correctly predict the difference between *happiness* and *stress* (see Figure 3.6c) along the pleasantness axis. Similarly, *relaxedness* predictions were frequently attributed to the *sadness* emotional states (see Figure 3.6e). This suggests that representations of end-effector trajectories alone hindered the participant's capacity to accurately judge differences along the valence axis. Hence, it is possible that motions cues necessary for the distinction between possive and negative emotional states might not be longer present in our partial-body stimuli. Lastly, in the case of *neutral* predictions, results shown in Figure 3.6a suggest that participants were merely guessing most of the time and could not categorized the particularities of *neutral* expressions with respect to other intended emotions.

3.8.2 The Effect of Representation and Emotion on Participants' Ratings

The main and interaction effects of motion representation, i.e., whole-body or partial-body stimuli, and intended emotion on the perception of emotionally expressive body movements were evaluated using mixed two-way repeated measures ANOVA. Since each participant only rated the videos belonging to one of the two possible motion representations, we defined *representation* as a between-subject factor and *intended emotion* as a within-subject factor. Five ANOVA's tests were performed on the average accuracy, intensity, valence, arousal and difficulty ratings reported by participants. Using the notation proposed by [211], we list below the null hypotheses evaluated in this user study:

$H_0(1, i)$: *The means of the participants' ratings of i for whole-body and end-effector trajectories stimuli are equal.*

$H_0(2, i)$: *The means of the participants' ratings of i for the different intended emotions are equal.*

$H_0(3, i)$: *Representation type and intended emotion are independent factors and no interaction between the two is present on the participants' ratings of i.*

With $i = \{$*accuracy, intensity, valence, arousal, difficulty*$\}$. Table 3.10 list the resulting F-statistics (degrees of freedom are explained in the corresponding footnote), p-values and effect sizes (η^2). Effects and interactions were evaluated at a significant level of $\alpha = 0.05$.

	Ratings(i):	*Accuracy*	*Intensity*	*Valence*	*Arousal*	*Difficulty*
Representation $H_0(1, i)$	F(1, 188)=	54.308	0.9157	2.477	29.732	1.028
	p-value =	**< 0.001**	0.339	0.117	**< 0.001**	0.311
	η^2 =	0.224	0.005	0.013	0.136	0.005
Intended emotion $H_0(2, i)$	F(4, 752) =	34.918	19.553	71.452	193.713	16.298
	p-value =	**<0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
	η^2 =	0.406	0.293	0.502	0.713	0.201
Intended emotion × representation $H_0(3, i)$	F(4, 752) =	2.951	2.181	14.16	12.981	2.545
	p-value =	**0.019**	0.069	**< 0.001**	**< 0.001**	**0.045**
	η^2 =	0.061	0.045	0.175	0.136	0.059

Table 3.10: F-statistics[4], p-values and effect size (η^2) results from two-way repeated measures ANOVA's for main effect of intended emotion and representation. Greenhouse-Geisser correction was used when sphericity assumption was violated. P-values indicating a significant difference at level of $\alpha = 0.05$ are highlighted.

Representation and Interactions

Contrary to what was expected, we observe that representation, i.e., whole-body or end-effector trajectory stimuli, has a significant effect only on accuracy and arousal ratings. Hence, we can only reject the null hypotheses $H_0(1,$ *accuracy*$)$ and $H_0(1,$ *arousal*$)$. We observe also that the main effect of representation is large on participant's accuracy ($\eta^2 > 0.16$) but medium ($0.06 < \eta^2 < 0.16$) on arousal ratings. In the one hand, the large effect on accuracy indicates that the visual stimuli presented to the observers was crucial for their perception of expressive movements and emotions. In the other hand, the medium effect on arousal ratings suggests that representation alone can not account for all the variability observed on the perception of the activation and kinematic patterns of our expressive motions. To further analyze and identify the significant differences detected by the ANOVA tests, we applied once again Tukey-HSD paired tests with Bonferroni correction on accuracy and arousal ratings.

In the case of participants' accuracy, the main differences are observed on the perception of motions depicting *sadness*, *happiness*, and the *neutral* state. As showed in Figure 3.7a, we find that the recognition rates of these three emotional states decreased between the two types of representations. This effect is much more significant for *sad* movements and for the

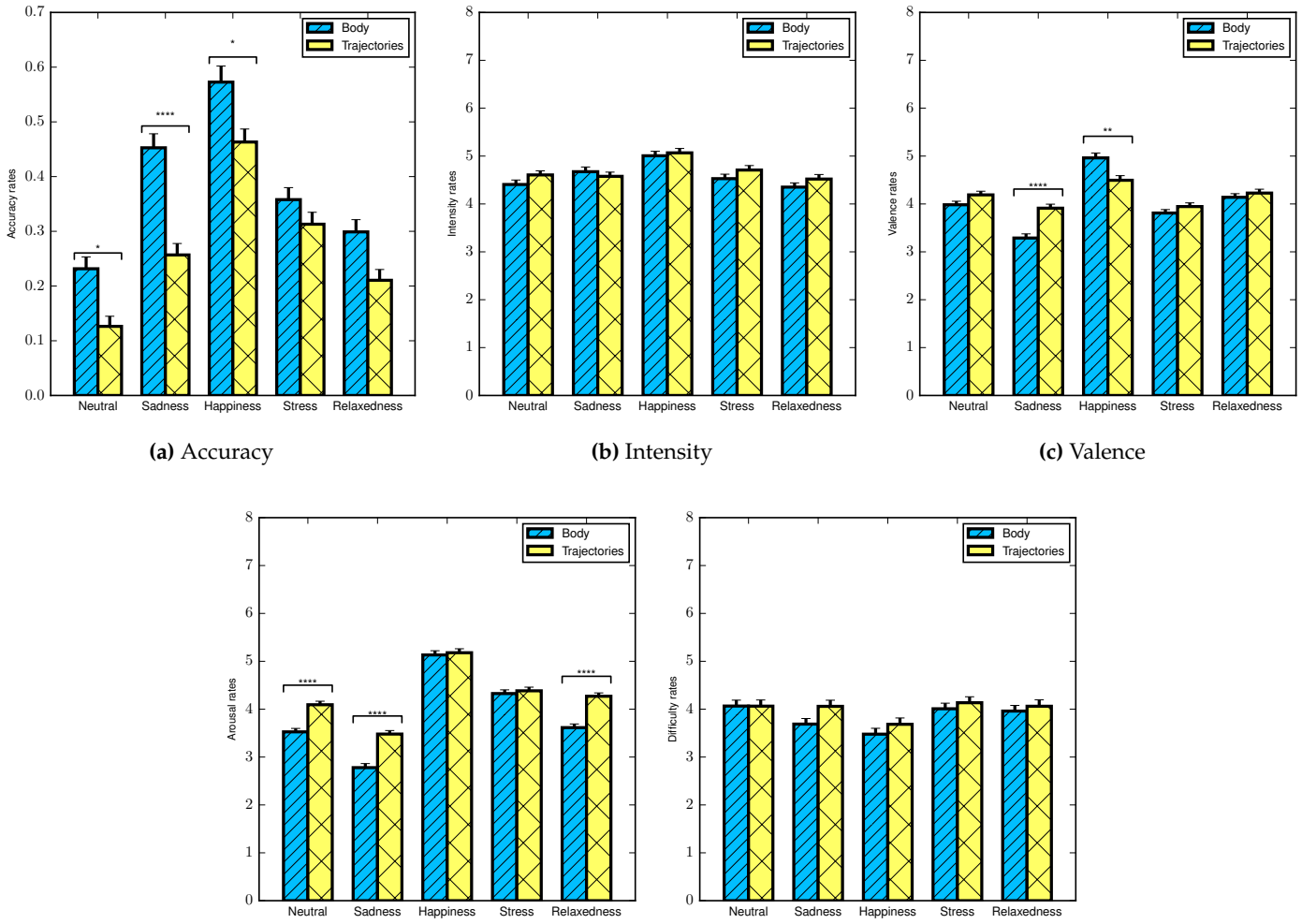
[4]Knowing that the answers of 190 subjects were analyzed, the degrees of freedom of these F-statistics are: *a.)* representation is a between-subjects factor with 2 levels (i.e., whole body or partial body): $df_{representation} = 2 - 1 = 1$ and $df_{error1} = 190 - 2 = 188$, *b.)* intended emotion is a within-subjects factor with 5 levels (i.e., *neutral*, *happiness*, etc.): $df_{emotion} = 5 - 1 = 4$ and $df_{error2} = df_{emotion} \times df_{error1} = 752$, and *c.)* the interaction between these two factors: $df_{interaction} = df_{representation} \times df_{emotions} = 4$ and $df_{error2} = 752$.

neutral state. The former showed a loss of accuracy of approximately 20%, while the latter was not longer recognized above chance level for the end-effector trajectories stimuli. *Happiness* accuracy also decreased, but it remained the best recognized emotional state among participants. Conversely, arousal ratings for partial-body stimuli (i.e., end-effector trajectories) showed an increase for the *neutral, sadness* and *relaxedness* emotional states with respect to whole-body representations. Arousal ratings for *happiness* and *stress* remained the same across the two type of representations, indicating that emotional states characterized by low-activation seem to be perceived as more energetic when representations with reduced body information are employed.

Although ANOVA tests reported that representation only had a significant effect on accuracy and arousal ratings, we find that interactions between this factor and intended emotion were tested as statistically significant at $p<0.05$ for four of the five dependent variables we analyze. Thus, we reject the following null hypotheses: $H_0(3, \mathit{accuracy})$, $H_0(3, \mathit{valence})$, $H_0(3, \mathit{arousal})$, and $H_0(3, \mathit{difficulty})$. This interaction registered, in average, a medium effect size ($\eta^2 \approx 0.107$), indicating that the interplay between the type of representation and the intended emotion is responsible for approximately 10% of the variability observed on participants' perception of expressive bodily emotions. Since the effects of this interaction on accuracy and arousal ratings were already discussed, we focus hereinafter on the analysis of valence and difficulty ratings.

Follow-up post-hoc paired Tukey-HSD tests with Bonferroni corrections on valence ratings over emotion and representation reported statistically significant differences ($p<0.01$) between the average valence rating of *sadness* and *happiness* (see Figure3.7c). We find that valence ratings of motions depicting *happiness* registered a significant decrease between whole-body and end-effector trajectories stimuli. Conversely, *sadness* motions were rated with a higher valence level when depicted through end-effector trajectories. Since observers were much more accurate on the recognition of these two emotional states (see Figure 3.7a) than for the other intended emotions, it is likely that their perception of the differences between them was considerably impaired by the change of representation. Hence, when presented with end-effector trajectories alone, participants were less accurate on the perception of both *happiness* and *sadness* as well as on their judgments about the pleasantness level associated to these two emotions. This claim is supported by the absence of significant differences between the valence ratings of *neutral*, *stress* and *relaxedness* emotional states for both stimuli. Namely, it seems that participants could not accurately decode the nuances, along the valence axis, necessary to distinguish between these three emotional states and resorted to rate them equally for both stimuli as can be seen in Figure 3.7c.

Although an ANOVA test indicated a significant interaction between the type of representation and intended emotion on difficulty ratings, pairwise post-hoc tests found no significant differences ($p>0.05$). Contrary to what was expected, the perception of emotionally expressive body motions was rated as equally difficult for both representations and all emotional states (see Figure 3.7e). Given the decrease on observers' accuracy when presented with end-effector trajectories, we expected higher difficulty values for this type of representation over all emotional states. After looking at the effect sizes reported for representation, intended emotions, and their interaction on difficulty ratings (last column of Table 3.10), we can see that these three factors combined account only for approximately 21% of the variability in the participants' difficulty assessments.



(a) Accuracy **(b)** Intensity **(c)** Valence

(d) Arousal **(e)** Difficulty

Figure 3.7: Mixed two-way ANOVA (intended emotion, representation) accuracy and other ratings for all participants. Significant differences were labeled following the convention listed in Table 3.9.

This indicates that other factors not considered during the ANOVA test (e.g., actors expressiveness, type of motions, errors in decoding emotions by participants [101]) might need to be analyzed in order to understand and explain the variability observed on participants' difficulty ratings.

Intended Emotion

Intended emotion has a significant effect on the participants' perception of expressive movements for all ratings i. Hence, we reject the null hypotheses $H_0(1, i)$ for all dependent variables. Similarly, the main effect of intended emotion is large ($\eta^2 > 0.16$) in all cases, indicating that the conveyed emotion was a critical factor on participants' accuracy and ratings on intensity, difficulty, arousal and valence for all visual stimuli.

Follow-up post-hoc analysis show that accuracy for all emotional states is above chance level (20%) for whole-body stimuli. However, only four of the five emotional states reported recognition rates superior to chance for the end-effector trajectories stimuli. Significant pairwise differences between participants' accuracy were found at a Bonferroni corrected significant level of $p < 0.01$ for most of the intended emotions. We observe for instance that for both representations, *happiness* registered the highest accuracy rate followed, in no particular order, by *sadness* and *stress*. No significant difference (p>0.05) between the accuracy rates of *relaxedness* and the *neutral* state were found for both representations. This suggests that observers found the same difficulties when trying to distinguish stimuli conveying these two emotional states from each other. A closer look at the proportion of stimuli labeled as either *relaxedness* or *neutral* (see Figures 3.5a and 3.5e for body representation, and Figures 3.6a and 3.6e for end-effector trajectories representation) show this is indeed the case. In other words, the accuracy and misclassification rates for these two emotional states are significantly close.

The pairwise differences on intensity ratings with respect to the intended emotion were found to be only significant ($p < 0.05$) for happiness and sadness on whole-body stimuli, and for happiness alone on end-effector trajectories. This indicates, as found in a previous study [75], that the body movements perceived as being the most expressive are the more easily recognized. This claim might explain why no significant differences on intensity ratings (p>0.05) were found for the emotional states the less accurately recognized (stress, neutral and relaxedness). We highlight that neither the type of representation or the interaction between this factor and intended emotion were found statistically significant for intensity ratings. This suggests that both the changes on participants' accuracy and the inexistent difference between the intensity ratings for both representations might be due to errors and discrepancies in the manner in which our actors encoded the different intended emotions through their movements.

A similar pattern was observed on the pairwise differences of valence ratings. For whole-body stimuli, we observe that both *happy* and *sad* movements have valence ratings significantly different (p<0.001) from each other and from the other intended emotions. This might be due to the fact that both *happiness* and *sadness* were more accurately recognized than other emotional states, making their assessment along the valence much easier and accurate for participants. No significant difference between the ratings of *neutral*, *relaxedness* and *stress* were observed. For partial-body stimuli, i.e., end-effector trajectories, only two pairwise differences were found as significant (p<0.001): *happiness-sadness* and *happiness-stress*. It seems that most of the cues used by our observers to assess valence ratings from emotion-

ally expressive body motions are not longer present or encoded in a visual representation containing only end-effector trajectories.

Pairwise differences on difficulty ratings were only significant ($p<0.01$) for the emotional states the best recognized. More precisely, we observe that difficulty ratings for *happiness* and *sadness* were substantially smaller than those of the other intended emotions for the whole-body representation. Similarly, we find that when presented with end-effector trajectories only, participants frequently ($p<0.001$) judged the perception of motions depicting *happiness* as less difficult than for the other intended emotions. This indicates that the more recognizable the expressive motions, the smaller the difficulty rating participants would attribute to it.

Post-hoc analysis of arousal ratings with respect to the intended emotion showed that for whole-body stimuli there are significant differences ($p<0.0001$) for almost all pairwise comparisons. This result is consistent with the findings reported by [197], which indicates that there is a direct relation between the corresponding activation of perceived emotional states and the kinematic information observed on body motions. Additionally, we find that there is no notable difference between the arousal ratings of *neutral* and *relaxedness* emotional states. This result support our claim about the reasons behind the confusion in the perception of these two emotional states. That is, since *neutral* and *relaxed* motions are kinematically very similar, it is both hard to distinguish, and to perceive and rate the subtle differences of activation between them. In the case of end-effector trajectories, we find that the pairwise differences on arousal ratings follow the same pattern than those on participants' accuracy rates. Namely, arousal ratings for *happiness* and *sadness* are significantly different ($p<0.0001$) in the light of other emotional states. However, it seems that most of the confusions on participants' perceptions of the intended emotions are due to ambiguities on arousal ratings. In particular we found that our observers did not perceive differences between the activation levels of the following pairs: *neutral-relaxedness*, *neutral-stress* and *relaxedness-stress*.

3.8.3 The Effect of Gender

In Section 3.8.2 it was mentioned that although representation and emotion explained a considerable portion of the variability observed in the participants' perception and ratings of expressive movements, there might be additional factors having an influence in our results. A recent study found that women and men perceive affective movements differently. Males surpass women on the recognition of happy actions, whereas females are better at perceiving hostile and non-expressive actions [227]. Hence, we decided to investigate the effect of gender on the perception of intended emotions for the two types of visual stimuli used in this study. In order to simplify the analysis and interpretations of the possible effects, we conducted two sets of two-way ANOVA tests. Namely, we separated our results by gender and performed two-way ANOVA's with representation and intended emotion as between and within factors respectively. Accuracy, difficulty, arousal, valence and intensity ratings were evaluated separately. Once again, we adopt the notation proposed by [211] to formulate the null hypotheses we aim to test:

> $H_0^G(1, i)$: *The means of the G participants' ratings of i for whole-body and end-effector trajectories stimuli are equal.*

$H_0^G(2,i)$: *The means of the G participants' ratings of i for the different intended emotions are equal.*

$H_0^G(3,i)$: *Representation type and intended emotion are independent factors and no interaction between them is present on the G participants' ratings of i.*

with $i = \{accuracy, intensity, valence, arousal, difficulty\}$ and $G = \{male, female\}$.

	Ratings(i):	*Accuracy*	*Intensity*	*Valence*	*Arousal*	*Difficulty*
Representation $H_0^{male}(1,i)$	F(1, 93) =	20.089	0.052	0.556	10.625	4.993
	p-value =	**< 0.001**	0.818	0.457	**0.002**	**0.027**
	η^2 =	0.177	0.001	0.005	10.253	0.051
Intended emotion $H_0^{male}(2,i)$	F(4, 372) =	21.947	8.336	36.126	77.599	16.185
	p-value =	**<0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
	η^2 =	0.469	0.293	0.504	0.625	0.168
Intended emotion × representation $H_0^{male}(3,i)$	F(4, 372) =	2.83	1.719	10.19	3.575	0.965
	p-value =	**0.024**	0.152	**< 0.001**	**0.007**	0.419
	η^2 =	0.107	0.052	0.235	0.075	0.041

(a) ANOVA results for male participants

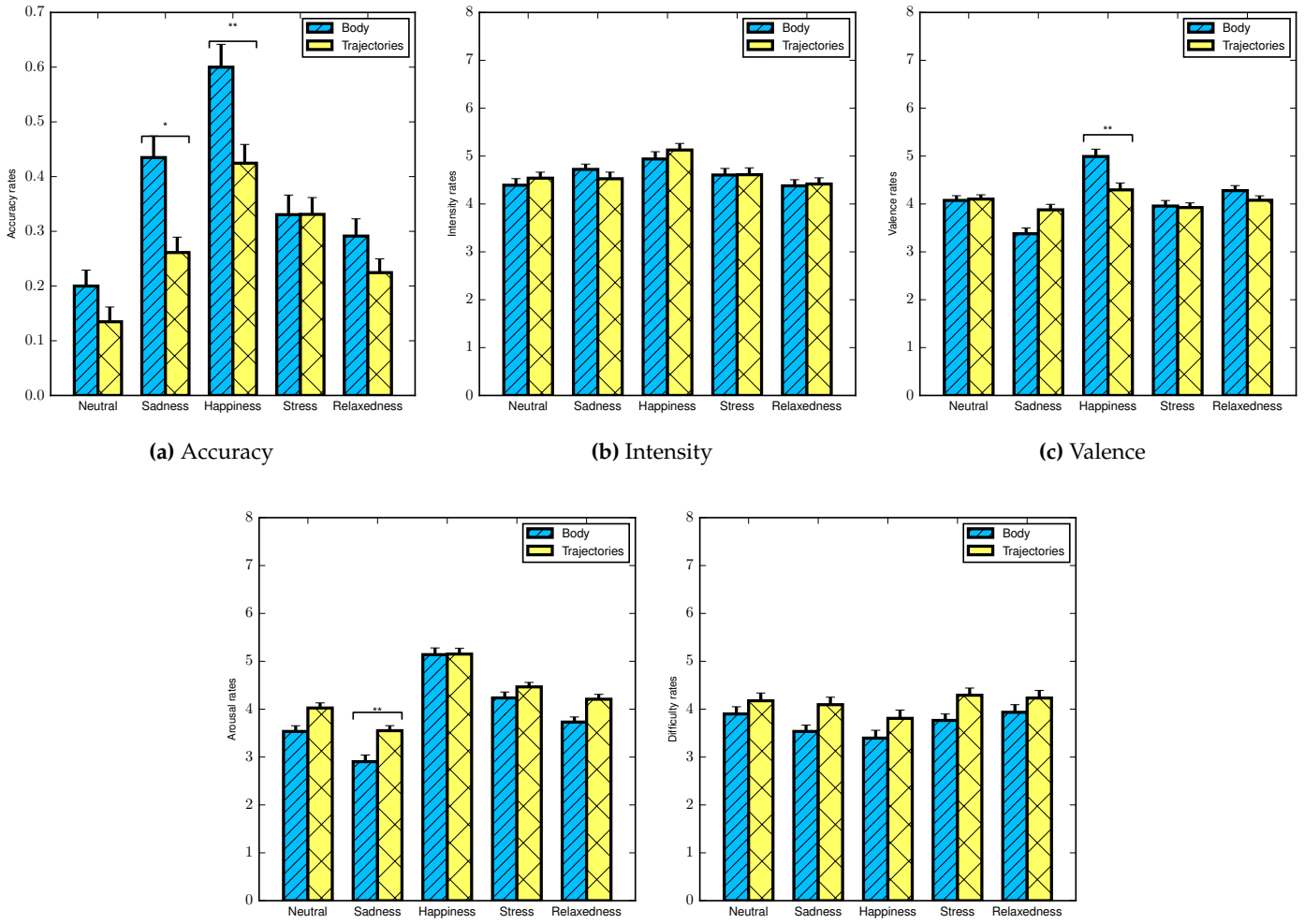
	Ratings(i):	*Accuracy*	*Intensity*	*Valence*	*Arousal*	*Difficulty*
Representation $H_0^{female}(1,i)$	F(1, 93) =	35.052	1.41	9.968	20.487	0.132
	p-value =	**< 0.001**	0.236	**0.002**	**< 0.001**	0.717
	η^2 =	0.262	0.015	0.098	0.181	0.001
Intended emotion $H_0^{female}(2,i)$	F(4, 372) =	13.913	11.885	35.635	122.449	12.746
	p-value =	**<0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
	η^2 =	0.359	0.326	0.534	0.810	0.261
Intended emotion × representation $H_0^{female}(3,i)$	F(4, 372) =	2.101	2.377	5.212	11.951	2.899
	p-value =	0.080	0.051	**< 0.001**	**< 0.001**	**0.031**
	η^2 =	0.076	0.094	0.148	0.377	0.115

(b) ANOVA results for female participants

Table 3.11: F-statistics [5], p-values and effect size (η^2) results from two-way repeated measures ANOVA's for main effect of intended emotion and representation. Analysis was done on males (top table) and females (bottom table) separately. Greenhouse-Geisser correction was used when sphericity assumption was violated. P-values indicating a significant difference at level of $\alpha = 0.05$ are highlighted.

From the ANOVA results summarized in Table 3.11, there is a significant interaction of representation and intended emotion in the males' accuracy, arousal, and difficulty ratings; hence we reject the null hypotheses $H_0^{male}(3,i)$ for $i \in \{accuracy, arousal, valence\}$ and retain the remaining two null hypotheses $H_0^{male}(3,i)$ for $i\{intensity, difficulty\}$.

[5]Knowing that the answers of an equal number of males and females (95) were analyzed, the degrees of freedom of these F-statistics are: *a.)* representation is a between-subjects factor with 2 levels (i.e., whole body or partial body): $df_{representation} = 2-1 = 1$ and $df_{error1} = 95-2 = 93$, *b.)* intended emotion is a within-subjects factor with 5 levels (i.e., *neutral*, *happiness*, etc.): $df_{emotion} = 5-1 = 4$ and $df_{error2} = df_{emotion} \times df_{error1} = 372$, and *c.)* the interaction between these two factors: $df_{interaction} = df_{representation} \times df_{emotions} = 4$ and $df_{error2} = 372$.



(a) Accuracy **(b)** Intensity **(c)** Valence

(d) Arousal **(e)** Difficulty

Figure 3.8: Mixed two-way ANOVA (intended emotion, representation) accuracy and other ratings for male participants. Significant differences were labeled following the convention listed in Table 3.9.

For the female participants in turn, no significant interaction was found for accuracy and intensity ratings, hence we reject $H_0^{female}(3,i)$ for $i \in \{valence,\ arousal,\ difficulty\}$ and retain $H_0^{female}(3,i)$ for $i \in \{accuracy,\ intensity\}$.

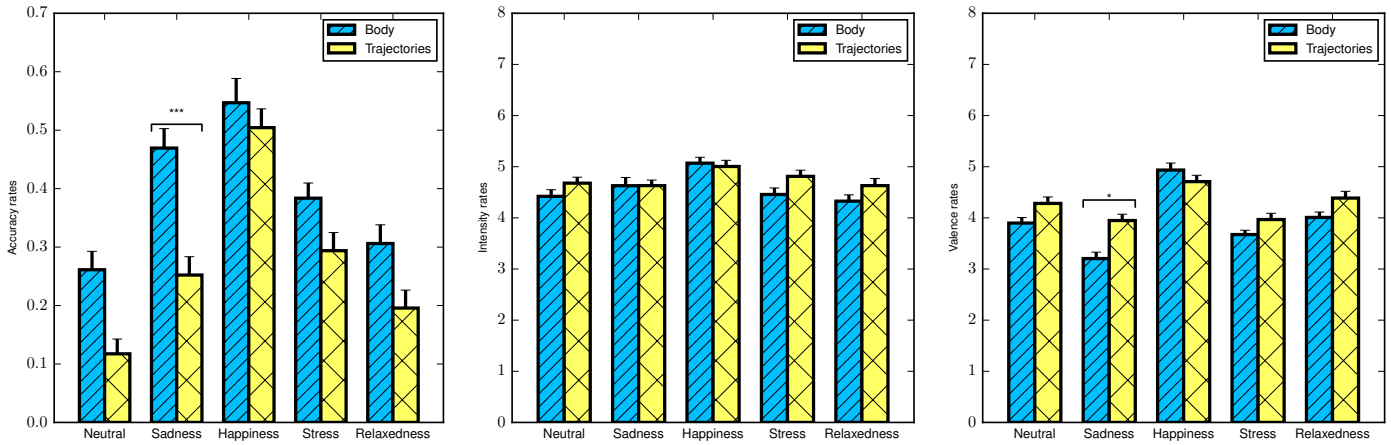
The intended emotion was found statistically significant for all ratings of both males and females ($p<0.001$). There are, in turn, differences in the main effect of representation between females and males. Whereas representation has a significant effect on males' accuracy, arousal and difficulty ratings, it was only found as significant on females' accuracy, valence and arousal ratings. Hence we reject the following null hypothesis: $H_0^{male}(1,i)$ for $i \in \{accuracy,\ arousal,\ difficulty\}$, $H_0^{female}(1,i)$ for $i \in \{accuracy,\ arousal,\ valence\}$ and $H_0^G(2,i)$ for all ratings on both males and females. For the remaining hypotheses, we did not find enough evidence to reject them in favor of the alternative hypotheses.

Bar charts of average ratings for each type of representation are shown in Figure 3.8 for males and in Figure 3.9 for females. Significant pair-wise differences between the same intended emotion for both type of representations are labeled using the notation described in Table 3.9. A thorough analysis of the differences observed between males and females on the perception of emotionally expressive body motions is discussed below.

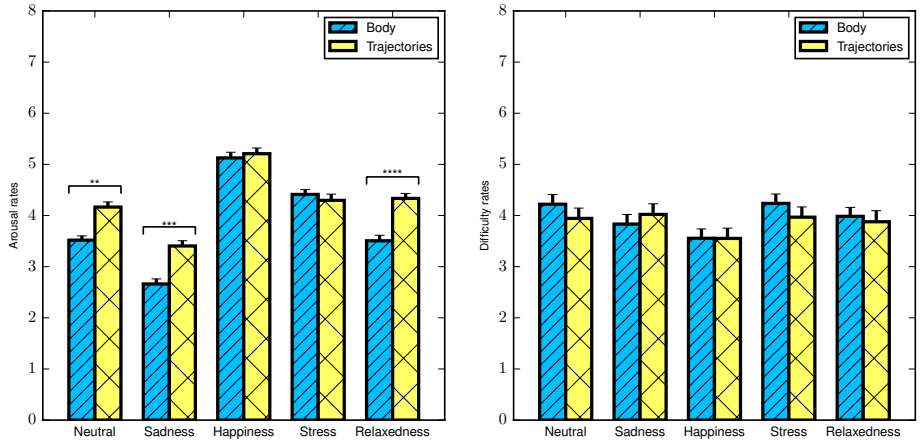
For the interactions common to both genders, i.e., valence and arousal, we observe that, in average, females increased their valence ratings when presented with end-effector trajectories (see Figure 3.9c). In the contrary, males would either rate emotions in the same manner for both representations or decrease their ratings, in particular for *happy* motions (see Figure 3.8c). Pair-wise post-hoc Tukey-HSD tests showed that, for whole-body stimuli, both males and females could correctly assess the differences between the valence levels of *sadness* and *happiness* ($p<0.0001$). However, they perceived no difference among the other three emotional states. We observe also that women noted *sadness* with smaller valence values than males and that males assigned greater valence ratings to *happiness* than females. This suggests that, as reported by [227], males are indeed more accurate on the perception of positive emotions such as *happiness*, while females are better at perceiving negative emotions such as *sadness*.

Both females and males were, in average, less accurate in their appreciations of arousal ratings when presented with end-effector trajectories. However, females would frequently assign, with respect to the whole-body representations, higher arousal values to partial-body stimuli than males. This can be seen on the three pair-wise significant differences ($p<0.01$) found for females (see Figure 3.9d) against one significant difference ($p<0.01$) on males' ratings (see Figure 3.8d). An interesting result shows that they both agreed on their perceptions about the activation level of *happy* motions across representations. This might be due to *happiness* being the emotional state the best recognized by both females and males.

Although intended emotion was found to have a main effect on intensity ratings, pair-wise post-hoc tests showed that, in the case of whole-body representations, only intensity ratings for movements depicting *happiness* were found as significantly different for both females and males ($p<0.05$). This can be due to *happiness* being the best recognized emotional state for both type representations. In the case of end-effector trajectories, only males perceived *happiness* as being more expressively compelling than the other intended emotions ($p<0.001$). Females, in turn, did not perceive differences on the expressiveness of the intended emotions for this type of representation.



(a) Accuracy **(b)** Intensity **(c)** Valence



(d) Arousal **(e)** Difficulty

Figure 3.9: Mixed two-way ANOVA (intended emotion, representation) accuracy and other ratings for female participants. Significant differences were labeled following the convention listed in Table 3.9.

Regarding difficulty ratings, we observe that, for whole-body stimuli, both females and males agreed on *happiness* being easier to recognize ($p<0.01$) with respect to *neutral* and *relaxedness*. However, only females found that making the distinction between *happiness* and *stress* was equally simple ($p<0.001$). No significant difference between *sadness* and *happiness* difficulty ratings was found for both males and females. In the case of end-effector trajectories representation, only females find happiness equally easy to recognize ($p<0.05$) in the light of the other intended emotions. This finding is supported by the small difference between the accuracy rates reported for *happiness* across all representations (see Figure 3.9a). In the contrary, as suggested by the loss of accuracy reported by males in Figure 3.8a, males found that the perception of emotions from end-effector trajectories is equally difficulty for all intended emotions.

Finally, post-hoc pairwise Tukey-HSD test with Bonferroni corrections on accuracy rates show that *sadness* and *happiness* are the two emotional states the best perceived from the motions in our database ($p<0.05$). We observe also that, in the case of *happy* motions, females seemed to be less sensible to the change of representation than males (see Figures 3.8a and 3.9a). Futhermore, while females recognized *neutral* examples better than males, males were much more accurate on the perception of motions conveying *stress*. However, no significant difference was found between the accuracy rates of motions conveying *neutral*, *relaxedness* and *stress* for both females and males. This suggests that either our observers could not decode the subtle differences between these three emotional states, or that the motions produced by our actors did not account for all the nuances necessary to distinguish between them.

3.9 Summary and General Discussion

In this chapter, a new motion capture (MoCap) database inspired by three of the key ideas of *physical theater* theory: *the expressive body*, *the corporal mime* and *the neutral mask*, was described and evaluated. This database was designed and built for both the analysis and synthesis of emotionally expressive bodily motions.

The proposed database consists of 5 different motion sequences: *i.)* three magic tricks: *the disappearing box*, *pulling a rabbit from a hat*, and *taking scarves from an empty jacket*, *ii.)* locomotion examples, and *iii.)* improvisation sketches. These sequences were performed by 7 laypersons under 5 emotional states: *happiness*, *sadness*, *stress*, *relaxedness*, and *neutral*. A combined mood induction procedure (story-based and imagination-based MIP) was used in order to facilitate the enacting of the selected emotions.

Two different user studies were conducted in order to validate the expressive content of the proposed database. Additionally, one of these studies was also employed to assess whether emotion perception was still feasible when only end-effectors and pelvis trajectories were displayed. The results obtained from this perceptual evaluation represented a first validation of the parameterization of expressive bodily motions proposed in this thesis.

The analysis of the results obtained from both studies shows that the expressive content conveyed by the motion sequences in our database was recognized above chance level (20%) with overall recognition rates of 53% for the first study and 38% for the second one. The difference between the recognition rates of the first and second user studies might be explained by the changes we later introduced in the second study. Namely, the *other* option was not

longer available among the possible answers, participants simultaneously rated motions of two different actors rather than one single actor as it was done in the first study, and the angle at which the animated character was displayed also changed. For both studies we found that *happiness*, *sadness*, and *stress* were among the best recognized emotional states. *Relaxedness* and *neutral* states, although recognized above chance level, were often confounded either with each other or with *sadness*.

Despite reporting recognition rates above chance level, we observed that the motion sequences within our database were less accurately recognized than movements considered in the databases surveyed in Section 3.2. We have identified three possible reasons:

i.) We have included two emotional states, i.e., *relaxedness* and *stress*, that are not often considered when investigating the perception of emotions from motion and/or when designing expressive MoCap corpora. Thus, there is little knowledge whether additional contextual factors (e.g., who the character is, where she/he is, what her/his current task is) and visual cues such as facial expressions, gaze direction, etc. [85], are needed for a more accurate identification of these two emotions [105].

ii.) Gunes *et al.* [105] pointed out that the understanding of the action being performed is critical for emotion perception from human body motion. It is likely that participants had a harder time trying to identify and understand the actions involved in the magician sequences and improvisation sketches. Thus, they were less accurate to ascribe the intended emotional states to the motion clips they were presented with. This claim is further supported by the results reported in [67], which suggests that there is a hierarchy of action recognition. For instance, the recognition of locomotion and functional daily movements as the ones found in the databases [87, 205, 130] is easier than the recognition of less fundamental motions such as social, artistic, and instrumental actions.

iii.) Actors were given a short scenario that helped them to better contextualize, interpret, and enact each of the intended emotions in the database. This kind of induction mood procedures privilege more natural and spontaneous expressive motions over more easily recognized movements [14]. As result, accuracy rates for this kind of expressive motion are often lower with respect to the ones reported for portrayed and acted motion databases [7, 75, 129, 187].

The perceptual evaluation of the proposed low-dimensional motion parameterization showed that although observers were less accurate in their judgments, 4 of the 5 intended emotional state were still recognized above chance level. Furthermore, we observed that *happiness* remained the best recognized emotional state, followed *stress* and *sadness*. Overall, we found that the emotional states characterized by lower arousal levels were the most affected by the change of representation. They were perceived as being more active, which in turn resulted in lower precision rates for *happiness* and *stress* predictions. Nonetheless, we notice that the chosen low-dimensional representation still encodes (with little loss) motion cues salient to the perception of emotion via bodily motions.

Chapter 4

Validation of Low-Dimensional Parameterization Through Classification

Contents

Motivated by the intuition that a good low-dimensional parameterization of expressive bodily motions should also provide good results in emotion recognition, in this chapter, we describe how, by means of an automatic affect classification model, we quantitatively validate the relevance and suitability of the motion model proposed in this thesis.

We address the following hypothesis:

"If the spatio-temporal trajectories of a manually selected subset of body joints (i.e., end-effectors and pelvis) encode most of the motion variability related to the expression of affect, the performance of an automatic affect classification model trained on features computed from trajectories should be close to the recognition rates obtained from: (a) human observers, and (b) the same classification model

trained either on an automatically selected subset of features or on features computed using the entire body".

The reasons behind this hypothesis are two-fold:

i.) The learning and posterior performance of a classifier are inherently determined by how the instances on which the classifier is trained and tested are represented. The better the representation separates the data, the easier it will be for a classifier to establish the boundaries between classes. Thus, if the same classifier shows similar performances on two distinct representations with equal or different dimensionality (e.g., manually selected subset vs. automatically selected subset or manually selected subset vs. whole set), it is because there is no significant loss of information when only the manually selected subset is considered.

ii.) Human motion is a highly correlated and redundant type of multidimensional data, which is actually constrained to a lower dimensional representation. An experimental evaluation based on classification results can help us to define a simpler motion model with negligible loss of affect-related information

In this chapter we describe the experimental protocol used in the validation of the aforementioned hypothesis. We review the challenges associated to the classification of motion examples as well as the approach we adopted for the definition of alternative body joint subsets.

4.1 Related Works

During the last decade there has been a growing interest in the automatic classification of affect and emotions based on bodily expressions. This growth has been driven by various advances (e.g., affordable devices for 2D and/or 3D body tracking) and demands (e.g., less controlled and more naturalistic interaction environments) made in the human-machine interaction domain [105].

Previous work on automatic recognition of affect from movements vary in terms of many different aspects: the manner in which emotions are represented (categorical [129] or dimensional models [130]), the number of subjects considered, the type (acted [129] or non-acted [215]) and kinematic variability (fixed and specified trajectories [21] or unconstrained movements [213]) in the movements to classify, the nature of the movements being analyzed (dance [41], theater gestures [220], daily actions [88], etc.), the classification approach (discriminative [130] and/or generative models [89]), and the manner in which movements are represented (feature [242] or model based [213]). Furthermore, the developed classification models are built so as to be person-specific (classifier trained and tested in the same subject) or interpersonal (training and testing samples come from different subjects), and movement-dependent (training and testing samples belong to the same motor behavior) or movement-invariant (testing samples belong to motor behaviors non observed during training). Nonetheless, all the approaches found in the existing literature aim to engineer a classification model whose generalization and recognition capabilities are the highest.

In this thesis, rather than designing and building a classification model with improved accuracy and generalization, we aim to analyze and study the behavior of a standard classification method when trained in different feature subsets. Particularly, we seek to evaluate from an automatic classification standpoint how informative is the proposed low-dimensional motion representation when compared with the entire body and/or automatically determined feature subsets. For this reason, in this section we briefly review the previous classification studies that are closer to the approach discussed in this chapter. We highlight the aspects of motion representation and feature design, data dimensionality, prediction models, subject and movement invariance and learning databases inspired by the performing arts. More complete and general reviews on automatic affect classification from movement can be found in [105, 131, 136].

In the literature, we can identify two main approaches to movement modeling for affect classification: feature-based and generative model based. While the former approximates motion data through a discrete number of features, the latter aims to estimate the stochastic process from which the observed movements were generated [211]. Feature-based approaches are frequently privileged because of their simplicity, interpretability and low computational cost. Furthermore, expert-knowledge can be easily incorporated into the choice of features used to represent the movement data.

A large variety of features can be computed from movement data. In general, the four following approaches are commonly used for defining the feature-based space in which classification is done: *i.)* Data analysis and modeling methods such as functional analysis [212] are used to automatically extract features relevant to affect classification. *ii.)* Manually-selected features frequently inspired by an understanding of the type of movements to analyze during classification, e.g., [129, 131, 251, 260]. *iii.)* High-level descriptors inspired by motion notation systems [4, 60, 88, 220, 242] such as Laban [147], body action and posture (BAP) [60], or multi-level body movement notation system (MLBNS) [86]. *iv.)* Manually-selected features grounded in psychological studies about the perception and expression of affect and emotion via bodily movements [20, 95].

The feature space defined through any of the last three approaches can often exhibit high-dimensionality and/or information redundancy, which in turn can result in poor generalization capabilities. Hence, transforming and reducing the feature space can improve classification results as well as provide an additional insight into the most salient features for the recognition and expression of affect. This transformation can be achieved through either dimensionality reduction techniques such as principal component analysis (PCA) [95, 126] and linear discriminant analysis (LDA) [130], or feature selection techniques such as filter [20] or wrapper [88] methods.

Once a compact feature base representation has been defined, a classification model that maps the selected features to affective or emotional states needs to be chosen. Traditional classifiers such as *Support Vector Machines* (SVM) [21, 213], *Neural Networks* (ANN) [215], *Naive Bayes* (NB) [130] and *Nearest Neighbors* (NN) [129] are among the most used. *Random Forest* (RF) methods [34] in particular have gained an increased interest in automatic classification of affect [4, 88, 242, 260] during the last years. They are not only easy to tune and computationally cheap, but also capable of defining complex class boundaries and known to achieve the best recognition performance [47, 82]. Furthermore, they can also be used as a feature selection method.

A major challenge in designing affect automatic classifiers for body motions is the large

amount of variability in human movement. A robust affect classifier should be able to recognize and generalize to different subjects and movement classes. To guarantee good generalization capabilities to several subjects, the common approach is to apply a double classification procedure. That is, a first classifier identifies the subject and a second classifier determines the emotional state conveyed by the subject's movement [131]. A similar approach is adopted for enhancing movement invariance. For instance, Bernhardt *et al.* [21] used Hidden Markov Models (HMM) to determine the action being performed and then employed a SVM classifier to identify the emotional state. Recently, Aristidiou and colleagues [4] achieved good movement generalization through the combination of a overlapping sliding window approach and a majority vote schema. The overlapped windows provided a good estimation of the temporal dynamics of the movements within the sequence and the majority vote smoothed the classifier predictions.

Finally, expressive motion corpora inspired by the performing arts have been lately used to train and test affect classifiers. For instance [4] used contemporary dance motions performed by 6 different professional dancers, [242] employed orchestra conductor gestures recorded during 8 different rehearsals, and [220] used the improvisation performances of 10 professional theater actors.

4.2 The Challenges of Motion Classification

A motion sequence $\mathbf{M}$ can be formally defined as a multidimensional time series or sequence of length T and dimensionality D – with D being determined by the type of body representation (e.g., full-body or end-effectors representation) – to which an emotion label is associated. Given $\mathbb{L}$ as a the set of emotion labels, the task of motion sequence classification is to learn a classifier $\mathcal{C}$, which is a function mapping a motion sequence $\mathbf{M}$ to a class label $l \in \mathbb{L}$, written as:

$$\mathcal{C} : \mathbf{M} \rightarrow l,\ l \in \mathbb{L} \tag{4.1}$$

However, as any other time series, a motion sequence is characterized by being large in data size, variable in length and high-dimensional [90]; making its classification through standard methods a difficult task. As presented by [267], there are three major challenges in time series classification:

- There are no explicit or standard features to represent time series and most of the commonly used classifier models (e.g., decision trees, support vector machines, neural networks, etc.) only take a vector of features as input data.
- Although the existing literature offers a wide range of techniques for extraction and selection of features, selecting an adequate approach is far from trivial. Aspects as number of samples, nature of the data, computational cost and dimensionality have to be taken into consideration.
- Besides a good performance, one often aims to get an interpretable classifier. However, any possible interpretation mainly depends on how the time series is represented.

An additional challenge, proper to sequences of bodily expressive motions, is that the phenomenon we wish to characterize and discriminate is not directly measurable from the motion sequence. Instead, the affective content is encoded in the action or behavior being executed and whichever representation we chose, it must account for this somehow *hidden* information.

We can see then that in order to accurately classify a time series, the choice of representation is of fundamental importance. A good representation should not only capture the nuances we seek to classify, but should also reduce, if redundancy exits, the dimensionality of the data. The existing approaches for time series classification can be divided into two large categories: those on which the classifiers are adapted to work with this type of data, distance-based models belong to this category since similarity metrics can be adapted to work with time series; and those on which the time series are transformed to an equivalent representation such that the existing classification models can be directly used, such as feature-based and model-based classifiers. We refer readers to [90, 267, 157] for a further exposition about this topic.

Between those two categories, we decided to use a *feature-based model* in which each motion sequence is transformed into a set of features vectors of fixed-length. The motivations behind this choice are:

- The interpretability of the resulting classifier. Numerous standard classifiers such as Random Forest, Kernel Rigde Regression, Support Vector Machines with a linear kernel, etc. offer an insight into the importance and relevance of all features in the separation among classes.
- The simpler and more understandable assumptions made by this kind of model. Feature-based models benefit from an *a priori* knowledge of the kind of phenomenon one is trying to explain and classify. Furthermore, in the case of expressive motion classification, feature-based movement representations can also be used to examine which motion aspects should be considered when generating new expressive motions.
- The suitability of this approach to the computations necessary to prove our hypothesis. State-of-the-art feature selection techniques can be directly used to generate the feature subsets against which we wish to compare the proposed parameterization.
- Its simplicity and low computational cost.
- Previous studies on the perception and automatic recognition of affect offer a good insight into the set of features that can be used to correctly characterize sequences of expressive body motions.
- We can use a discriminative classification model as our affect classifier. This type of models is known to provide better classification results than generative approaches such as Hidden Markov Models (HMM) [213].

4.3 Feature-Based Representation of Motion Sequences

We look for a set of features, also referred to as cues, suitable for the classification of affective content embedded in expressive bodily motions. The selected features should not

only maximize the differences between distinct emotional states, but also minimize the variability (i.e., the action, subject's style, etc.) among the motions depicting the same emotion. Additionally, they should be independent of the functional behavior described by the body movements we analyze and easily computed for whichever motion representation we use.

4.3.1 Lessons From Psychology: How Humans Perceive Emotions?

When considering body motion as a visual modality in non-verbal and affective communication, Dittrich and Atkinson [68] highlighted three main sources of relevant information: *i.)* posture, also called form, and its changes over time (e.g., the position of body parts relative to themselves and to the whole body); *ii.)* kinematics (quantities related to how people move their bodies, e.g., velocity and acceleration); and *iii.)* dynamics (motion specified in terms of mass and forces). Each one of these sources can be in turn characterized by a large number of potentially relevant features.

To consider all possible features that previous perceptual studies have reported to be correlated with the perception and expression of emotion through body motions proves to be an inefficient approach. Not only we will increase the already high dimensionality of the data, but we will also be confronted with attributes that are not meaningful to the type of motion behaviors we study in this thesis (for example, elbow flexion and stride length are mainly relevant when considering gait patterns [130, 205]) and that are not possible to compute from the low-dimensional representation we are trying to validate (features such as hand-to-shoulder distance or shoulder orientation's axis [63] can not be defined when only end-effector trajectories are used). Since the recordings in our MoCap database only provide us with two of these three main classes of information (standard motion capture devices do not provide mass and force information), we have narrowed the set of possible features to posture and kinematics cues only.

Although there is compelling evidence of the importance of kinematics [62] and form [56] information in emotion perception from body motion, the contribution and relative importance of cues defined from both sources remain unclear. Some authors [149, 117] suggest that form information can be instrumental in the recognition of biological motion and more sophisticated tasks such as emotion perception [63], while motion information may be partially redundant to form information and only used to resolve inconsistencies. Other authors [7, 197, 216] suggest that the kinematics of body motion – either from whole body or specific body parts – are at least sufficient in providing cues for the perception of emotional expressions. They support this assertion through results reported by perceptual studies in emotion perception in which human observers were presented with stimuli for which form information was substantially reduced [7, 197]. This claim seems in agreement with other psychological findings which state that certain emotions such as fear and disgust can not be distinguished if kinematic information is missing [56, 99]. Furthermore, findings by Pollick et al. [197] point also that the activation dimension in the circumplex model of affect [207] seems to relate directly with kinematic information.

A recent study by Atkinson and colleagues [6] determined that observers rely on both form and kinematic information during emotion perception. The authors demonstrated that there is a substantial reduction on emotion classification accuracy from whole-body motion when patch-light and full-light stimuli were inverted and/or played backwards. The effect was stronger on patch-light displays, which attests to the importance of form-related cues in

emotion perception. Nevertheless, even when both stimulus manipulation where combined, observers accuracy was still above chance level. The authors concluded that kinematics information alone can help to distinguish basic emotions from whole-body motions [68]. From this and since form information is considerable diminished when end-effector and pelvis trajectories are used, we decided to characterize the motion sequences in our database through kinematic-related features only.

4.3.2 Motion Sequences as Ensembles of Kinematic Features

Among all the features that can be used to characterize the kinematic qualities of body movements, we have decided to represent motion sequences through *velocity*, *acceleration*, and *jerk*. They have been systematically reported as relevant in the perception of emotions from body movements. For example, Sawada et colleagues [216] showed that arm movements made with the intention of conveying certain emotions mainly varied in their velocity, acceleration, and displacement. Similarly, [59] and [205] observed that changes of velocity in gait motions are affected by emotional states. Moreover, recent studies in automatic classification of affect have reported high accuracy levels for emotion portrayals [129], non-stylized body motions [21], and stylized dance motions [40] with classifiers trained on these kinematic features or on qualitative cues (e.g., Laban notation components) derived from them. It seems then that *velocity*, *acceleration*, and *jerk* are not only good for affect discrimination, but they are also consistent among different body movements (e.g., gait, dance, daily arm motions, etc.). Hence they can be easily applied to the theatrical movements in our database and for both full-body and end-effector trajectories representations.

When characterizing motion trajectories, specially end-effector trajectories, it has been shown that the kinematic qualities (e.g., velocity) along a trajectory are explicitly constrained by the geometric properties (e.g., curvature) of the trajectory itself [252]. For example, in a curved motion, portions with high curvature will entail a reduction of velocity and a change in direction [83]. Both [95] and [208] observed that curvature had a significance influence in the clustering of emotion portrayals and in the generation of affective-like motions for non-anthropomorphic robots respectively. To capture this relation between the kinematic and geometric properties of a motion trajectory, we have included *curvature* in the set of features that characterize our motion sequences.

Mathematically speaking, the features we have chosen have straightforward definitions. *Velocity*, *acceleration* and *jerk* correspond to the first up to third derivatives of position over time. *Curvature*, in turn, measures how fast a curve is turning or equivalently how much a point deviates from following a straight line. Nevertheless, when used to characterize body motion and joint trajectories, they provide us with good approximations of the different motion qualities that might define the underlying emotional states we wish to classify. For example, *velocity* informs us about the level of energy with which a body joint followed its trajectory, with the highest values being related to emotions like joy and anger, while the lowest values correspond to sadness and boredom. In the same manner, *acceleration* and *jerk* account respectively for the smoothness and fluidity of the motion described by any body joint, i.e., if the motion was sudden or sustained. *Curvature*, although being related with the three kinematic qualities we have just described, partially accounts for the form and geometric properties of the trajectory itself.

4.3.3 A Global Classification From A Local Representation

Given a motion $\mathbf{M} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_T)$ of T body postures, each posture $\mathbf{X}_t$ with $1 \leq t \leq T$ is a vector of 3D Cartesian coordinates. Namely, $\mathbf{X}_t = (\mathbf{x}_{i,t}, ..., \mathbf{x}_{n,t})$, where n corresponds to the number of body joints recorded during the MoCap sessions ($n = 27$ in our MoCap database) and $\mathbf{x}_{i,t} = (x_{i,t}, y_{i,t}, z_{i,t})$. Thus $\mathbf{X}_t \in \mathbb{R}^{3\times n}, \forall_t \in \{1, .., T\}$ and $\mathbf{x}_{i,t} \in \mathbb{R}^3, \forall_i \in \{1, .., n\}$. Determining the chosen features for each posture $\mathbf{X}_t \in \mathbf{M}$ consist in computing *velocity* (Eq. 4.2), *acceleration* (Eq. 4.3), *jerk* (Eq. 4.4), and *curvature* (Eq. 4.5) for each one of the n joints in the i-th body posture. Since MoCap data corresponds to regularly sampled time series data, we can approximate each feature by taking the finite differences between neighboring postures according to the following equations:

$$\dot{\vec{x}}_{i,t} = \frac{x_{i,t+1} + x_{i,t-1}}{2\Delta t} \tag{4.2}$$

$$\ddot{\vec{x}}_{i,t} = \frac{x_{i,t+1} - 2x_{i,t} + x_{i,t-1}}{\Delta t^2} \tag{4.3}$$

$$\dddot{\vec{x}}_{i,t} = \frac{-x_{i,t-2} + 2x_{i,t-1} - 2x_{i,t+1} + x_{i,t+2}}{2\Delta t^3} \tag{4.4}$$

$$\kappa_{i,t} = \frac{\|\dot{\vec{x}}_{i,t} \times \ddot{\vec{x}}_{i,t}\|}{\|\dot{\vec{x}}_{i,t}\|^3} \tag{4.5}$$

Before computing all kinematic features, we applied a Gaussian filter of order=0 and $\sigma = 10$ on all joint positions data. We did so as to smooth out most of the noise introduced when markers' gaps were filled and joints' centroids were determined by averaging clusters of markers.

Although we have now a set of descriptors that might account for the underlying phenomenon we seek to classify, i.e., the emotion-related variations in whole-body motions, we have incurred in a considerable increase of dimensionality. In our initial representation, we had that each posture $\mathbf{X}_t$ corresponded to a vector in $\mathbb{R}^{n\times 3}$ with $n = 27$. A motion $\mathbf{M}$ would be then defined in a space $\mathbb{R}^D$, with $D = n \times 3 \times T$ and T being the motion duration and variable among all motions. Now, if we consider the features derived from all n joints, we have that each posture is represented by a feature vector $\mathbf{F}_t$ defined in $\mathbb{R}^{n\times 10}$ – *velocity*, *acceleration* and *jerk* are 3D vectors and *curvature* is a scalar – and a motion $\mathbf{M}$ is now defined in a feature space of dimensionality $d = n \times 10 \times T$, with $d \approx 3 \times D$. Since most of the standard classifiers have problems dealing with datasets for which the number of samples to classify is significantly smaller than the dimensionality of the samples, i.e., $p \ll d$, where p is the number of samples in the dataset, we have adopted a different approach. This new way of generating feature vectors will provide us in addition with the fixed-length samples necessary for the type of classifier we use.

Given a motion $\mathbf{M}$ of length T postures, a *motion chunk* $\mathbf{C}_i \subset \mathbf{M}$ is a sampling of length $w \leq T$ of contiguous postures from $\mathbf{M}$, that is:

$$\mathbf{C}_i = (\mathbf{X}_i, \mathbf{X}_{i+1}, ..., \mathbf{X}_{i+w-1}) \; for \; 1 \leq i \leq T - w + 1 \tag{4.6}$$

Motion chunks are generated through overlapping sliding windows. That is, given a user-defined subsequence length w (in frames) and an overlap ratio r with $0.0 \leq r < 1.0$ [1], all possible *motion chunks* can be built by "sliding a window" across $\mathbf{M}$. A *motion chunk* $\mathbf{C}_i$ of length w is defined every time the "window" has slided $w \times (1 - r)$ frames forward, that is, any two contiguous *motion chunks* are $w \times (1 - r)$ frames apart from each other. A graphical representation of two *motion chunks* being defined and the sliding window approach is presented in Figure 4.1.

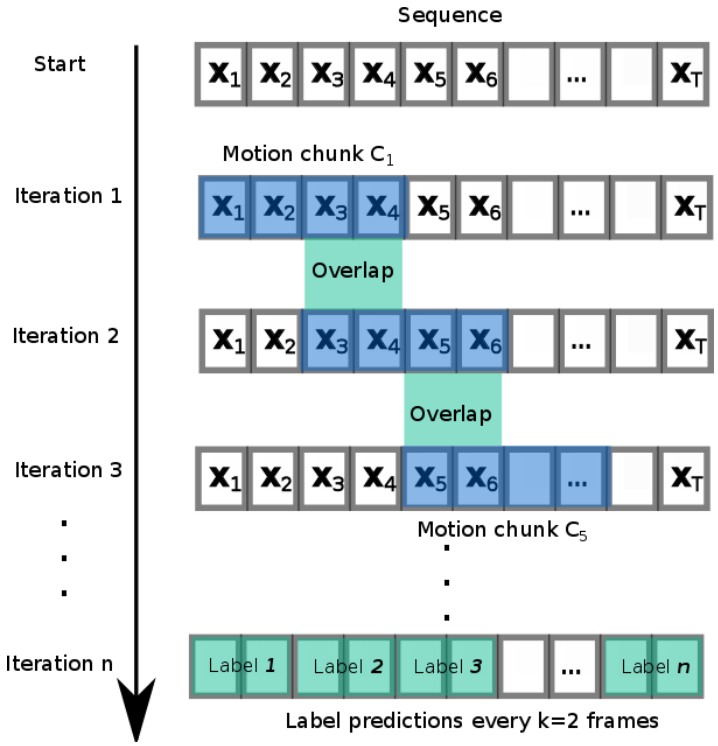


Figure 4.1: *Motion chunks* $\mathbf{C}_1, \mathbf{C}_3, \mathbf{C}_5, ..., \mathbf{C}_{T-4}$ are defined using a sliding window of length $w = 4$ frames and an overlap $r = 0.5$. That is, at each iteration, the "window" (in blue) slides forward $4 \times (1 - 0.5) = 2$ frames along the motion sequence $\mathbf{M}$. Two contiguous *motion chunks* share an overlap (in aquamarine) that is 2 frames long. Prediction can be done at sequence, *motion chunk* and overlap level.

Motion chunks provide us with a partial solution to the dimensionality problem since they increase the number of samples in our dataset. They also solve the problem of classifiers input, since they all correspond to an equal fixed-length vector; the classifiers can then be trained and tested at the *motion chunk* level rather than at motion sequence level. However, the problem of $p \ll d$ (i.e., the number of samples is considerable smaller than the dimensionality of the feature space) remains; in their initial representation, i.e, as a fixed-length subsequence of w contiguous postures, a *motion chunk* is a vector defined in $\mathbb{R}^{n \times 3 \times w}$. If we represent them in feature space, $\mathbf{C}_i = (\mathbf{F}_i, \mathbf{F}_{i+1}, ..., \mathbf{F}_{i+w-1})$, a *motion chunk* is now a feature vector in $\mathbb{R}^{n \times 10 \times w}$. Fortunately, it is still possible to further discretize our feature vectors such that their dimensionality is considerably reduced while preserving the kinematic and temporal patterns characterizing the emotion-related content in the motion examples.

[1] Note that in order to limit the redundancy between successive *motion chunks*, r cannot be large.

Rather than registering the exact values of *velocity*, *acceleration*, *jerk*, and *curvature* during a time interval, we are interested in how these quantities evolve along time. The variations and tendencies of these quantities can be captured by statistical measures such as the mean and the standard deviation. Consequently, we can redefined the representation of a *motion chunk* $\mathbf{C}_i$ into feature space as follow:

- Compute *velocity*, *acceleration*, *jerk*, and *curvature* for all $\mathbf{X}_t \in \mathbf{C}_i$, with $t \in \{i, ..., i + w - 1\}$.
- For all 3D descriptors, compute their magnitudes.
- Compute mean and standard deviation for the magnitudes of each descriptor.
- Define feature vector $\mathbf{c}_i \in \mathbb{R}^{8\times n}$ associated to $\mathbf{C}_i$ from descriptors means and standard deviations.

Considering the feature vectors derived as described above would give rise to a new feature space of dimensionality at most $d = 216$ (27 joints $\times$ 4 features $\times$ 2 statistical measures) if all body joints are taken into consideration. With this approach, we have a fixed-length feature vector which dimensionality is considerably lower than the number of samples in the dataset.

Beyond the "mathematical" reasons that motivated our choice of subsampling each motion sequence into a finite set of overlapping *motion chunks*, this approach also allows us to capture both the short and long terms variations due to the emotion-related characteristics of the motion, while smoothing possible fluctuations coming from noise still present in the data. Furthermore, since we are interested in measuring the suitability of end-effector trajectories for the synthesis of expressive bodily motion, regardless of the semantic elements of the motion, a sliding window supposes no prior knowledge about the structure of the expressive motion data. Furthermore, it will likely smooth out variations related to the transitions between semantic actions while preserving the underlying affective content. Additionally, this new representation preserves the temporal context necessary for the classification of motion sequences and provides us with a means to classify a motion example at different granularity levels.

We recall that a motion sequence $\mathbf{M}$ is associated to a class label l which corresponds to the emotional state elicited when the motion was recorded. All *motion chunks* subsampled from this motion, $\mathbf{C}_i \subset \mathbf{M}$, are immediately associated to the same class label l. Thus, it is possible to make a prediction about a motion sequence's label based on the individual predictions made by a classifier for motion chunks $\mathbf{C}_i$ from $\mathbf{M}$. It suffices to apply a majority vote scheme on $\mathbf{C}_i$'s predictions. Similarly, since overlapping windows are used, it is possible to have $\lfloor \frac{1}{1-r} \rfloor$ predictions every $w \times (1 - r)$ frames and apply the same majority vote scheme on the overlapped segments [4]. With this procedure, we can make predictions on motions sequences as we will do on stream data. Figure 4.1 depicts the different granularity levels in a motion sequence at which it is possible to make predictions about one or several emotional states conveyed along the motion example $\mathbf{M}$. The majority vote scheme used for determining motion and overlap labels will be presented further in this chapter (see Section 4.5).

4.4 Classification Model and Feature Subsets Definition

We aim to quantify how much information might be lost when only features computed from partial-body representations we manually selected (e.g., end-effector trajectories) are used. To do so, we will compare the performance of a classifier trained on: the whole feature set, subsets automatically defined by some feature selection method, henceforward referred as to ***ranked*** group, and subsets manually built, hereinafter called ***adhoc*** group. The idea of comparing ***ranked*** group against ***adhoc*** group comes from common knowledge about the bio-mechanical constraints governing the human body. We know for example that elbow motion is clearly not independent from hand motion [9] or legs ands arm operate in a coordinate way in most cases [210]. These and other constraints suggest that it may exist one or several subsets of joints and consequently of feature subsets, other than the ones we manually defined, that may be equally useful in the construction of a good affect classifier. Automatic feature selection methods are the means we chose to uncover these alternative subsets.

4.4.1 An Overview of Feature Selection

Feature selection aims to choose a small subset of relevant features from the original set according to some relevance measure. e.g., correlation, accuracy, Information Gain, etc. The potential benefits of working with the resulting subset are: an improved prediction performance, lower computational cost, a lower risk at over-fitting the learning models, the dimensionality reduction, a simpler model easier to understand, and better data interpretability [229]. Feature selection methods can be roughly grouped into three main categories: filters, wrappers and embedded methods [107]. *Filter methods* are independent of the classification model. They score features on a particular metric calculated directly from the data (e.g., Pearson's correlation coefficient, entropy, etc.) and seek to filter out those descriptors that seem to be not useful, i.e., those whose scores are below some established threshold [69]. Although filter methods are computationally simple and fast, and can easily scale to very high-dimensional datasets, they exhibit two main drawbacks: *i.)* redundancy between features is not taken into account, thus descriptors that are individually relevant may lead to worse classification performance when compared to other types of feature selection techniques [209]; *ii.)* dependency between features is ignored, two features that are not individually relevant may become relevant together [107]. Contrary to filter methods, both *wrapper* and *embedded* techniques make use of the classification model during feature subset determination and take into account feature dependencies. However they do it differently. In one hand, *wrapper methods* aim to find a subset of features that maximize the classifier's performance. The selection algorithm searches for a good subset of features using the classification model as a black box that evaluates the usefulness of possible optimal subsets [107]. Thus, wrapper methods are usually tailored to a specific classification algorithm [209]. In the other hand, *embedded methods* do not separate the learning process from the feature selection part, hence the internal structure of the classifier plays a crucial role on the search of an optimal subset of features [148]. For example, decision trees such as Classification and Regression Tress-CART [36] have a built-in mechanism to perform variable selection [107]. Both wrappers and embedded methods interact with the target classification algorithm during feature selection and potentially achieve better results since they do not make assumptions of fea-

ture independence as filter methods do.

We are interested in automatically generating subsets of features, and implicitly subsets of body joints, useful for both the construction of a good affect classifier and the quantitative comparison with our manually selected subsets. Only after comparing the ***adhoc group*** against feature subsets containing highly informative predictors, we will be in measure of evaluating its suitability. Thus we focus our interest on *wrappers* and *embedded methods*, since they both consider the interaction with the classification model during the selection process. Both *wrappers* and *embedded methods* share the ability of taking into account feature dependencies thus yielding more compact subsets of features. They also have a higher risk of over-fitting when small amounts of training data are available [209]. They potentially suffer of high variance and small changes in the training data lead to large changes in the resulting feature subset [244]. However, there are also strong differences between them. *Wrapper methods* are slower and more computationally intensive, since they typically need to evaluate a cross-validation scheme at every interaction in order to obtain an estimate of the usefulness of a possible optimal subset [148]. *Embedded methods*, in the contrary, are more efficient and make a better use ot the available data by not needing to split the training data into training and validation sets [107]. Although our strategy of generating *motion chunks* from all the examples in the MoCap database allow us to increase the number of samples in the dataset, due to the temporal correlation and dependency among chunks issued from the same sequence, we are constrained to split the data into training and testing set at the level of motion sequences instead of at chunk level. Thus, we are still limited by the amount of available data and we will benefit from a feature selection method that makes the most efficient use of it. Hence, we have decided to use an *embedded method* coupled with a forward selection strategy [148] for the definition of the subsets of features belonging to what we have referred to as ***ranked group***.

Among the embedded methods, decision trees such as CART are one of the most successful and thus attractive techniques at our disposition. Decision Trees can model arbitrarily complex relations, as it is the case for body motion and emotions, without prior assumptions, handle heterogeneous and noisy data, and deal with feature redundancy [162]. They are scale invariant and robust to outliers, missing values and errors in labels. Nonetheless, decision trees, as all embedded methods, are prone to high-variance and thus a change in the data will produce an entirely different model. Fortunately, feature selection is a special case of the model selection problem and thus it benefits from the regularization effect produced by ensemble methods [244]. That is, by combining a set of simple base learners, e.g., decision trees, the strengths of the base learners are preserved while reducing their variance and their instability. *Random Forest*, henceforth RF, is one of the most suitable and up-to-date ensemble method to be applied on decision trees, since its learning principle seems to work especially well for high-variance, low-bias models, such as trees [110]. RF is not only known for its ability to provide robust and accurate models [245], but also has previously produced good results in the recognition of affect from bodily motions [88, 260]. Most importantly, RF intrinsically preserves the feature selection mechanisms of decision trees and can be used both as a feature selection method and as a classifier. For these reasons we chose to use RF as both feature selection method and classifier.

4.4.2 Empirical Reasons for Using Random Forests

Before explaining in detail the principles and mechanisms behind the *Random Forest* learning algorithm, we would like to present some experimental results and additional practical reasons for which we chose RF not only as the method to use in the definition of ***ranked*** feature subsets but also as our automatic affect classifier model.

In the results we published in [44], we compared the performances of two different classifiers models *Linear Discriminant Analysis* (LDA) and SVM when trained either on whole-body features – all 216 features – or on features computed end-effector trajectories only, that is, only 40 features were considered. We obtained similar performances for both models, however a more detailed inspection of LDA's behavior when the model complexity was iteratively increased (see Figure 4.2) showed that the difference between the training and testing error rates raised with the number of features considered in the representation of the learning samples. Thus, LDA seems to over-fit the training data and generalize poorly to the independent testing set. This might be due to the reduced number of samples compared to the number of predictors representing them. In the contrary, as will be explained in the next section, two of the most attractive features of RF model are its low-variance and its capacity to handle high-dimensional data.

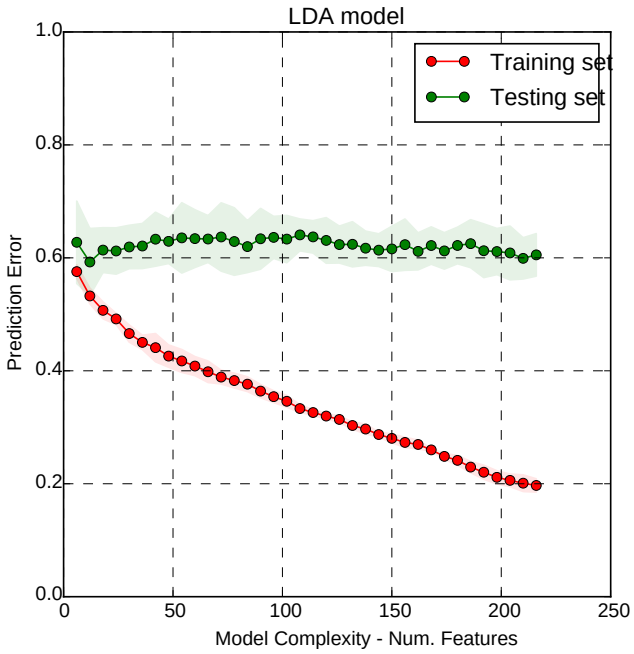


Figure 4.2: Behavior of the test sample (green curves) and training sample (red curves) error rates as LDA's model complexity, i.e, number of predictors, is increased. Results correspond to motion chunk predictions and were estimated using 10 repetitions of stratified 5-fold cross-validation on LC actor's data.

Although SVM has been successfully used in the classification of affect from body motion cues [21, 130], a good SVM classifier depends on the choice of its hyper-parameters [19], e.g., kernel function, cost parameter C, etc. In the context of our application, this implies that for

each subset in either ***adhoc*** or ***ranked*** group, an additional internal parameter search should be performed before estimating the classifier's generalization error. As we have already mentioned, although the *motion chunks* strategy increases the number of observations in the learning set, there is a temporal correlation to be considered when splitting the observations into training and testing sets. Thus, the amount of data we have at our disposition might not be sufficient to do both parameter search and generalization error estimation. RF, conversely, has been reported to be less sensitive to its hyper-parameters than SVM and good results can be obtained with default values [158]. Finally, being based on decision trees, RF is much more easier to interpret than SVM.

4.4.3 Random Forest and Feature Subsets

Proposed by L. Breiman in 2001 [34], RF is a very popular and effective learning model that has been repeatedly reported to be: (a) extremely successful as a general purpose classification and regression approach, (b) easy to train and tune, and (c) able to deal with small sample sizes and high-dimensional feature spaces without overfitting [23]. The three main essential ingredients in *Random Forest* are: bagging, the CART-split criterion and un-pruned randomized trees. Bagging or bootstrap aggregation [35] is a general aggregation technique which generates bootstrap samples (i.e., samples are randomly drawn with replacement) from the original dataset, builds a base learner for each sample, and use their averaged vote to predict new data. The CART-split criterion is used in the construction of the individual randomized trees to choose the best binary split. At each node of the tree, the best split is selected by optimizing an impurity measure. The smaller the impurity index associated to each node t, the purer the node and the better the predictions for all samples falling into t. Thus, when building a tree in the forest, RF learning algorithm seeks to maximize the impurity decrease which is equivalent to minimize the generalization error [162]. Following the nomenclature and definitions given in [162], the decrease of impurity of a binary split on feature f_i is formally defined as:

$$\Delta i(f_i, t) = i(t) - p_{Left} i(t_{Left}) - p_{Right} i(t_{Right}) \tag{4.7}$$

where p_{Left} (resp. p_{Right}) corresponds to the proportion $\frac{N_{t_{Left}}}{N_t}$ (resp. $\frac{N_{t_{Right}}}{N_t}$) of learning samples ($N_t$) falling into node t and going either to the left ($N_{t_{Left}}$) or to the right ($N_{t_{Right}}$), and $i(t)$ denotes an impurity measure. The most common impurity criterion used for growing trees inside a RF is the *Gini index*:

$$i_G(t) = \sum_{k=1}^{|\mathbb{L}|} p(l_k \mid t)(1 - p(l_k \mid t)) \tag{4.8}$$

where $\mathbb{L}$ is the set of possible labels, $|\mathbb{L}|$ corresponds to its cardinality, and $p(l_k \mid t) = \frac{N_{l_k}}{N_t}$ is the probability of label l_k in node t. The *Gini-based* impurity criterion $i_G(t)$ measures how often a randomly chosen sample would be incorrectly classified if it were randomly labeled according to the distribution $\mathcal{P}(\mathbb{L} \mid t)$ [162]. Particularly, it is zero when node t has observations only from one class, and it attains its maximum when classes are perfectly equally [245].

In standard trees, each node is split using the best optimized split along all d features. Once the tree has grown to the desired depth, its nodes are pruned. That is, nodes that degrade the generalization error – estimated in an independent validation set – are sequentially removed until the optimal tree is found [162]. In a randomized tree, as the ones used as base learners in RF, each node is split using the best among a subset of predictors randomly selected at that node. Similarly, trees are grown un-pruned [158].

An additional characteristic of RF is its out-of-bag (OOB) error. For each bootstrap training sample, and consequently each randomized tree, about one third of the observations in the dataset are left out and can be considered as an internal validation set [30]. The OOB error is simply the average error obtained when the observations in the training dataset are predicted using the trees for which they are OOB. See Figure 4.3 [2] for a graphical description of the RF internal learning algorithm.

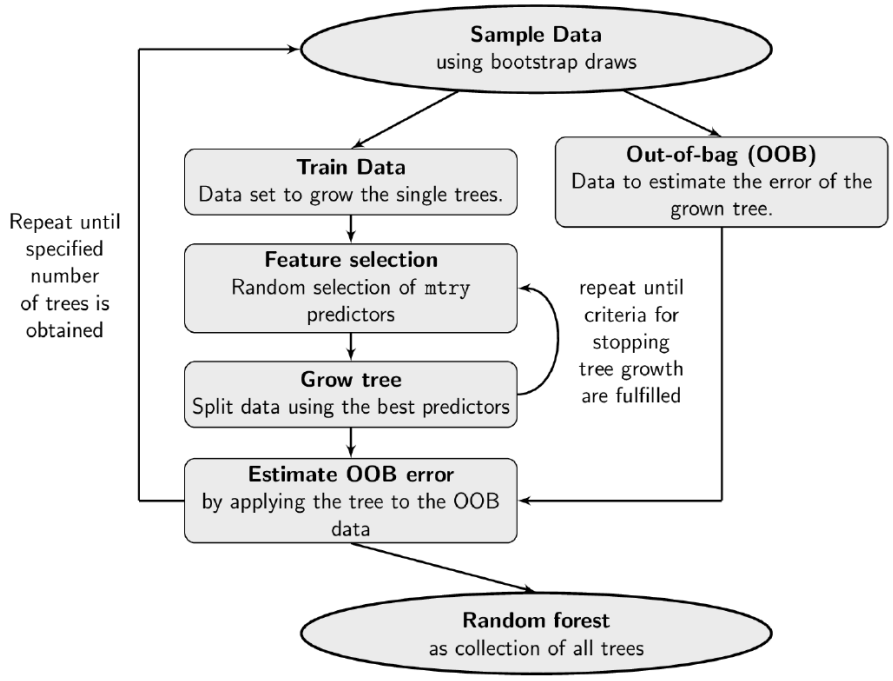


Figure 4.3: Graphical representation of the RF model's learning algorithm taken from [30]. Note how all main elements (bagging, split-criterion, randomized tree and OOB) are combined together in the construction of the resulting RF.

Recall that at each step in the construction of either a standard or randomized tree φ_h, an exhaustive search for the split that achieves the maximum impurity decrease is done. Under this procedure, the impurity decrease due to a split on a specific feature f_i for $1 \leq i \leq d$ indicates the relative importance of f_i in the tree φ_h. In the context of RF, Breiman proposed to evaluate the importance of a feature f_i for the prediction of the label l by computing the weighted impurity decrease $p(t)\Delta i(f_i, t)$ for all nodes where the feature f_i was used as optimal split, averaged over all trees φ_h (for $h \in \{1, ..., H\}$, with H as the total number of trees in the RF model) in the forest [162, 245]. Formally and following the definition provided

[2] Consent to use this figure was granted by John Wiley and Sons under license 3861801463020.

in [162], the measure of a features's importance (FI) is defined as follows:

$$FI(f_i) = \frac{1}{H} \sum_{h=1}^{H} \sum_{t \in \varphi_h} 1(f_t = f_i)[p(t)\Delta i(f_i, t)] \tag{4.9}$$

where $p(t)$ is the proportion $\frac{N_t}{N}$ of observations reaching node t and $1(f_t = f_i)$ denotes if feature f_i was used for splitting node t. If the *Gini index* is used as impurity criterion, this measure of variable importance is called *Mean Decrease Gini*, henceforth MDG.

RF implicitly provides us with a relative feature ranking, based on Eq. 4.9, that incorporates interactions of any order between features. However, in order to produce smaller predictor subsets, it is necessary to combine RF's ranking with either a backward or forward selection strategy (i.e, feature are iteratively eliminated or added until the optimal feature subset is found) [98]. We have decided to use a forward selection approach based on the work done by [91]. In this approach, MDG variable importance is computed only at the initialization of the algorithm and then the predictors are iteratively added according to their relative importance until all d variables are considered. Note that we aim to automatically generate subsets of the most important predictor and of the same cardinality than those belonging to the ***adhoc group***. The approach can be summarized as follows:

- Compute the RF scores of variable importance and rank variables according to it.
- Define a feature subset using the k first most important variables, for all $k \in \vartheta_r$.

where $\vartheta_r = \{1, 16, 24, 32, 40, 48, 64, 80, 128, 216\}$ corresponds to the set of cardinalities we want to consider in the ***ranked group***. Correspondingly, the ***adhoc group*** comprises only features computed from the body joints listed in Table 4.1 (ϑ_a denotes the cardinalities of these subsets).

ϑ_a	Features from:
24	Head, hands
32	Head, hands, pelvis
40	Head, hands, toes
48	Head, hands, toes, pelvis
64	Head, elbows, hands, pelvis, toes
80	Head, elbows, hands, pelvis, knees, toes

Table 4.1: Joint and number of features for ***ad-hoc*** group.

The subsets in ***adhoc group*** were defined by iteratively adding body end-effectors and other joints which might help us to constrain the synthesized bodily motions to the space of plausible body postures. The synthesis methods propose in this thesis will be described in Chapter 5.

4.5 Sequence and Overlap Predictions

As mentioned earlier, the *motion chunk* strategy described in Section 4.3.3 allow us to make emotion predictions at three deferent levels: sequence, *motion chunk*, and overlap. Making predictions at sequence level is necessary if we want to compare the classifier's performance with the base line provided by the perceptual studies described and analyzed in Chapter 3. Indeed, in both studies annotators were asked to rate the emotional content for the whole motion sequences rather than for individual actions. Similarly, analyzing the classifier's recognition rates at the overlap level (as we will see in Section 4.8.1) provides a continuous stream of emotion predictions that give us some insight into whether the elicited emotion was sustained and effectively conveyed along the sequence.

To predict a sequence's label, we need a means to combine the predictions made for all *motion chunks* extracted from the same sequence. Similarly, to predict an overlap's label, the predictions of all the $\lfloor \frac{1}{1-r} \rfloor$ (for $0.0 \leq r < 1.0$) contiguous *motion chunks* for which $w \times (1-r)$ frames (where w denotes the sliding window's length) are common need to be considered. Loosely speaking, the majority vote scheme we propose to use for making label predictions at sequence level consists on averaging the confidence levels provided by the RF classifier across all the *motion chunks* belonging to the same motion sequence. The class label associated to the highest confidence level is taken as final prediction. Overlaps' predictions are determined in the same way. The only difference is that for each overlap, we only consider the *motion chunks* common to it.

Formally, the majority vote scheme we have adopted can be formulated as follows: suppose that *motion chunks'* predictions correspond to the decisions made by a set of experts $\mathbb{E} = \{\mathcal{C}(\mathbf{C}_1), ..., \mathcal{C}(\mathbf{C}_P)\}$ (where P is the number of *motion chunks* in a sequence $\mathbf{M}$ and $\mathcal{C}(\mathbf{C}_i)$ stands for the prediction made by the classifier $\mathcal{C}$ about the *motion chunk* $\mathbf{C}_i$) about the class $l \in \mathbb{L}$ to which a sequence $\mathbf{M}$ should belong. Based on the literature about classifier fusion, depending on the type of output provided by the experts, there are different strategies to combine their decisions [269]. *Random Forest* can supplies us with what [269] denoted as measurement level or type 3, that is, it produces a $|\mathbb{L}|$-dimensional vector $[\omega_{i,l_1}, ..., \omega_{i,l_{|\mathbb{L}|}}]^T$ in which ω_{i,l_k} represents an approximation of the degree of belief (or confidence level) the feature vector $\mathbf{c}_i$ computed from the the i-th *motion chunk* $\mathbf{C}_i \subset \mathbf{M}$ comes from the class l_k. Given these measures, it is possible to determine a new estimate or prediction common to all experts, i.e., to all *motion chunks* as follows [269]:

$$[\omega_{\mathbb{E},l_1}, ..., \omega_{\mathbb{E},l_{|\mathbb{L}|}}] = \left[\frac{1}{P} \sum_{i=1}^{P} \omega_{i,l_k}, \; with \; k = 1, .., |\mathbb{L}| \right] \tag{4.10}$$

Thus, the final decision made by the set of experts $\mathbb{E}$, and consequently the label associated to the sequence $\mathbf{M}$ from which all P *motion chunks* were extracted, is given by:

$$\mathbb{E}(\mathbf{M}) = l^* = \arg\max_{l \in \mathbb{L}} (\omega_{\mathbb{E},l}) \tag{4.11}$$

Decisions about overlaps' predictions are made in the same way. The only difference is that for each overlap, the set of experts $\mathbb{E}$ will comprise only predictions of the *motion chunks* common to it.

4.6 Classification Tasks

Between the first and second set of MoCap recordings, we introduced significant differences in the emotion elicitation procedure, the capture protocol and the actors' recruitment process. Thus, in order to test the classifier under homogeneous conditions, all the affect classification tasks to be described focus on the motion data recorded during the second Mocap sessions. The data is as follows:

- 5 actors (3 men and 2 women)
- 5 distinct motion examples: three magician sequences, one walk, and one free improvisation exercise. All sequences have different time durations.
- 3 trials for each magician sequence. A single trial for the other two types of motion sequences.
- Trials were recorded for each of the 5 emotion classes: happiness, sadness, relaxedness, neutral, stressed. Thus, our learning dataset comprises in total 275 motion sequences.

Each one of the 5 recorded actors in our dataset can be considered as a *Random Process* from which some instances[3] of expressive bodily motion are known. Thus, each motion sequence corresponds to one of these known instances. According to the way in which the learning dataset is partitioned into training and testing sets, we can define different emotion classification tasks [130]. Each task will evaluate, under different settings, the suitability (i.e., how much expressive information is preserved) and generalization capabilities to different subjects and movements of the motion model we propose in this thesis. We distinguish three possible main tasks, with two of them being further divided into two sub-tasks:

- **Single subject or person-dependent recognition**: the motion data produced by an actor is considered as a whole learning dataset in itself, i.e., motion observations for the same subject are included in both the training and testing sets. This is the simplest task, since it allows the classifier to learn the style in which the subject expressed each emotional state. This task is performed for all 5 subjects.
- **Within subject classification**: we take all the subjects together as learning dataset, i.e., motion instances for all the subjects are included in both the training and testing sets. This might be a significantly more difficult task, since the classifier will be exposed to all the within class variability due to the actor. Although actors were asked to perform the same type of sequences and submitted to the same elicitation procedure for the same target emotion, it is very likely that depictions of the same emotional state vary considerably from actor to actor.
- **Between subject classification**: also known as leave-one-subject-out. All the data produced by a single subject is considered as test set and the classifier is trained on the remaining subjects' data. This is the hardest task, since we ask the classifier to recognize the affective state of a completely unknown actor. This task not only evaluates the overall generalization capability of the low-dimensional model we propose, but is also

[3] Also referred to as observations or samples

the closest to the perceptual study described in Chapter 3 and used as base line. We can run 5 instances of this classification task since our learning dataset contains data from 5 different actors.

Our database comprises 5 different motions sequences, 3 of them count with several repetitions by emotional state, belong to the same semantic context, i.e., a magician's performance, and share up to 3 individual actions. Furthermore, due to the temporal correlation between the *motion chunks* issued from the same motion sequence, training and testing dataset partitions are done at sequence level. That is, cross-validation splits are defined on a learning dataset with 275 instances only. In this context both single and within subject classification tasks can be further divided as follows:

- **One sequence out**: all the repetitions for all emotional states of a single sequence example are considered as testing set. Each classifier is then trained on the examples of the other 4 sequences. We rotate through all sequences, which results in 5 instances of either the single or within subject classification task. This partitioning of the learning dataset allows us to measure, to some extent, how well a classifier trained on either a subset from the ***ranked group*** or ***adhoc group*** generalizes to unseen types of actions and/or sequences. In other words, we measure the movement-dependency of the classifier.
- **One repetition out**: one example of each sequence for affective state makes part of the testing set. This is the simplest task setting at hand, since all subjects and sequence types are seen by the classifier during its training. Together with the perceptual study, it provides a base line for evaluating and comparing the classifier's performance in the other more difficult and complex tasks.

To summarize, both ***adhoc*** and ***ranked groups*** will be compared in 5 different classification tasks and their performance will be measured at 2 different levels (see Table 4.2). These tasks measure not only how much information might be lost when manually selected subset of body joints are used to characterize expressive bodily motion, but also the action-based and actor-based generalization capabilities of both subsets groups. Furthermore, we can also quantitatively evaluate the relevance of our database and the type of actions (inspired from physical theater) within it.

Main Task	Sub-tasks	Performance levels
Single subject recognition	One sequence out	Sequence
Within subject recognition	One repetition out	*Motion chunk*
Between subject recognition	None	

Table 4.2: Different classification tasks and dataset granularities on which ***adhoc*** and ***ranked groups*** will be compared.

4.7 Experimental Setup

In this section we will describe in detail how the emotion recognition framework we have just introduced is evaluated. The classifier's performance will be evaluated using recogni-

tion rates, also called accuracy rates, on a separate testing set. Since we do not count with a large learning set, we will use repeated cross-validation for the definition of the training and testing sets as well as for the estimation of the expected test accuracy of our classifier. Depending on the task at hand, we will use either stratified K-fold cross-validation [4] on class labels or leave-one-out cross-validation [5] on sequence types and/or actors.

We signal to the reader that we are using cross-validation as the means to asses the performance of a same classifier algorithm trained on different variable subsets rather than to fit and choose the best model. We do so for the following reasons: *(a)* our aim is to measure the differences in performance when using predictor subsets defined either automatically or manually, not to select the best affect classification model, and *(b)* we dispose of a very limited amount of data and to perform both model selection and model assessment will require to carry out a nested cross-validation procedure [143] in which the training set might be too small to characterize the subtle variations related to affect expression.

In their book *The Elements of Statistical Learning* Hastie et al. [110] provided general guidelines in regard to the correct way to carry out cross-validation in a context similar to ours. We have adapted their algorithm to our specific case. The exact procedure is presented below:

In this procedure, the way in which the learning dataset (denoted as Ω in the algorithm) is divided into K-folds depends on the classification task being considered. For assessing the classifier performances in both *single subject* and *within subject* recognition, we apply stratified K-fold cross-validation when working on the *one-repetition-out* sub-task as suggested by [143]. Specifically, we stratify the dataset according to emotion labels and sequence types. The dataset is then randomly split into K folds in such a way that each fold contains the same proportion of the different stratified tuples. We found that by setting the number of folds equal to 3, we will obtain a test set with an example by emotion and by sequence type for either one actor or all of them. In the same way, when working on *between subject recognition*, or *one sequence out* for either one or all subjects, we use a leave-one-out cross-validation split pattern. The dataset will be split according to the actor or sequence type categories. In both cases, this division results in 5 folds, since there are 5 different actors and 5 distinct sequences (see Table 4.3 for a summary of the cross-validation schemes used in each classification task).

An additional step we consider before computing feature rankings and training affect classifiers is feature standardization. Since we are working with different higher order derivatives of position or geometric quantities, each feature $f_i \in \mathbf{c}_j$ have a different order of magnitude. We use the classical centering and scaling procedure: $f_i' = (f_i - \mu_i)/\sigma_i$, where μ_i and σ_i are the mean and standard deviation of feature f_i over the training observations. The standardization 'learned' on the training set is then applied to the test observations previous classification.

The classification tasks summarized in Table 4.2 are evaluated and analyzed in four different contexts. There are:

1. **Emotion categorization from different feature subsets:** we consider the recognition

[4] The learning data is partitioned into K equal sized fold such that each fold contains roughly the same proportions of class labels. A single fold is retained as the test data and the remaining K-1 folds are used as training data.

[5] All instances of a given sequence (e.g., improvisation sketches) or actor are used as testing data. The remaining 4 sequences/actors' data is used as training data

Data: Ω: original dataset, ϑ_r: cardinalities in ***ranked group***, Γ: feature subsets in ***adhoc group***, *n_rep*: number of runs of cross-validation procedure, d: original dataset dimensionality
Result: Average accuracies and confusion matrices for all elements of ϑ_r and Γ
for *J from 1 to n_rep* **do**
  Divide dataset Ω into K-folds;
  for *I from 1 to K* **do**
    Define set T as dataset Ω with the I-th fold;
    Define set S as the I-th fold of the dataset X;
    Build RF model $f' = f(T, d)$;
    Compute ranking of all d features in Ω;
    for *all* $\theta \in \vartheta_r$ **do**
      Define T' as set T with only first θ features selected from ranking;
      Define S' as set S with only first θ features selected from ranking;
      Build RF model $f' = f(T', \omega)$;
      Apply f' on S' and store accuracy and confusion matrix;
    end
    Calculate average accuracies and confusion matrices for all elements of ϑ_r across folds;
    Repeat the same procedure for all γ subsets in Γ;
    Calculate average accuracies and confusion matrices for subsets in Γ across folds;
  end
  Calculate average accuracy and confusion matrices for both ϑ_r and Γ across all repetitions
end

Algorithm 2: Cross-validation procedure used for comparing ***adhoc*** and ***ranked groups*** on each one of the classification tasks listed in Sec. 4.6

rates of the same classifier model, i.e, RF, trained on each of the subsets in ***adhoc*** and ***ranked groups***. The goal is to determine whether the first part of the hypothesis: *"The accuracy rates of a classifier trained on features computed from the proposed low-dimensional motion representation are close to the rates obtained from the same classifier trained on features computed from the entire body and/or automatically selected feature subsets, i.e.,* ***ranked group****"*, holds and to quantify the possible loss of affect-related information we might have introduced when considering features from end-effector trajectories only.

2. **Comparison with human performance:** Emotional states are abstract concepts that are highly subjective and require human validation. In this experiment human evaluation is used as a basis for accuracy comparison. We aim to determine two things: (i) whether our classifier, independently of the dataset representation employed during learning, performs at least as well as human annotators did/do, and (ii) whether humans can still recognize affect from impoverished body motion representations as the one obtained when only end-effector trajectories are visualized.

3. **Sliding window parameters effects:** We use overlapping sliding windows, and consequently what we referred to as *motion chunks*, as a means to capture both the local

Task	Cross-Validation Scheme
Single subject recognition: *one sequence out*	Leave-one-out (5 folds)
Within subject recognition: *one sequence out*	Leave-one-out (5 folds)
Single subject recognition: *one repetition out*	Stratified 3-folds
Within subject recognition: *one repetition out*	Stratified 3-folds
Between subject recognition	Leave-one-out (5 folds)

Table 4.3: Cross-validation schemes used to split learning dataset into training and testing sets for each classification task.

and global patterns encoding the emotional content present in a sequence. They also provide a temporal context for the classification of the motion examples. However, this approach depends on two parameters: the window size w and the overlap percentage r. A window too small might not be able to register the long-term fluctuations related to emotion expression, whereas a window too large will considerably reduce the number of samples in the learning dataset and important information might be lost when feature discretization is applied. In the same way, a low overlap percentage might not provide enough context for sequence classification, while a high percentage will produce *motion chunks* too similar between them and consequently overestimate the true classifier accuracy. We seek to study the effect of these two parameters in the classifier recognition rates.

4. **Model hyper-parameters effects:** There are two main hyper-parameters to specify when working with RF: (1) the number of trees to grow in the forest, *n_trees*, and (2) the number of features to consider in the search of a node's optimal split, *m_try*. In this experiment we aim to analyze how sensitive are both the features subsets generated using RF variable importance ranking and the estimated test accuracies to these hyper-parameters.

4.8 Results

In this section we present the results obtained for the 2 of the 4 experiments outlined in Sec. 4.7. Results obtained for the analysis of sliding window and RF parameters are presented in Appendix B and Appendix C respectively. All experiments were carried out using the free machine learning library scikit-learn [219] and the RF implementation developed by [162]. For all experiments we will mainly analyze the average micro accuracy rate obtained from all subsets in both ***adhoc*** and ***ranked groups***. These average rates have been estimated through ten repetitions of the cross-validation schemes presented in Table 4.3. When studying the effect of both RF and sliding window parameters, we will focus on between subject and within subject *one sequence out* classification tasks, since they both account for the generalization capabilities of the end-effector motion model we propose. We will also observe if a

classifier trained on feature subsets coming from both ***adhoc*** and ***ranked groups*** is invariant to the differences among the sequences and actors in the learning set.

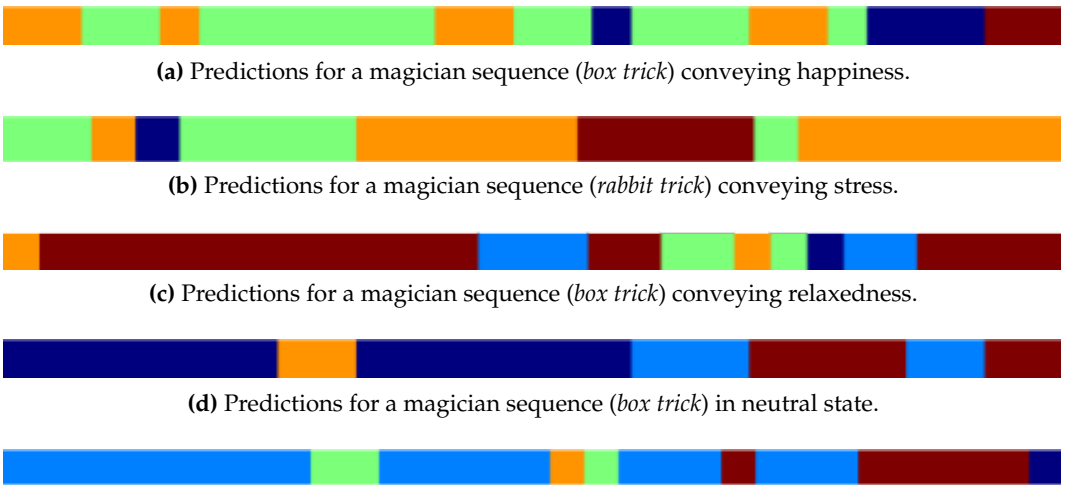
4.8.1 Emotion Categorization from Different Feature Subsets

For this experiment *motion chunks* and their corresponding feature vectors, $\mathbf{c}_i$, were computed using a window of $w = 500$ frames length and an $r = 50\%$ overlap. Both parameters were defined based on our knowledge about the MoCap database we used. An 500 frames window represents the average duration of an individual action in the magician sequences (approx. 2.5 s.), and an 50% overlap provides enough temporal correlation for the classification of a sequence, while making contiguous *motion chunks* different enough from each other. Results were computed using a RF model with default hyper-parameters ($ntrees = 500$, $mtyr = \sqrt{d}$) [34]. For the estimation of ***ranked*** subsets, d was equal to the initial dimensions of the motion chunks, i.e., 216 features. However, during classification, d changed according to the cardinality of the feature subset being evaluated. Figure 4.5 shows the average behavior of test sample accuracy for both *motion chunks* and sequences as the feature subsets representing the observations on our MoCap database changed.

At a large scale, we notice that emotion recognition performance for *motion chunks* was systematically lower than for sequences, independently of the classification task and the feature subset group. If we analyze the stream of emotion labels, as shown in Figure 4.4, generated for a sequence through the classification of all overlapping segments, we observe that although the conveyed emotional state takes precedence over the other possible emotions, there are moments along the sequence in which the classifier judged that a change of emotion took place. This implies that, although coming from the same motion sequence, *motion chunks* placed near and along the regions in which a change of emotion might have taken place had a higher chance to be misclassified. Thus, it seems reasonable that the classifier's accuracy on *motion chunks* is always lower compared to the classification of whole motion sequences. Also, averaging the confidence interval values along the sequence filters part of the misclassification noise.

The continuous change of the classifier's predictions along the same motion sequence suggests that although actors were encouraged to maintain the same emotional state for the whole sequence duration, there were moments in time for which their affective state unconsciously changed or drifted away. Furthermore, since the motion sequences in our database are considerable long and combine several distinct actions, it also possible that: (a) some temporal segments are characterized by few or none body motion activity – such as the preparation of an individual action or the transition between two different actions –, and hence the classifier tried to attribute an emotional state to a non-expressive segment; (b) or our feature-based representation failed to capture and separate the affect-related variations present in some *motion chunks* from the semantic-related content of the action being played. A thorough analysis of this scenario is done in Section 4.8.4.

If we look back at Figure 4.5 (results for ***adhoc*** and ***ranked groups*** are presented at the left and right respectively), we observe also how accuracy rates decrease while the complexity of the classification tasks increases. Note for example how for both feature subset groups the recognition rates on sequences varies from 0.78 ± 0.01 (Figures 4.5a and 4.5b) in the simplest task (single subject *one repetition out* classification) to 0.46 ± 0.02 (***adhoc group***, Figure 4.5i) and 0.48 ± 0.02 (***ranked group***, Figure 4.5j) for the most difficult task (between subject clas-



(a) Predictions for a magician sequence (*box trick*) conveying happiness.

(b) Predictions for a magician sequence (*rabbit trick*) conveying stress.

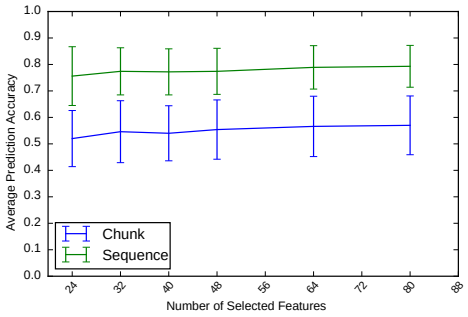
(c) Predictions for a magician sequence (*box trick*) conveying relaxedness.

(d) Predictions for a magician sequence (*box trick*) in neutral state.

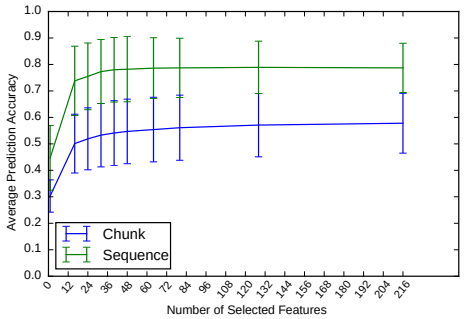
(e) Predictions for a magician sequence (*rabbit trick*) conveying sadness.

Figure 4.4: Stream of predictions generated for five different sequences for LC actor. Color code is as follows: *neutral:* dark blue, *sadness*: light blue, *happiness*: green, *stress*: orange, and *relaxedness*: brown.

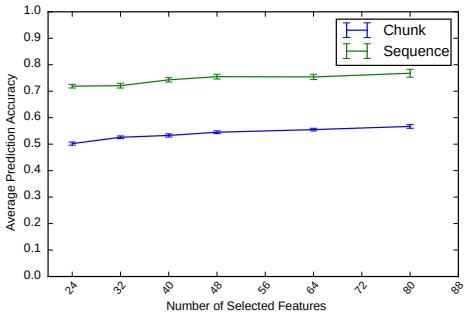
sification). However, it is important to notice that the average recognition rates obtained for both ***adhoc*** and ***ranked groups*** in the latter task are consistent with previous studies (approx. $45\% - 50\%$) [20, 130] for which no action or actor identification was performed ahead of emotion recognition. Similarly, our results are comparable to the human recognition rates reported in Chapter 3 and discussed in detail in Section 4.8.5.



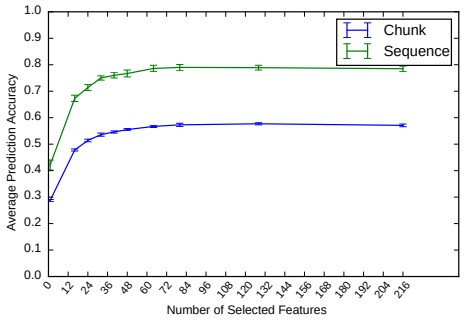
(a) Single subject *one repetition out* for ***adhoc group***



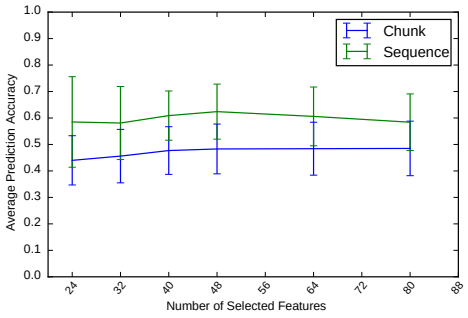
(b) Single subject *one repetition out* for ***ranked group***



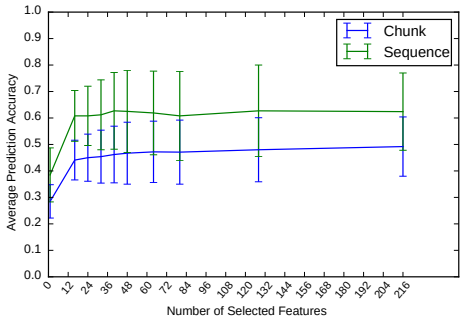
(c) Within subject *one repetition out* for ***adhoc group***



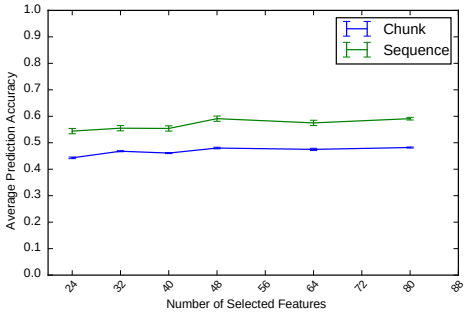
(d) Within *one repetition out* for ***ranked group***



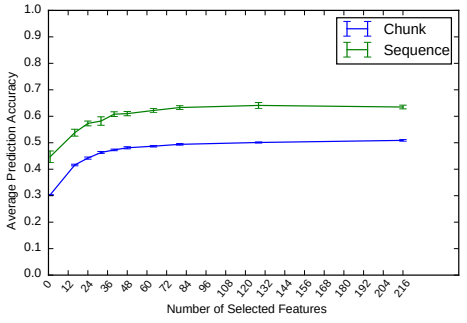
(e) Single subject *one sequence out* for ***adhoc group***



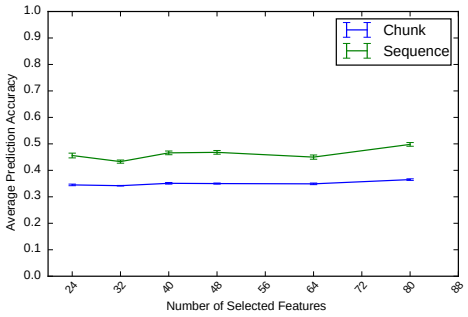
(f) Single subject *one sequence out* for ***ranked group***



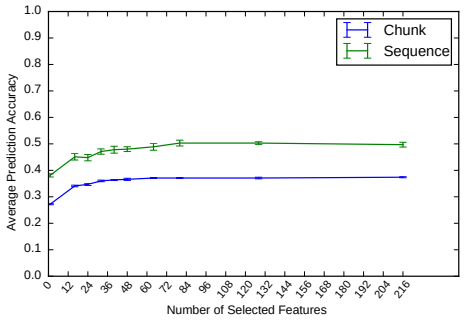
(g) Within subject *one sequence out* for ***adhoc group***



(h) Within subject *one sequence out* for ***ranked group***



(i) Between subject for ***adhoc group***



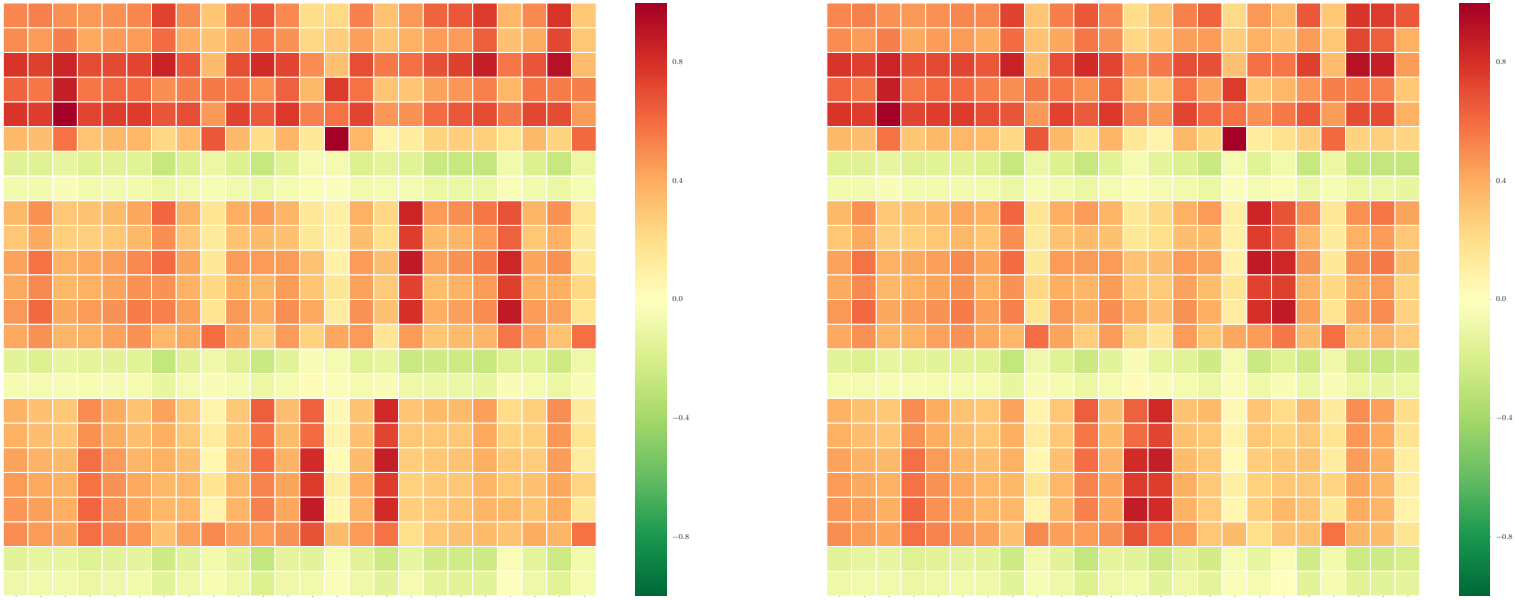
(j) Between subject for ***ranked group***

Figure 4.5: Behavior of test sample accuracy for both *motion chunks* (blue curves) and sequences (green curves). Accuracy rates are presented for each one of the classification tasks defined in Sec. 4.6. Results for ***adhoc group*** are presented on the left. Similarly, results for ***ranked group*** are presented on the right.

Overall, we observe in Figure 4.5 that the performance of a classifier trained on feature subsets defined in the ***adhoc group***, i.e., using end-effector trajectories, is relatively close to the performance of the same classifier while trained on the subsets defined through the procedure described in Section 4.7, i.e., ***ranked group***. As shown in Table 4.4, for four of the five classification tasks considered in this experiment, the average accuracy for all subsets in ***adhoc group*** is within one standard deviation of the mean accuracy rate reported for the subsets of same cardinality belonging to the ***ranked group***. For the remaining task, i.e, within subject *one-sequence-out*, the accuracy rate of the ***adhoc group*** is within two standard deviations of its counterpart's mean accuracy rate.

Task	Level	Avg. accuracy ***adhoc***	Avg. accuracy ***ranked***	Relative difference
Single subject *one repetition out*	*Motion chunk*	0.549 ± 0.018	0.542 ± 0.015	0.007 ± 0.005
	Sequence	0.776 ± 0.013	0.777 ± 0.012	0.005 ± 0.003
Within subject *one repetition out*	*Motion chunk*	0.538 ± 0.023	0.548 ± 0.022	0.010 ± 0.002
	Sequence	0.743 ± 0.020	0.761 ± 0.027	0.020 ± 0.010
Single subject *one sequence out*	*Motion chunk*	0.471 ± 0.019	0.462 ± 0.009	0.012 ± 0.005
	Sequence	0.598 ± 0.017	0.616 ± 0.008	0.018 ± 0.010
Within subject *one sequence out*	*Motion chunk*	0.468 ± 0.015	0.473 ± 0.019	0.007 ± 0.005
	Sequence	0.568 ± 0.020	0.605 ± 0.023	0.036 ± 0.013
Between subject	*Motion chunk*	0.350 ± 0.008	0.361 ± 0.009	0.017 ± 0.006
	Sequence	0.462 ± 0.022	0.478 ± 0.018	0.019 ± 0.015

Table 4.4: Average accuracy and relative accuracy differences for both *motion chunks* and sequences. Only common cardinalities, i.e., $\vartheta_r \cap \vartheta_a = \{24, 32, 40, 48, 64, 80\}$, between ***adhoc*** and ***ranked groups*** are considered. Results are presented for all classification tasks.



(a) Between subject task

(b) Within subject *one sequence out* task

Figure 4.6: Correlation matrix between features in ***adhoc group*** (rows) and features in ***ranked group*** (columns) for cardinality 24. Observations for all actors, sequences and emotions were considered during the computation of the correlation coefficients. Features on ranked group were defined based on the average feature ranking obtained for both between subject (a) and within subject *one sequence out* (b) classification tasks.

Figure 4.6 suggests that the close similarity between the accuracy rates of both feature subset groups can be explained by the high correlation and information redundancy between body joints and limbs. To evaluate this similarity, we built and analyzed the correlation matrix between features in ***adhoc*** and ***ranked groups***. More precisely, we computed Pearson's correlation coefficient between the 24 features selected as the most important using RF-based variable ranking and their equivalent in the ***adhoc group***, i.e., features computed only from head and hands trajectories. Coefficients were estimated for both the average rankings generated during the between subject (Figure 4.6a) and within subject *one sequence out* (Figure 4.6b) cross-validation procedures. As it can be observed on Figure 4.6, most of the features on both groups share between moderate (coefficients around ± 0.5) to strong (coefficients above ± 0.8) linear relationships, which indicates the strength of the association between both feature subsets. Thus, it is likely that both feature subsets explain the same amount of variation related to the expression of emotions.

In order to determine whether the differences between the average accuracy rates obtained for both ***adhoc*** and ***ranked groups*** are statistically significant, we conducted paired statistical tests on common cardinalities, i.e., $\vartheta_r \cap \vartheta_a = \{24, 32, 40, 48, 64, 80\}$. However, for each classification task we measured at most of k-folds $\times$ 10 repetitions of recognition performances for each feature subset (see Tablename 4.3 for a remainder of the number of folds defined for each classification task.). Thus, we count with a limited number of samples of two random variables[6] whose differences do not usually respect the assumptions required by the widely used parametric paired t-tests [64]. Following the advice provided by [64], we employed instead Wilcoxon Signed-Ranks statistical test.

The Wilcoxon signed-ranks test is a non-parametric alternative to the paired t-test with no assumptions about the population's probability distribution. It ranks the differences in performance of two classifiers (one classifier by feature subset in ***adhoc*** and ***ranked groups***) for each test dataset, ignoring the signs, and compares the ranks of the positive and the negative differences. If the difference between the positive and negative ranks approximates zero, within the limits of random variability, the null hypothesis cannot be rejected.

In the context of our work, we aim to evaluate if the same classification model performs equally well for both feature groups. Formally we state the following hypothesis:

$$H_0 : OA_{ranked} - OA_{adhoc} = 0 \qquad \text{Feature subset's performances are not different}$$

$$H_1 : OA_{ranked} - OA_{adhoc} \neq 0 \qquad \text{Feature subset's performances are different}$$

where OA_{ranked} and OA_{adhoc} represent average accuracy for subsets of same cardinality in ***ranked*** and ***adhoc groups*** respectively. That is, we compare each common cardinality independently on both *motion chunks* and sequences predictions. Differences between feature subset groups are considered significant at $p < 0.01$. Table 4.5 shows the $p-$values obtained for each classification task and each cardinality. Conditions for which the null hypothesis could not be rejected are highlighted. We observe that for at least half of the conditions we evaluated, we failed to reject the null hypothesis, thus we do not have evidence to suggest that there is a significant difference between the performance of the feature subsets belonging to ***ranked group*** and their equivalent/images in the ***adhoc group***.

[6] Accuracies for each subset group are considered as random variables

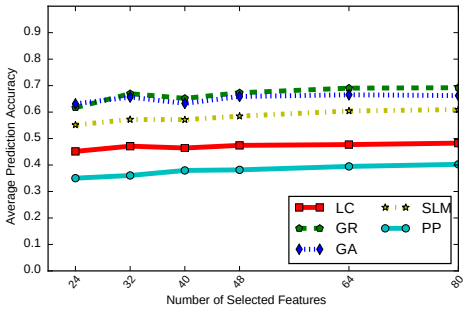
Task	Level	Subsets cardinalities					
		24	32	40	48	64	80
Single subject *one repetition out*	*Motion chunk*	**0.48**	5.1e−10	**0.46**	1.5e−3	1.3e−9	2.7e−5
	Sequence	**0.55**	**0.86**	**0.14**	**0.17**	**1.0**	**0.15**
Within subject *one repetition out*	*Motion chunk*	1.1e−5	1.3e−4	3.2e−6	1.6e−6	1.3e−5	6.6e−4
	Sequence	**0.47**	7.3e−7	**0.02**	**0.15**	1.6e−5	3.3e−3
Single subject *one sequence out*	*Motion chunk*	3.7e−3	**0.14**	**0.01**	1.6e−4	**0.27**	**0.15**
	Sequence	**0.06**	**0.01**	**0.93**	**0.78**	**0.27**	**0.03**
Within subject *one sequence out*	*Motion chunk*	**0.28**	**0.99**	4.7e−7	**0.16**	4.9e−4	1.2e−7
	Sequence	**0.09**	**0.01**	4.1e−6	**0.09**	1.1e−4	3.2e−8
Between subject	*Motion chunk*	**0.85**	8.6e−10	4.2e−7	3.5e−8	4.2e−12	**0.08**
	Sequence	**0.23**	2.1e−8	**0.15**	**0.21**	2.4e−8	**0.42**

Table 4.5: Significant levels (p-values) for difference of average accuracy rate between ***adhoc*** and ***ranked*** subsets of same cardinality. Two-tailed Wilcoxon Signed-Ranks statistical test at $\alpha = 0.01$ level of significance was employed. Conditions for which no significant difference was found are highlighted.

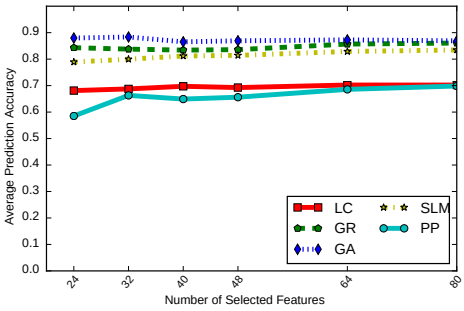
4.8.2 Effect of Actor on Classifier's Accuracy

If we carefully observe the average accuracy reported for single subject *one repetition out* classification task (Figures 4.5a and 4.5b) we find that the standard deviation for both feature groups is significantly high even after performing 10 repetitions of their respective cross-validation scheme. This suggests that, among the five subjects whose data is used for the classifier training and testing, there might be some actors for whom the boundaries between the different classes were not well defined. This fact might explain the substantial decrease on the RF classifier's performance, as shown in Figures 4.5i and 4.5j, when presented with the between subject classification task.

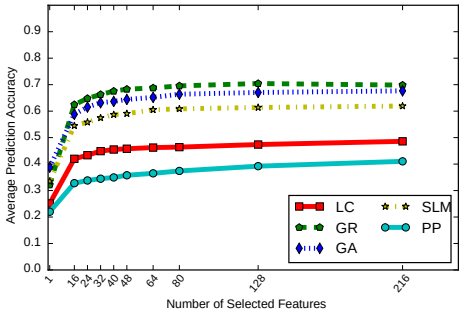
A detailed representation of the performance reported by the classifier for each actor on the single subject *one repetition out* task is depicted in Figure 4.7. Observe that for all combinations of feature subset groups and representations (i.e., *motion chunks* and sequences), there is a significant difference, approximately 30%, between the actors for which the highest (GR and GA) and lowest (PP) performances were registered. In this task there are not unknown sources of variation during the test stage, i.e., the classifier has seen examples of all sequences and all actor-dependent information throughout its training. Thus, the substantial difference of accuracy showed by the classifier when trained and tested on each actor's data suggests that although all actors were submitted to the same emotion elicitation procedure, some of them were much less expressive and eventually showed mixed interpretations of the target emotional states.



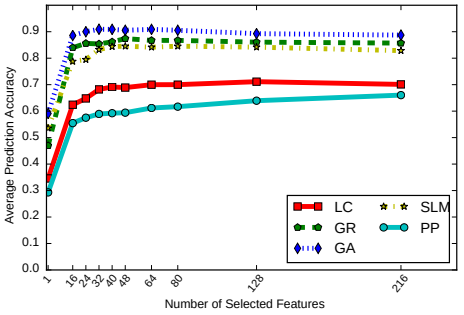
(a) *Motion chunk* accuracy for ***adhoc group***



(b) Sequence accuracy for ***adhoc group***



(c) *Motion chunk* accuracy for ***ranked group***



(d) Sequence accuracy for ***ranked group***

Figure 4.7: Behavior of average accuracy on test set for each one of the five actors whose data makes part of the learning dataset. Each curve corresponds to one actor. Both *motion chunks* (left) and sequences (right) were considered for both feature groups. Measurements were obtained from single subject *one repetition out* classification task.

4.8.3 Effect of Sequence on Classifier's Accuracy

Until now we have seen that the accuracy results obtained for the feature subsets computed from end-effector trajectories are relatively close to those obtained from the automatically defined feature subsets (see Figure 4.5). We have also seen that although actors expressed the same emotional states in different manners, the ***adhoc group*** reached accuracy rates similar to those obtained for the ***ranked group***. Furthermore, both feature groups exhibited the same behavior with respect to the five actors; emotions for actors *GA, GR* and *SLM* were better recognized than emotions for actors *PP* and *LC*. Nonetheless, we still do not know whether end-effector trajectories preserve equally well the affective content in all types of body movements.

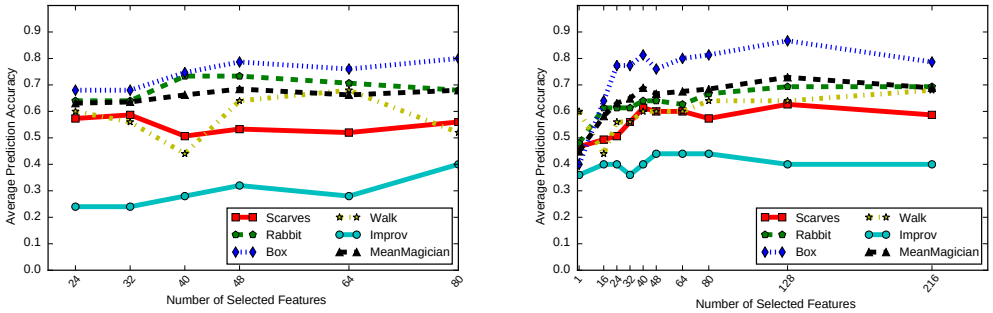
Our database has three main types of body sequences: magician tricks, walk examples and short improvisation sketches. Although the magician tricks belong to the same scenario, there are at least 4 distinct actions in each one of them. Similarly, all improvisation sketches contain different body movements since they were freely chosen by each actor. Since we are validating the proposed end-effector representation by means of automatic classification of affect, we can assess the generalization capabilities to different body movements of the end-effector trajectories through the evaluation of the classifier's action dependency. More precisely, we can gather a general understanding about how the proposed motion model generalizes to notably different body movements by analyzing the RF classifier's performance when the test set contains actions that are kinematically and semantically different from the examples seen by the classifier during its training.

Two of the classification tasks described in Section 4.6 provide this information: single subject *one sequence out* and within subject *one sequence out*. However, in the case of the former task, the results obtained when all observations of each sequence example were used as test set are likely over-optimistic. The classifier might have adjusted itself to the particularities of each actor, rather than learned class boundaries which are action-independent. Henceforth, we based our analysis on the results obtained from the within subject *one sequence out* classification task.

Figure 4.8 shows the average accuracy registered from the RF classifier for each feature subsets' group and for each sequence example. At first glance, we can observe that the body motions in the *disappearing box* trick obtained the highest accuracy rate for both ***adhoc*** and ***ranked group***. Similarly, this sequence is also the best recognized among all magician tricks. We find also that the average accuracy rate for all magician tricks is relatively the same (around 65%) between the feature subsets computed from the end-effector trajectories and those determined through feature selection techniques.

Interestingly, there is a particular pattern common to both feature groups: the recognition rates for the magician and walking examples are much more higher than those obtained for the improvisation sequences. This behavior can be due to several factors:

- Both feature groups generalize well to body movements coming from the same context and exhibiting the same co-articulation and amount of motion quantity between body's limbs. Both the magician sequences and the walking examples required actors to move arms and legs simultaneously and to make displacements around the space. In contrast, three of the five improvisation sequences only employed the arms and did not involve body displacements on ground space.



(a) Sequence accuracy for ***adhoc group***

(b) Sequence accuracy for ***ranked group***

Figure 4.8: Behavior of average accuracy when each one of the five sequence examples was individually employed as test set. *Green*, *red*, *blue*, *yellow* and *cyan* curves correspond to each one of the sequence examples. The *black* curve depicts the average accuracy registered for all sequences in magician scenario, i.e., *box*, *rabbit* and *scarves*. Measurements were obtained from within subject *one sequence out* classification task.

- It might be possible that the kinematic features we used to represent body joints' trajectories failed to abstract the motion cues related to the expression of affect from those inherent to the movement being performed. A detailed discussion and analysis of this issue will be presented in Section 4.8.4.

- There are differences of emotion expression between actors. As we saw in Section 4.8.2, it seems that some actors were less expressive than others. It is highly probable that this difference was even more accentuated on actions freely chosen by the actors.

In conclusion, as long as the body motions and/or actions, for which we seek to classify emotion-related content, employ the same body's limbs during their execution, the end-effector motion parameterization can successfully characterize and preserve their expressive content.

4.8.4 Analysis of Emotion Misclassification

Among the five tasks on which ***adhoc*** and ***ranked*** feature subsets were evaluated, between subject and within subject *one sequence out* are the two scenarios that provide us with greater insight into the generalization capabilities of the motion model we proposed. The former shows us how well the proposed low-dimensional representations preserves the expressive content of movements performed by different subjects, while the latter investigates its suitability to movements/motor behaviors of diverse types. In this section we present a detailed discussion of the classification performance of both tasks for all emotional states. We focus on the confusion matrices obtained for the feature subsets of equal cardinality, i.e., $\vartheta_r \cap \vartheta_a = \{24, 32, 40, 48, 64, 80\}$. Confusion matrices from the within subject *one sequence out* task are presented in Table 4.6 and, and those obtained from the between subject task can be seen in Table 4.7.

At large scale, we note that *sadness* and *happiness* are the emotions with the highest accuracy rates for both tasks and feature groups. This suggests that the expressions of these two emotions were well distinguished among actors and sequences. They are then followed, in no particular order, by *relaxedness* and *stress*. After looking carefully to the misclassification patterns associated to these two emotions in Table 4.6, we can see that they were often labeled as sadness and happiness respectively. This suggests that emotions sharing the same activation level were not correctly separated by the RF classifier. Thus, independently of the feature subsets used to represent all the observations in the learning set, the RF classifier seems to be much better at separating emotions along the activation axis than emotions placed at the extremes of the valence axis. A possible reason for this could be the choice of kinematic quantities as three of the four features that characterize joint trajectories. This claim is supported by the work of Pollick and colleagues [197]. They showed that, for natural arm movements like knocking and drinking, the activation of perceived affect is directly related to the movement kinematics. Thus, our features might not provide enough information for separating emotions along the valence/pleasantness axis.

In the case of within subject *one sequence out* task, the main difference between ***adhoc*** and ***ranked groups*** lays in the results obtained for the *neutral* state. From the left side of Table 4.6, it is clear that for all subsets belonging to ***adhoc group***, the *neutral* state was never recognized above chance level (20%). This suggests that information necessary for the correct discrimination of this state was missing from the features computed from end-effector trajectories. If we look carefully to the first row of the individual confusion matrices shown in Table 4.6, we observe that the *neutral* state was frequently misclassified as either *sadness* or *relaxedness*. It seems then that a RF classifier trained on information extracted from the end-effector trajectories can accurately recognize the differences of the *neutral* state with respect to highly activated emotions (*happiness* and *stress*), but fails to discriminate it from emotional states with low activation and opposed along the valence dimension. In [197, 205], authors suggested that pleasantness information might be directly related to relations between the different limb segments. It is highly probable that this information was implicitly preserved for body joints others than the end-effectors, which will explain why ***ranked subsets*** performed relatively better on the discrimination of the neutral state than the ***adhoc subsets***. However, even for ***ranked*** feature subsets, the recognition rates for the *neutral* state remain the lowest among all intended emotional states (around 29%).

0.17	0.36	0.06	0.09	0.32
0.0	**0.92**	0.0	0.02	0.06
0.05	0.04	**0.71**	0.18	0.02
0.01	0.15	0.17	**0.62**	0.05
0.02	0.37	0.04	0.04	**0.53**
average rate: 0.59				

(a) ***Adhoc subset***: 24 features

0.23	0.33	0.12	0.05	0.27
0.0	**0.93**	0.0	0.0	0.07
0.01	0.05	**0.78**	0.10	0.06
0.0	0.16	0.23	**0.58**	0.03
0.02	0.29	0.12	0.06	**0.51**
average rate: 0.61				

(b) ***Ranked subset***: 24 features

0.18	0.38	0.11	0.05	0.28
0.0	**0.91**	0.0	0.04	0.06
0.03	0.04	**0.75**	0.14	0.04
0.01	0.15	0.22	**0.60**	0.03
0.01	0.32	0.04	0.04	**0.59**
average rate: 0.61				

(c) ***Adhoc subset***: 32 features

0.28	0.31	0.09	0.04	0.28
0.0	**0.92**	0.0	0.0	0.08
0.01	0.04	**0.82**	0.08	0.05
0.0	0.16	0.21	**0.60**	0.03
0.02	0.31	0.12	0.06	**0.49**
average rate: 0.62				

(d) ***Ranked subset***: 32 features

0.18	0.46	0.05	0.07	0.24
0.0	**0.96**	0.0	0.0	0.04
0.03	0.06	**0.75**	0.10	0.07
0.00	0.18	0.19	**0.57**	0.06
0.0	0.30	0.07	0.04	**0.59**
average rate: 0.61				

(e) ***Adhoc subset***: 40 features

0.28	0.31	0.10	0.01	0.30
0.0	**0.92**	0.0	0.0	0.08
0.02	0.03	**0.86**	0.04	0.05
0.0	0.16	0.18	**0.61**	0.05
0.02	0.31	0.08	0.04	**0.55**
average rate: 0.64				

(f) ***Ranked subset***: 40 features

0.18	0.42	0.06	0.05	0.29
0.01	**0.96**	0.00	0.01	0.02
0.0	0.04	**0.82**	0.06	0.08
0.01	0.16	0.17	**0.61**	0.05
0.02	0.30	0.06	0.03	**0.59**
average rate: 0.64				

(g) ***Adhoc subset***: 48 features

0.29	0.30	0.10	0.0	0.31
0.0	**0.93**	0.0	0.0	0.07
0.02	0.02	**0.87**	0.04	0.05
0.00	0.16	0.18	**0.61**	0.05
0.02	0.34	0.08	0.04	**0.52**
average rate: 0.64				

(h) ***Ranked subset***: 48 features

0.18	0.42	0.08	0.03	0.29
0.01	**0.96**	0.0	0.01	0.02
0.00	0.04	**0.81**	0.08	0.07
0.00	0.19	0.20	**0.56**	0.05
0.00	0.28	0.11	0.02	**0.59**
average rate: 0.62				

(i) ***Adhoc subset***: 64 features

0.29	0.32	0.11	0.00	0.28
0.00	**0.94**	0.0	0.0	0.06
0.02	0.03	**0.87**	0.04	0.04
0.01	0.14	0.16	**0.64**	0.05
0.02	0.32	0.06	0.05	**0.55**
average rate: 0.66				

(j) ***Ranked subset***: 64 features

0.19	0.40	0.10	0.01	0.30
0.0	**0.95**	0.0	0.02	0.03
0.02	0.04	**0.87**	0.03	0.04
0.00	0.17	0.18	**0.57**	0.08
0.00	0.29	0.13	0.0	**0.58**
average rate: 0.63				

(k) ***Adhoc subset***: 80 features

0.30	0.31	0.12	0.0	0.27
0.0	**0.94**	0.0	0.0	0.06
0.01	0.03	**0.87**	0.04	0.05
0.01	0.14	0.14	**0.66**	0.05
0.02	0.31	0.07	0.04	**0.56**
average rate: 0.66				

(l) ***Ranked subset***: 80 features

Table 4.6: Confusion matrices for within subject *one sequence out* classification task. Only common cardinalities, i.e., $\vartheta_r \cap \vartheta_a = \{24, 32, 40, 48, 64, 80\}$, between ***adhoc*** and ***ranked groups*** are considered. Individual confusion matrices list emotions in the order *neutral*, *sadness*, *happiness*, *stress*, and *relaxedness*.

0.17	0.38	0.12	0.15	0.18
0.16	**0.79**	0.0	0.0	0.05
0.02	0.09	**0.68**	0.17	0.04
0.06	0.12	0.39	**0.35**	0.08
0.09	0.36	0.10	0.16	**0.29**
average rate: 0.46				

(a) ***Adhoc subset***: 24 features

0.19	0.39	0.08	0.17	0.17
0.07	**0.84**	0.0	0.02	0.07
0.03	0.06	**0.68**	0.15	0.08
0.06	0.14	0.38	**0.29**	0.13
0.14	0.40	0.08	0.14	**0.24**
average rate: 0.45				

(b) ***Ranked subset***: 24 features

0.18	0.36	0.14	0.11	0.21
0.20	**0.66**	0.0	0.02	0.12
0.02	0.05	**0.68**	0.14	0.11
0.03	0.09	0.5	**0.28**	0.10
0.11	0.31	0.12	0.09	**0.37**
average rate: 0.43				

(c) ***Adhoc subset***: 32 features

0.23	0.36	0.07	0.14	0.20
0.08	**0.80**	0.0	0.03	0.09
0.03	0.07	**0.71**	0.12	0.07
0.08	0.12	0.34	**0.34**	0.12
0.21	0.34	0.09	0.09	**0.27**
average rate: 0.47				

(d) ***Ranked subset***: 32 features

0.24	0.27	0.08	0.16	0.25
0.10	**0.75**	0.0	0.02	0.13
0.03	0.04	**0.67**	0.16	0.10
0.13	0.08	0.38	**0.37**	0.02
0.21	0.26	0.15	0.08	**0.30**
average rate: 0.47				

(e) ***Adhoc subset***: 40 features

0.28	0.31	0.07	0.14	0.20
0.09	**0.80**	0.0	0.03	0.08
0.03	0.04	**0.74**	0.10	0.08
0.10	0.10	0.36	**0.30**	0.14
0.29	0.28	0.08	0.08	**0.27**
average rate: 0.48				

(f) ***Ranked subset***: 40 features

0.26	0.24	0.13	0.11	0.26
0.14	**0.73**	0.0	0.02	0.11
0.04	0.02	**0.70**	0.12	0.12
0.15	0.06	0.38	**0.36**	0.06
0.28	0.24	0.12	0.07	**0.29**
average rate: 0.47				

(g) ***Adhoc subset***: 48 features

0.26	0.31	0.08	0.14	0.21
0.12	**0.77**	0.0	0.03	0.08
0.04	0.03	**0.77**	0.09	0.07
0.10	0.10	0.37	**0.28**	0.15
0.26	0.27	0.08	0.08	**0.31**
average rate: 0.48				

(h) ***Ranked subset***: 48 features

0.20	0.36	0.11	0.12	0.21
0.18	**0.73**	0.0	0.02	0.07
0.03	0.04	**0.71**	0.15	0.07
0.12	0.09	0.38	**0.34**	0.07
0.23	0.29	0.09	0.12	**0.27**
average rate: 0.45				

(i) ***Adhoc subset***: 64 features

0.27	0.29	0.08	0.14	0.22
0.15	**0.75**	0.0	0.03	0.07
0.05	0.03	**0.79**	0.08	0.05
0.11	0.10	0.37	**0.31**	0.11
0.28	0.23	0.09	0.07	**0.33**
average rate: 0.49				

(j) ***Ranked subset***: 64 features

0.28	0.32	0.11	0.13	0.16
0.20	**0.75**	0.0	0.02	0.03
0.05	0.04	**0.79**	0.07	0.05
0.11	0.09	0.37	**0.34**	0.09
0.21	0.28	0.09	0.09	**0.33**
average rate: 0.50				

(k) ***Adhoc subset***: 80 features

0.31	0.29	0.08	0.14	0.18
0.15	**0.77**	0.0	0.02	0.06
0.03	0.03	**0.79**	0.08	0.07
0.12	0.11	0.38	**0.31**	0.08
0.28	0.26	0.08	0.06	**0.32**
average rate: 0.50				

(l) ***Ranked subset***: 80 features

Table 4.7: Confusion matrices for between subject classification task. Only common cardinalities, i.e., $\vartheta_r \cap \vartheta_a = \{24, 32, 40, 48, 64, 80\}$, between ***adhoc*** and ***ranked groups*** are considered. Individual confusion matrices list emotions in the order *neutral*, *sadness*, *happiness*, *stress*, and *relaxedness*.

Compared to the results obtained for within subject *one sequence out* task (see Table 4.6), between subject classification (see Table 4.7) is characterized by an overall decrease on the classifier's accuracy for both ***adhoc*** and ***ranked subsets***. Nevertheless, we find the same tendency on the observed recognition rates. *Sadness* and *happiness* are the best recognized emotions, followed by *stress* and *relaxedness*. Similarly, we observe that the rates for *happiness* and *sadness* are relatively close (around $70\% - 80\%$) to those registered by previous studies in which between subject classification of emotional states was performed ([4, 20, 260]).

The *neutral* state reports anew the lowest accuracy rate and keeps being frequently misclassified as either *sadness* or *relaxedness*. However, contrary to the within subject *one sequence out* task, we find that the inverse behavior took place for the between subject task. That is, *relaxedness* and *sadness* misclassification as *neutral* went from none to approximately $25\% - 30\%$ for both feature groups. This fact can be due to: (a) the inability of the selected kinematic features to capture the different nuances of emotions lying along the pleasantness axis, or (b) the subject bias. As it was discussed in Section 4.8.2, some actors showed mixed interpretations of the five target emotional states and even though they were tested against their own referential (see Figure 4.7), the classifier reporter considerably lower accuracy rates for them. Hence, it is possible that by training the classifier on such distinct examples for the same class, the RF model could not effectively separate them.

The considerable gap between the recognition rates for each emotional state makes difficult to reach a final conclusion about the generalization capabilities – to both actions and subjects – of the proposed motion model. However, since the same behavior is observed on the subsets defined through feature selection techniques, it is possible to say that, in the context of the current conditions and limitations, features computed from end-effector trajectories seem to provide the same amount of information about the expressive content of different movement and for different subjects.

4.8.5 Comparison with Human Perceptual Evaluation

During the training of our classifier, we used the elicited emotions as ground truth data. We then evaluated the classifier's overall behavior as well as the relevance of end-effector trajectories in function of these annotations. If the classifier is able to accurately predict a high proportion of these ground truth labels, we consider it is a good predictor of the affective state contained in body movements. Similarly, if features extracted from end-effector trajectories produce also accurate predictions, we rate it as a good motion parameterization of expressive motions. However, we cannot forget that the recognition of emotional states from bodily motions is by definition an extremely subjective task, since humans interpret, perceive and convey emotions differently one from another. For this reason, we decided to use human evaluation as a base line for assessing both the classifier's accuracy and the suitability of the motion model proposed in this thesis.

The comparison between humans and the RF classifier is possible since the features on which the latter has been trained were computed using the same body representations presented to the human participants as described in Chapter 3-Section 3.7. Results obtained for both the human observers and the RF classifier are summarized and presented in the form of confusion matrices in Table 4.8 and Table 4.9 respectively. We consider both the whole-body and end-effector trajectories (*head, hands, feet, and pelvis*) representations. The classifier results come from the between subject classification task, since it is the task that approaches

the most to the conditions in which human raters were evaluated.

At a large scale, we observe the same overall behavior for both representations and features sets. *Sadness* and *happiness* are the two best recognized emotional states, followed by *stress* and *relaxedness*. The *neutral* state, although recognized above chance level in three of the four cases, is the state in which both humans and the classifier were the less accurate. We find also that emotional states are particularly well discriminated along the arousal/activation axis. However, most of the misclassification for both humans and the classifier happen at the pleasantness/valence level. More precisely, *happiness* is more often mixed with *stress* than with any other emotional state. The same pattern is observed between *sadness* and *relaxedness*. These results suggest that: (a) the depictions/expressions of *happiness* and *sadness* were consistent among actors, thus making their discrimination much more easier for both humans and the classifier; (b) the gap between the accuracy rates of *sadness* and *happiness*, and the other emotional states (i.e., *stress*, *relaxedness*, and *neutral* state) reported by the RF classifier (see Table 4.9) can be due to the differences among actors and their interpretations of these three elicited emotions rather than to the kinematic features we selected.

0.23	0.22	0.17	0.11	0.27
0.16	**0.45**	0.08	0.10	0.21
0.10	0.06	**0.57**	0.16	0.11
0.15	0.13	0.22	**0.36**	0.14
0.22	0.19	0.19	0.10	**0.30**
average rate: 0.38				

(a) Perceptual study: whole-body stimuli

0.13	0.18	0.28	0.20	0.21
0.14	**0.26**	0.13	0.14	0.33
0.06	0.10	**0.46**	0.29	0.09
0.11	0.15	0.27	**0.32**	0.15
0.12	0.17	0.30	0.20	**0.21**
average rate: 0.28				

(b) Perceptual study: end-effector trajectories

Table 4.8: Confusion matrices from perceptual study described in Section 3.7 in Chapter 3, i.e., whole-body stimuli (left) and end-effector trajectories (right). Emotions are listed in the order *neutral*, *sadness*, *happiness*, *stress*, and *relaxedness*.

0.31	0.30	0.08	0.14	0.17
0.13	**0.76**	0.00	0.02	0.09
0.04	0.02	**0.77**	0.10	0.07
0.12	0.09	0.39	**0.30**	0.10
0.28	0.25	0.08	0.04	**0.35**
average rate: 0.49				

(a) Classifier: whole-body features (cardinality 216)

0.26	0.24	0.13	0.11	0.26
0.14	**0.73**	0.0	0.02	0.11
0.04	0.02	**0.70**	0.12	0.12
0.15	0.06	0.38	**0.36**	0.06
0.28	0.24	0.12	0.07	**0.29**
average rate: 0.47				

(b) Classifier: end-effectors features (cardinality 48)

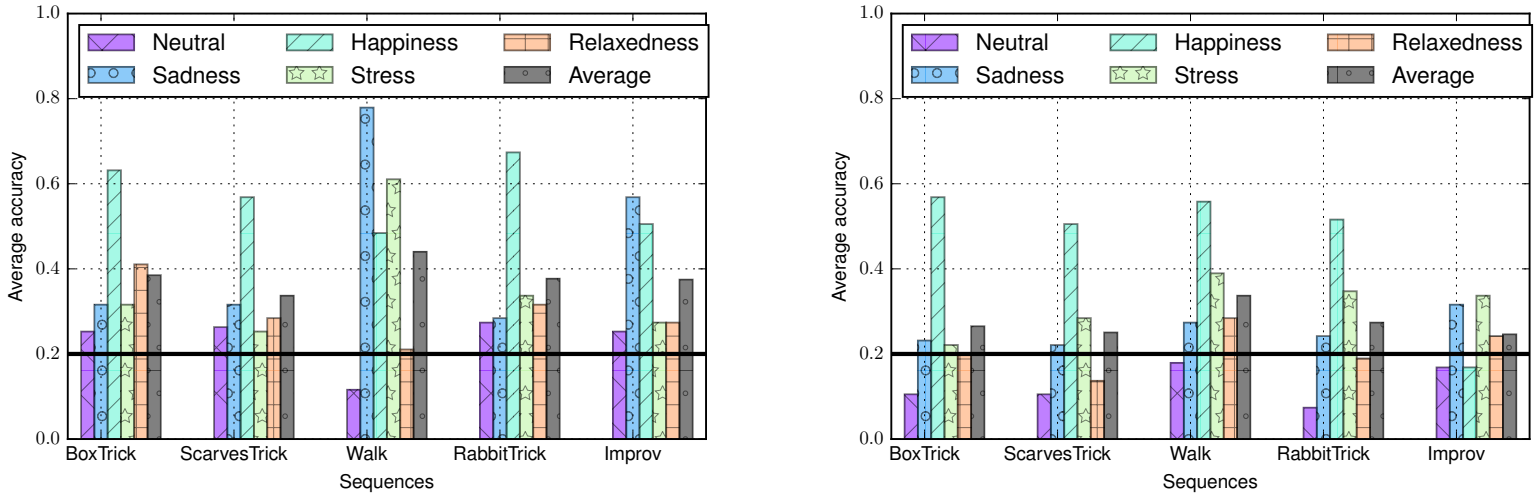
Table 4.9: Confusion matrices for classifiers trained on body representations that are equivalent to those presented during the perceptual study described in Section 3.7. Emotions are listed in the order *neutral*, *sadness*, *happiness*, *stress*, and *relaxedness*.

The average recognition rates between human observers (Table 4.8a) and the RF classifier (Table 4.9a), for the whole-body representation, show that the classifier was, in average, 11% much more accurate than humans. A possible explanation for this is the lack of a training stage for the human observers who rated our database. Contrary to other classifier-human comparisons in which raters saw between 10 [129] to 60 [130] trials by emotional state, due to

the long duration of the motion sequences in our database, we presented each rater with at most 5 examples for each emotional state. Thus, human observers had fewer opportunities to grasp the particularities of the actors, sequences and emotions than the classifier.

After comparing the average rates for the end-effector trajectories representation, we observed the same behavior. This time however, the RF classifier (see Table 4.9a) outperforms by 19% the human observers (see Table 4.8b). In addition to the lack of training for the human observers, it seems that, from a perceptual point of view, end-effector trajectories alone do not provide as much affect-related information to humans as they do to a classifier. This impairment of human observer's capacity to discriminate affect when presented with end-effector trajectories only can be explained by the few or none structural and body-form information provided by this type of visual stimuli. More precisely, in spite of previous perceptual studies showing that humans can still recognize emotions from coarse body representations such as point-light displays [6, 197], it seems that supplementary form information is still needed for sophisticated tasks such as an accurate recognition of emotional states [6]. Nonetheless, we highlight that human observers were still able to recognize 4 of the 5 elicited emotional states above chance.

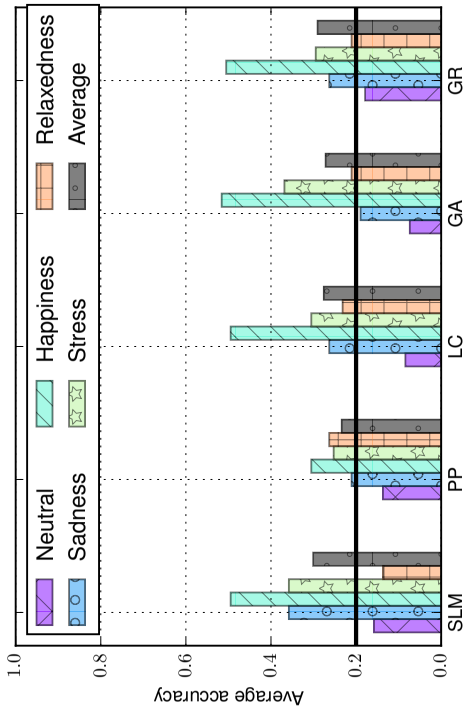
In addition to the analysis of the difference of overall performance between human observers and our RF classifier model, we are also interested in studying how raters's accuracy was influenced by sequences and actors. We wish to determine if the sequences and actors for which the classifier was the most accurate are those who reported the highest perception rates. Figure 4.9 and Figure 4.10 show the recognition rates obtained from human observers for each one of the five examples of motion sequences and each one of the five actors in the MoCap database respectively. For the whole-body representation, we observe that, contrary to the results obtained from the classifier (see Figure 4.8), the walk examples were the sequence for which emotions were the most accurately recognized. The *box trick*, *rabbit trick* and the improvisation sketches follow closely. The sequence that reported the lowest accurate rate is *scarves trick*. In the case of end-effector trajectories stimuli (see Figure 4.9b), the relative order between motion sequences given by the human observers' rates is closer to the one obtained from the RF classifier (see Figure 4.8b). Both the improvisation sketches and the *scarves trick* are the motion examples for which emotion recognition seems to be the hardest. In general, we find that the relative ordering of the magician sequences is common to both classifiers and human annotators, and that whereas the classifier performs its best on the theatrical gestures, the human observers seem to be more adept at recognizing emotions from less elaborated body motions; possibly because humans are more "trained" at decoding and understanding locomotion movements [67].



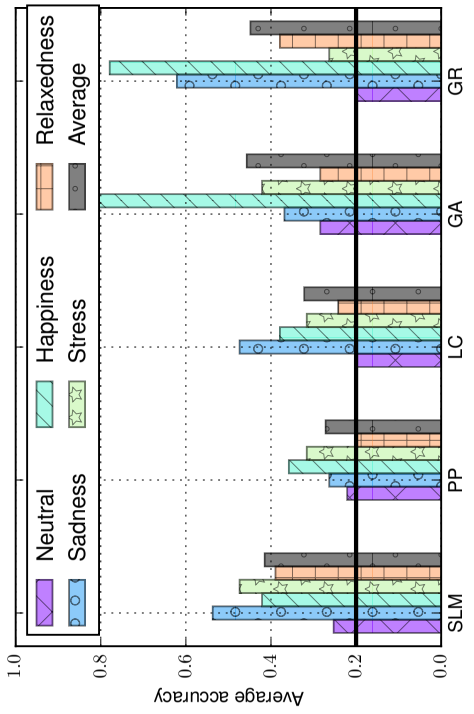
(a) Whole-body stimuli

(b) End-effector trajectories stimuli

Figure 4.9: Accuracy rates reported by human observes for each one of the five examples of motion sequences in the MoCap database



(b) End-effector trajectories stimuli



(a) Whole-body stimuli

Figure 4.10: Accuracy rates reported by human observes for each one of the five actors in the MoCap database

With respect to the influence of the actors, in the case of whole-body stimuli (see Figure 4.10a), it seems that human observers perceived the same differences in the actors' expressiveness than the RF classifier (see Figure 4.7). The actors GR, GA, and SLM reported higher accuracy rates in comparisons to the other two actors LC and PP. This result confirms, to some extent, our previous assumptions. It is possible that the substantial decrease in the classifier's accuracy during the between subject tasks and the low perception rates reported by the raters are due to how differently the actors interpreted and conveyed the same elicited emotion. Thus, since the variance within the examples of the same emotional state is so large, both the classifier and the human observers made more mistakes in their judgments. Finally, other than a slight increase in the recognition rates obtained for LC actor, we observe the same relative differences between actors expressiveness in the end-effector trajectories stimuli (see Figure 4.10b).

4.9 Summary

This chapter presented the classification protocol used to quantitatively validate the suitability of end-effectors trajectories for the parameterization of expressively body motions. After taking into consideration the technical challenges associated to the classification of time series: variable length, high-dimensionality, and ease of interpretation, a feature-based representation of motion sequences was defined. In this representation, the spatial-temporal trajectories of all body joints were characterized by three kinematic (velocity, acceleration, jerk) and one geometric (curvature) quantities. The choice of this representation was motivated by previous studies showing the importance of kinematic information in the perception of emotion from body movements. Furthermore, the chosen quantities could be easily computed from whole (all body joints) or partial (end-effector trajectories) body representations, which made the evaluation of the proposed motion model much more easier. We later detailed how fixed-length feature vectors were estimated from the four kinematic quantities describing the 3D trajectory of each body joint.

The main idea behind the use of automatic affect classification was to compare the performance of a single classifier when trained on two different groups of features. The first group consisted of features computed from end-effector trajectories only, whereas the second group comprised subsets of features automatically defined from all body joints. Since features subsets were to be defined, a brief review of the main approaches on automatic feature selection was presented. From then on, given the context of our study and the limited amount of data to use during learning and testing, it was argued that the combination of an ensemble classifier and a forward selection strategy would yield good results. A theoretical and empirical justification for the selection of Random Forest as the ensemble classifier to be used was then presented.

After introducing and describing the classifier, feature-based representation, and the feature selection procedure to be used, we presented a thorough description of the five classification tasks on which the relevance of end-effector trajectories was to be evaluated. These tasks, as a whole, provided us with enough information about the generalization capabilities of the motion model proposed in this thesis. We evaluated general behavior of a classifier trained on end-effector trajectories when different sequences of actions and actors were used as test sets. Similarly, we evaluated their robustness to different classification settings such

as classifier hyper-parameters and windows' lengths and overlaps (see Appendices B and C). Finally, we compared the overall classifier performance, for the whole-body and the end-effector trajectories against the recognition rates obtained from a user study.

4.10 Discussion

The analysis of the accuracy rates obtained for both ***adhoc*** and ***ranked groups*** show that, in the majority of the cases, the performance of a classifier trained on end-effector trajectories-related information could not be judged statistically different, at a significant level of $\alpha = 0.01$, from the their equivalent in the *ranked group* (see Table 4.5). For the cases in which the differences were found to be significant, we observed that the recognition rates obtained from the feature subsets in ***adhoc group*** were usually within one standard deviation (Table 4.4) of the accuracy rates registered by the ***ranked group***. We observed also that when comparing the average recognition rates obtained in the most difficult of all the classification task we evaluated, i.e., between subject task, for the largest cardinalities of each group – 80 features for the ***adhoc group*** and 216 (whole-body) for the ***ranked group*** –, there was only a slight decrease (2% approx.) in the classifier's overall accuracy (see Table 4.9). Thus, we can conclude that the first part of the hypothesis which stated that: *"the performance of a classifier trained on features computed form the proposed low-dimensional motion representation is close to the performance obtained for the same classifier trained using features computed from the entire-body or feature subsets automatically selected"* was confirmed by our results.

The second part of our hypothesis stated that the classifier's accuracy rates when trained on end-effector trajectories should be close to the rates obtained during a perceptual study. In Section 4.8.5, we pointed out that both feature groups, in particular the ***adhoc group***, outperformed human observers for both stimuli. We hypothesize that this noteworthy difference was due to the few examples shown to the observers in comparison to the dozens seen by the classifiers during its training phase. Nonetheless the second part of our hypothesis was also confirmed and we can conclude that, from a quantitative point of view, end-effector trajectories preserve most of the information related to the expression of affect in body motions.

A detailed analysis of the individual recognition rates of both classifier and user study pointed out that there were notable differences among emotions, actors and sequences. We observed that for both classifier and raters, *sadness* and *happiness* recognition rates were at least two times larger than the rates obtained for the other emotional states. In the case of the RF classifier, since most of the misclassification happened along the pleasantness/valence axis, we hypothesized that this difference was due to our decision of using only kinematic features as representations of body joints trajectories. However, after observing the same pattern in the user study, it seems that the consistent differences between the recognition rates of the five distinct emotional states can be due to differences in actors performances. Instead of a lack of discriminative power for both the RF classifier and the chosen features, it seems that actors showed different expressiveness capacities, and understood and interpreted the elicited emotions differently. This claim is sustained by the differences between actors recognition rates shown in Figures 4.7 and 4.10. We also found significant differences on the accuracies of both classifiers and raters across sequences. However, these differences were not consistent between the two of them. The classifier highest rates came from exam-

ples of the magician sequences, while the user study highest rates were observed for the walk and improvisation examples.

The strength of end-effector trajectories performance was also tested under different parameters of the classification protocol (see Appendices B and C) described in this chapter. The results obtained for different parameterizations of the sliding window – method used in the definition of feature vectors – and of the RF hyper-parameters indicated that, from an automatic classification perspective, both our experimental setup and the end-effector trajectories were already at their best performance. However, from a feature selection standpoint, it seemed that the procedure used in the definition of the ***ranked group*** can still be improved and thus yield smallest and more informative feature subsets. Nonetheless, the optimization of a feature selection approach is beyond the objectives of this chapter.

Chapter 5

Motion Synthesis through Inverse Kinematics and a Random Walk

Contents

One of the main objectives of this thesis is to generate expressive motions for virtual human-like characters. By motion we refer to the specification of a character's posture through time and space. The resulting motion should not only look natural and plausible but also satisfy some user-defined constraints such as follow a specific path in space, exhibit a particular style pattern, e.g., emotion-related content, or respect a determined postural configuration, e.g., raise one arm during half of the motion.

As already mentioned, in character motion synthesis, virtual characters are usually represented as skeletons, i.e., articulated bodies. An articulated body is in turn defined as a hierarchical structure composed of rigid objects, called *links* – analogous to human bones – connected together by *joints*. A *joint* is the component concerned with motion since it allows some degree of relative movement between two rigid segments. Thus, in a broad sense, animating a virtual character comes to define the 3D transformations that need to be applied to each one of the *joints* in the character's skeleton such that a determined configuration, i.e., posture, is attained.

The definition of an articulated body as a hierarchical structure suppose that: (a) all links or bones have at most one parent and any number of children, (b) any translation and/or rotation on the i-th joint affects the translation and rotation of any joint placed later in the articulated body, and (c) there is at least one bone with no children further in the hierarchy; this bone is identified as *end-effector*.

A character's skeleton can be decomposed into multiple and simpler articulated bodies, hereinafter referred to as articulated chains. A chain can be built for each *end-effector* in the character's body by moving back through the skeleton, going from parent to parent, until the start of the chain (the root) is reached. Head, hands and feet are some of the most common *end-effectors*. Figure 5.1 shows an example of a character's skeleton and the articulated chains that can be defined from it.

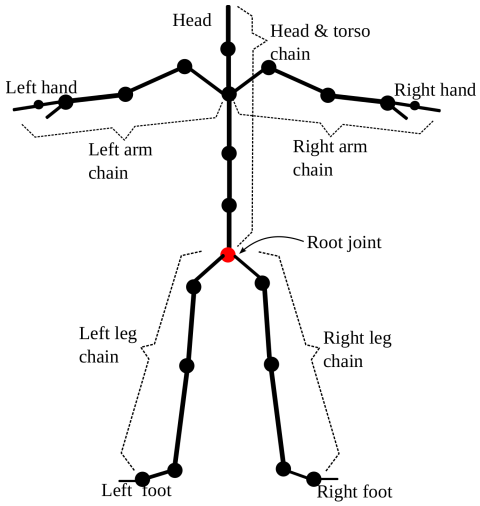


Figure 5.1: Example of virtual character's skeleton and articulated chains that can be defined from it. Head, hands and feet are the corresponding end-effectors.

Formally, an articulated chain $\mathcal{B}$ is composed of n joints. The configuration of $\mathcal{B}$ at frame t is defined by both the rotation vector $\mathbf{Q}_t = (\mathbf{q}_1, ..., \mathbf{q}_n) \in SO(3)^n$, where $\mathbf{q}_i$ is the unit quaternion [179] that defines the orientation of the i-th joint, and the translation $\mathbf{x}_{1,t}$ of the chain's root joint:

$$\mathbf{X}_t = (\mathbf{x}_{1,t}, \mathbf{Q}_t) \tag{5.1}$$

Given this formulation, generating an articulated body's motion can be reduced to solve either the *forward* or *inverse* kinematics problem. The former amounts to compute, at each time step t, the resulting posture once the root's translation $\mathbf{x}_{1,t}$ and the rotation vector $\mathbf{Q}_t$ have been set. The latter describes the process of determining an appropriate configuration for which the end-effectors move to a desired position or follow a determined trajectory over time.

In this chapter we briefly review the existing approaches to generate articulated body motions that exhibit expressive and/or stylistic content. We then describe the synthesis approach implemented in this thesis as well as the quantitative and qualitative evaluations of the motions generated with this approach.

5.1 Synthesis of Expressive Body Motions: a Survey

In the context of expressive motion synthesis, in addition to determine the appropriate skeletal configuration at each frame, we are also interested in modulating how this configuration changes over time. By doing so, we expect to accurately convey emotion-related content to the user with whom/who the virtual character might interact.

We have identified two main approaches for the generation of expressive motions [1]: *rule-based methods*, also known as procedural animation, and *example-based methods*. In this section we review the main principles behind these two approaches as well as several examples.

5.1.1 Rule-based methods

Rule-based methods propose an ensemble of motion generation and editing rules that specify how a set of motion parameters can be mapped to emotionally expressive body motions. These rules are often guided by a perceptual standpoint and reflect a thorough understanding of the motion characteristics employed by human observers when perceiving and decoding affect and emotions. Methods belonging to this category generate new motions by solving variants of the inverse kinematics problem. Several examples can be found in both the affective computing and computer animation literature.

Using a parameterization based on results found in the psychology literature, Hartmann and colleagues [109] defined six qualitative parameters: overall activation, spatial intent, temporal extent, fluidity, power, and repetition, in order to specify gesture expressiveness. These high-level parameters are latter mapped into low-level animation parameters that modify according to some rules of equivalence defined by the authors the wrist control trajectories and the body postures provided by an inverse kinematics analytical solver. This system has been extensively used for analyzing emotion-related gesture expressiveness in human users [169] and for synthesizing expressive gestures for embodied conversational agents [186].

The EMOTE model for effort and shape, developed by Chi *et al.* [52], uses components of the Laban Movement Analysis system to control the form and the execution of qualitative aspects of movements. These parameters are mapped to low-level movement rules that are then applied to independently defined underlying movements. The underlying movements were generated by a combination of key-poses, interpolation, and *inverse kinematics*.

Neff and colleagues [184, 185] propose a supporting software system for creating expressive character animations. This system uses key aspects of expressive movements (the amplitude of the motion, the amount of muscular tension, a particular class of posture, etc.) inspired by the study of artistic performance literature. These aspects, called movement properties, provide handles to specify expressive movements, a character's movement style, and to edit and refine an animation. Two levels of representation are employed: a lower-level or base representation that specifies the motion of a single joint or group of joints, and a higher-level that incorporates ideas from the arts literature and defines the structure of a motion. Higher-level parameters map to lower-level parameters through a script created by an animator. Final animations are created by defining key-poses and transitioning between them according to the lower and higher level parameters provided by the animator.

[1] Also referred to as stylistic motions in the computer animation domain.

Finally, Alaoui and colleagues [1, 2] measured different motion qualities in order to characterize the dance motions of a user interacting with an artistic installation. The author then defined a set of rules that mapped the measured qualities to the movement and behavior of two non-anthropomorphic physical systems. The mapping rules were manually defined according to the physical parameters and properties of each representation.

5.1.2 Example-based methods

Example-based methods use existing expressive motion examples to guide the editing and synthesis process. The examples encode the motion features that need to be present in a motion sequence so as to be categorized and/or perceived as expressive and/or stylistic. It is worth to notice that methods belonging to this category employ a more vague definition of expressive motions. They often refer to *motion style* rather than to emotions or any other affective phenomena, and define it as the particular manner in which a motion/action is performed. Therefore emotionally expressive motions are considered as a particular case of *motions with style*. Studies belonging to this category generate new motions by implicitly solving the forward kinematics problem. Within this category we have identified four main groups: motion blending, component models, style translation techniques, and stochastic generative models.

Motion Blending

New motions are generated through weighted interpolation of structurally similar (i.e., motions that depict the same action with different styles) but distinctive motion examples. Models within this group have to solve the inverse motion interpolation problem: given the values of a set of predefined control parameters, a set of motion examples and their blending weights have to be found so that the new motion follows the control parameters [250]. Different manners of determining blending weights and their mapping to control parameters have been proposed along the years.

Rose *et al.* [206] were the first to propose motion blending as a way to generate whole-body motions that exhibit emotional expressiveness. They proposed to build a space of different styles and variations (adverbs) for a single action (verbs). To do so, sets of similar, but distinct motion examples were parameterized through adverbs. A combination of radial basis functions and low order polynomial was used to determine the blend weights as well as the mapping between them and the adverbs. A similar method in which motion style is controlled with semantically meaningful commands (e.g., 'do the same, but more sadly') analogous to adverbs was propose by Forger *et al.* [84]. They introduced a correspondence between calculated motion features and high-level natural language commands. Given a desired style, the corresponding motion features are computed. The system, through an iterative and interactive process, automatically defines and tunes the necessary interpolation weights so that the resulting motion best exhibits the desired features.

Other motion blending approaches seek to learn the function that maps control parameters to interpolation weights. For instance, in [238] motion examples are labeled using Laban (LMA) Effort dimensions: flow, weight, time and space; the LMA annotations serve as control parameters. A blend motion is created for each pair of kinematically similar examples. The blend is then annotated using LMA parameters. By doing so, a parametric non-linear

function between the blending weights and the LMA parameter values of the blended motion is learned. This function quantifies how changes in the interpolation weights influence the style of the blended motion. The learned model can then be used to apply a variety of styles to other pre-recorded yet unseen motions. In another study, Ma and colleagues [164] use a Kriging model to determine the blending weights to use during interpolation. The proposed model takes as input a set of latent variation parameters that parameterize and control the style and variation of a predefined joint group. Furthermore, they introduce an additional mapping between user-defined constraints and the latent variation parameters; this mapping is in turn approximated through a Bayesian network.

Component Models

Introduced by [247], component models see human motion as the combination of many different, sometimes mutually independent, motion components. Some of them relate to the content of the motion, the identity of the performer of the motion, the style (the particular manner of the motion) or any other element that determines the motions being performed. New motions are generated by exchanging, merging and interpolating one or several components according to some high-level constraints or control parameters. Different methods, most of them coming from the statistics field, have been employed/proposed to obtain these so-called *motion components*.

Urtasun *et al.* [248] make the assumption that motions can be represented as linear sums of principal components that encapsulate and preserve most of the nuances and patterns found in the motion examples [202]. Following this idea, [248] propose to build a motion space by applying Principal Component Analysis (PCA) on examples of several subjects performing walking and jumping motions at different speeds and lengths. This space defines all possible motions that can be generated for a new actor. Given an example from an unseen subject, synthesized motions at different walking speeds or jumping lengths can be generated by projecting the new example into the PCA space and computing the respective sum coefficients. These coefficients are determined as functions of the distance between the projected motion and the motions within the PCA space.

Shapiro and colleagues [221] propose an interactive method for generating stylistic motions. They employ another linear model, Independent Component Analysis (ICA), to separate data motion into visually meaningful components called style components. New motions are generated using the style components that a user interactively identifies as suitable for the synthesized motion. Extensions to the use of ICA for motion decomposition have been proposed by [133] and [160]. While [133] proposed to apply ICA to body parts' motion rather than to the whole body movement as done by [221], [160] used a modified ICA procedure in order to obtain a better definition of the motion components related to style.

Rather than letting the decomposition method to determine the number of components underlying a motion, He *et al.* [112] suggested that human movement is composed of three main mutually independent elements: content (the intrinsic properties of the motion being performed), identity (all properties inherent to the performer of the motion), and style (the particular manner in which a motion is performed). These three components can be obtained through a combination of non-linear dimensionality reduction techniques and multi-linear tensor analysis. In the one hand, non-linear dimensionality reduction extracts the content of the motion while simultaneously preserving the geometry of the underlying manifold and

a direct mapping from low-to-high dimensional space. On the other hand, multi-linear tensor analysis decomposes multiple orthogonal factors which represent the motion variations related to style and identity components. New motions are then generated by exchanging, merging and interpolating one, two or all factors according to some user-defined constraints.

Style Translation

Style translation is the process of transforming an input motion into a new style while preserving its original content [119]. This transformation is estimated via the analysis of the differences between realizations of the same content in two different styles, e.g., a neutral and sad walk. This approach usually involves two main stages: *i.)* spatio-temporal alignment between the two motions [114] and *ii.)* modeling the stylistic differences. The main difference between the different implementations of style translation principles lies in this latter stage.

For instance, Amaya *et al.* [3] employ ideas from signal processing to compute emotional transformations that capture the differences between a neutral and emotional movement with respect to speed (timing) and spatial amplitude (range). These transformations are then applied to existing neutral motions in order to produce the same motions, but with an emotional quality. Hsu *et al.* [119] propose to use Linear Time Invariant (LTI) models to learn the differences between the input and output styles. Once the LTIs' parameters have been estimated from the training data, the model translates new motions with simple linear transformations. The resulting motion retains its content, but differs in its style of execution. In the work done by Xia and colleagues [266], differences between input and output styles are approximated through an online learning algorithm that builds a series of local mixtures of autoregressive models. Once the models parameters are estimated from the training data, the poses in the input motion are transformed to the desired output style with simple linear transformations. A particularity of the method proposed by [266] is that local regression models are build on the fly from the closest examples of each input pose in the database. Finally, Etemad and colleagues [79] use an ensemble of Gaussian RBF neural networks to model the differences between neutral and stylistic motion sequences. Prior to training the neural networks, time warping and principal components analysis are used. The trained model is capable of translating the learned styles to neutral input sequences.

Stochastic Generative Models

Another category of approaches used for editing and/or synthesizing stylistic motions is that one of stochastic generative models. That is, statistical models that capture the data's essential structure, i.e, the spatial and temporal variations defining both content and style. The appeal behind these models is justified by their capacity to generate motions distinct from the training data and from the increased number of motion variations that can be obtained with a small number of hidden variables [176]. Generative models that implicitly estimate of a low-dimensional hidden space and its corresponding mapping to high-dimensional movement space are among the most used. Below we discuss some of the most popular models belonging to this group.

Brand *et al.* [33] used Hidden Markov Models (HMM) to learn motion patterns from a highly varied set of motion captures. A multidimensional style variable that can be used to

vary the HMM parameters was added to the the standard HMM model. While the hidden states of the trained HMMs capture the content of the motion, the style variables model the differences between motions. The resulting models can be used to apply a given style to a motion sequence, to generate new sequences by doing a random walk on the trained HMMs, or to create new styles by interpolating or extrapolating within the space defined by the multidimensional style variables. Other extensions of HMM models for the generation of stylistic motions sequences were proposed by [235] and [211].

Motion textures [156] and *Motion graphs++* [176] are stochastic adaptations of a motion graph [142]. Motion sequences are segmented into motion textons [156] or morphable motion primitives [176] that capture repetitive patterns within the MoCap database. Each motion texton or morphable motion primitive represents a node within the motion graph and is used to approximate a local generative model. The local models represent all the style variations observed for a given texton or primitive. Li *et al.* [156] used linear dynamic systems to estimate them, while Min *et al.* [176] employed a combination of functional data analysis and Gaussian Mixture Models (GMM). The weight associated to each edge in the motion graph corresponds to the transition probability between textons or primitives. Whereas new motion stylistic variations can be obtained via probabilistic sampling, new motion sequences are generated via graph walks.

Taylor and colleagues [232] proposed to use Conditional Restricted Boltzmann Machines (CRBM), a deep neural network model, for exact inference and generation of stylistic human movements. They later added a set of style variables to gate the connections between the different layers within the neural network model [231]. By doing so, a much more powerful generative model was obtained. This new model allows controlled transitioning and blending of different styles. Specifically, a change in the style variable induces a change in the effective weights of the network. Thus, changing these style-based variables during generation induces natural transitions between different styles and permits interpolation and extrapolation of styles observed in the training data [231]. Recently, Holden *et al.* [118] proposed a deep learning framework for the motion synthesis and editing. In this framework, a convolutional autoencoder and a deep feedforward neural network are combined. The resulting model generates new stylistic motions by combining the style of one motion with the timing of another.

Finally, Gaussian Processes [201] have been used to imbue IK solutions with the style variations present in a MoCap database [100] and to generate new motion sequences that combine variations related to factors such as style, identity and content [259].

5.1.3 Discussion

As it was previously stated, this thesis lies in the intersection of two research domains: *affective computing* and *computer animation*. The generation of believable and expressive animated characters is a common interest between these two domains. However, while the affective computing domain supports most of its work on a profound study and understanding of what makes human motions to be perceived as emotionally expressive, computer animation rather strives to generate human motions as visually rich and detailed as those observed in MoCap databases. Thus, we observe that the affective computing domain favors rule-based synthesis methods, whereas most of the work done in computer animation belongs to the example-based category.

Each approach has its own merits and drawbacks. In the one hand, the strength behind ruled-based methods lies in their capacity to offer much better control and in the flexibility and variability of the motions that can be obtained. However, these motions are often described as stiff and less visually appealing. Furthermore, the definition of the relations and rules that map control parameters to motion features and motion features to emotional states is a complex task [131]. On the other hand, example-based methods are capable of generating novel movements with a high level of details and a great realism. Given a MoCap database containing examples of perceptually validated expressive motions, we can be sure that the novel movements generated from these examples will be equally expressive. Nevertheless, the flexibility and variability of example-based methods entirely depends on the richness and vastness of the labeled dataset used in these approaches. Furthermore, example-based methods often provide a very limited control over the possible output motions and styles, and can be rarely used to analyze and understand what exactly makes a motion to be perceived as expressive.

In order to benefit from the strengths and advantages of both approaches, while addressing their shortcomings, we propose to combine their main principles. Namely, we introduce a synthesis system in which the control and flexibility of the rule-based methods, and the visual appeal of the example-based techniques are combined. First, we propose to use end-effector trajectories as control signals. They are intuitive, easy to specify (e.g., through low-cost motion capture systems such as the Kinect sensor), and provide high-level control since there is a direct relationship between the input parameters and the resulting motion. Furthermore, since they will be extracted and/or generated from a labeled and perceptually validated dataset, we preserve most of the visual appeal of example-based methods. Second, we propose to use a procedural method (inverse kinematics) as the function that maps end-effector trajectories to whole-body motions. By doing so, we retain the control and flexibility of the rule-based approaches and decrease the dependency of the example-based methods on the motion database.

5.2 Mapping from End-Effector Trajectories to Whole Body Motions

When high-dimensional data has to be generated using only low-dimensional control signals (e.g., end-effector trajectories), two types of approaches are usually privileged: data-driven reconstruction methods [230, 49, 161] and Inverse Kinematics solvers [236, 12]. In the former, example postures closely approximating the trajectories described by the control signals are retrieved from a large MoCap database. Those examples are later used to build local models capable of producing continuous and smooth whole body motions. The resulting movements exhibit the same stylistic characteristics of the examples contained in the database.

In the latter approach, an iterative optimization solver computes the motion, posture by posture, for which the end effectors follow as smoothly and accurately as possible the control signals. Other than the knowledge about the hierarchical representation of the human body and its degrees of freedom, an IK solver has no prior information about the stylistic characteristics of the motions to be produced. Thus, if the resulting bodily motion exhibits stylistic variations related to a particular emotional state, it is because the relevant expressive cues were encoded in the control signals. For this reason, an IK-based solver has been

adopted as part of the synthesis method presented in this thesis.

In this section, we introduce the theoretical formulation of the inverse kinematics problem as well as the most common solution approach. We then outline what are the main challenges associated to the generation of full-body posture and how we address them in our synthesis approach.

5.2.1 Theoretical Background on Inverse Kinematics

Given the rotation vector $\mathbf{Q}$ and the root position $\mathbf{x}_1$ associated to the articulated chain $\mathcal{B}$, it is possible to define $\mathcal{F}$ as the forward kinematic operator used to compute the pose of the end-effector associated to $\mathcal{B}$. Although an end-effector's pose is usually defined by an orientation and a position, in this thesis we are only interested in the end-effector's position. Thus $\mathbf{z} \in \mathbb{R}^3$ and determines what we referred to as task space:

$$\mathcal{F} : \mathbb{R}^3 \times SO(3) \longmapsto \mathbb{R}^3 \tag{5.2}$$

$$(\mathbf{x}_1, \mathbf{Q}) \longmapsto \mathbf{z} \tag{5.3}$$

Inverse Kinematics (IK) can be defined as the problem of controlling an articulated chain through the specification of a target positions for the chain's end-effector. That is, IK aims to find a rotation vector $\mathbf{Q}$ such that $\mathbf{z}$ reaches a desired configuration $\mathbf{z}_d$. Hence $\mathbf{z}_d$ corresponds to the control signal that guides the *inverse kinematics* process and whose dimensionality is much lower than the dimensionality of $(\mathbf{x}_1, \mathbf{Q})$ combined.

For simplicity we summarize hereinafter the forward kinematics operator by $\mathbf{z} = \mathcal{F}(\mathbf{Q})$. IK can be then formulated as the following non-linear inverse problem:

$$\mathbf{Q} = \mathcal{F}^{-1}(\mathbf{z}_d) \tag{5.4}$$

Solving this inverse problem (5.4) is not an easy task. Since the function $\mathcal{F}$ is non-linear, there may not always be a solution or there may not be a unique (best) solution [37]. The most direct approach for solving the IK problem would be to obtain a closed-form solution for Equation 5.4. However, even in well-behaved situations, this kind of solution cannot be generally achieved. Furthermore, the larger and more complex the articulated chain we aim to control, the smaller the likelihood of finding an analytical solution [261]. Therefore, the IK problem is commonly solved through numerical approximation.

The most popular numerical approach consists on linearizing the IK problem about the current configuration $\mathbf{Q}$ using the Jacobian matrix $\mathbf{J}$. Then it is possible, through a simple iterative scheme, to converge towards the desired target $\mathbf{z}_d$ by computing small variations $\delta\mathbf{Q}$ of the rotation vector $\mathbf{Q}$. This approach ensures (most of the times) that the regulation from $\mathbf{z}$ to $\mathbf{z}_d$. Formally, the IK problem can be reformulated as follows:

$$\delta\mathbf{Q} = -\lambda \mathbf{J}_{\mathbf{Q}}^{-1}(\mathbf{z}_d - \mathcal{F}(\mathbf{Q})) \tag{5.5}$$

$$\delta\mathbf{Q} = -\lambda \mathbf{J}_{\mathbf{Q}}^{-1}\delta\mathbf{z} \tag{5.6}$$

where $\delta\mathbf{z}$ is the desired change on the end-effector's configuration, $\mathbf{J}_{\mathbf{Q}}^{-1}$ is the inverse of the Jacobian of the articulated chain $\mathcal{B}$ evaluated around the current configuration $\mathbf{Q}$, and λ is a scalar quantity which defines the rate of convergence. The iterative scheme we have just mentioned amounts to the following algorithm:

Data: $\mathbf{z_d}$: target configuration, $\mathbf{Q}$: current rotation vector
repeat
  Compute $\delta\mathbf{z} = \mathbf{z}_d - \mathcal{F}(\mathbf{Q})$;
  Calculate $\mathbf{J}$ for current $\mathbf{Q}$;
  Invert $\mathbf{J_Q}$;
  Calculate $\delta\mathbf{Q}$ using Equation 5.6;
  Compute new configuration $\mathbf{Q} \mathrel{+}= \delta\mathbf{Q}$;
until $\mathcal{F}(\mathbf{Q})$ *is sufficiently close to* $\mathbf{z_d}$;
Algorithm 3: Iterative solution of the IK problem through Jacobian inverse methods.

Although the Jacobian solution is fairly easy to implement, the Jacobian matrix may not be square or invertible, and difficulties might arise when the articulated chain is highly redundant (i.e., it has more degrees of freedom than are necessary to specify a target for the end-effector). Because of this, $\mathbf{J}_\mathbf{Q}^{-1}$ is usually replaced by some generalized inverse $\mathbf{J}_\mathbf{Q}^{\dagger}$ approximation:

$$\delta\mathbf{Q} = -\lambda\mathbf{J}_\mathbf{Q}^{\dagger}\delta\mathbf{z} \tag{5.7}$$

Some of the generalized inverses proposed along the years are: Jacobian transpose [264], Singular Value Decomposition (SVD) [166], the pseudo-inverse also known as the Moore-Penrose inverse of the Jacobian, the damped pseudo-inverse [256], among others. The main differences among all these approximations of the Jacobian's inverse are their convergence rate and their behavior around singularities, i.e., situations in which no changes in the rotation vector $\mathbf{Q}$ achieve the desired change in the chain's end-effector configuration.

Since the number of degrees of freedom in a character's articulated body is most of the times greater than the dimensionality of target space, the number of solutions to the IK problem is usually infinite. Therefore, although the Jacobian inverse approach provides us with a minimal norm solution [261], it is possible that the resulting body configuration may not look as humanly natural and plausible as we expect them to be. Fortunately, this same redundancy can be exploited for adding secondary constraints or tasks that will reduce the space of possible solutions selected by the Jacobian approach to a smaller and more desirable set.

These additional constraints or tasks can be added through a projection operator ($\mathbf{I}_n - \mathbf{J}_\mathbf{Q}^{\dagger}\mathbf{J}_\mathbf{Q}$) that allows us to project them on the null-space of $\mathbf{J}$, that is, the components of the secondary task that do not change the end-effector's configuration are selected. Hence, the new solutions is given by:

$$\delta\mathbf{Q} = -\lambda\mathbf{J}_\mathbf{Q}^{\dagger}\delta\mathbf{z} + (\mathbf{I}_n - \mathbf{J}_\mathbf{Q}^{\dagger}\mathbf{J}_\mathbf{Q})\nabla\mathcal{H}(\mathbf{Q}) \tag{5.8}$$

where $\mathbf{I}_n$ is an $n \times n$ identity matrix and the function $\mathcal{H}(\mathbf{Q})$ is a cost function to be minimized and subject to satisfy the primary configuration task ($\delta\mathbf{z} \approx 0$). This null-space method has been extensively used, notably to enforce joint limits [270], avoid singular configurations [94], control the position of a character's center of mass [31] or to assign priorities to different tasks [12, 270].

5.2.2 Main Challenges of IK-Based Synthesis of Whole-Body Motions

When using *inverse kinematics* for the generation of whole-body postures and/or whole-body motions several challenges need to be addressed. Here we outline the main three aspects that we consider are crucial for the generation of expressive bodily motions from expressive end-effector trajectories:

CH1 Generating whole-body postures using *inverse kinematics* supposes to specify and handle multiple end-effector configurations simultaneously. In the case of articulated body models such as the one presented in Figure 5.1, at least five different target configurations, one for each arm and leg, and one for the head, need to be provided to the synthesis system. It might happen that all or some of these targets, also referred to as constraints, conflict with each other and cannot be satisfied at the same time. Thus, a strategy for solving these conflicting situations is necessary.

CH2 Due to the excess of degrees of freedom in a character's articulated body, motions generated via purely procedural techniques such as IK are often considered as mechanical and unusual. Additional constraints such as explicit bio-mechanical joint limits are necessary in order to enhance the realism of the synthesized motions.

CH3 Our aim is the generation of whole-body motions that convey the internal emotional state of a virtual character. Hence, it is crucial that all features associated to the emotion-related content embedded in the end-effector trajectories can be successfully preserved during the IK reconstruction. Furthermore, these features should also propagate through the character's whole-body posture and the resulting motion.

In the remaining of this section, we present and describe the simple yet powerful *inverse kinematics* implementation we have used as the function doing the mapping from expressive end-effector trajectories (low-dimensional space R^d) to expressive full-body postures (high-dimensional space R^D).

5.2.3 Controlling Articulated Chains Independently

When multiple end-effector tasks are defined, there are two main strategies for the formulation of the *inverse kinematics* problem. In the first strategy, all the joints in the character's articulated body are involved in the satisfaction of all tasks or constraints [12, 270]. In the second one, the character's skeleton is partitioned into several articulated chains or joint groups not necessarily independent from each other (i.e., they can share common body joints). Each articulated chain is dedicated to specific tasks and has an IK solver associated to it [24, 146, 222, 236].

Although the first approach is often preferred for its capacity to preserve the synergy among joint groups, it requires a higher computational cost (the Jacobian matrix is larger and hence its inversion requires more operations) and a more detailed, sometimes expert, control of the priority between multiple, occasionally opposing, constraints. For instance, if both arms must reach configurations on opposite directions, it is necessary to determine which end-effector has a higher priority. Suppose for instance that the right arm target is considered as more important, hence the left arm target is taken into consideration through its projection

into the Jacobian's null space (see Equation 5.8). However, with the increase of the number of constraints, the instability around singularities of the Jacobian inverse approach is harder to solve [12] or avoid [270]. Furthermore, the priorities among all tasks need to be determined *a priori* the IK computations. This suppose to either define heuristics that will automatically determine the relative importance among all constraints as done in [224] or to manually set them [12].

Conversely, when applied to human-like articulated bodies, such as the one considered in this thesis, the second approach makes the computations easier and simpler [48]. This is mainly due to the reduction of both the number of dimensions in the Jacobian matrix and the sum of constraints to be considered at each step. Additionally, since it is possible to define which joints we wish to consider for each articulated chain, a more efficient and detailed control can be applied [153, 150]. However, it is still necessary to define the order in which each articulated chain must be treated. Fortunately, *Choi et al.* [53] pointed out that finding an inverse kinematic solution for multiple chains, each one with an end-effector target associated to it, does not require any priority assignment between the chains as long as all constraints remain attainable. As we will show later, this is the case when the end-effector tasks assigned to each chain are generated using examples from an expressive MoCap database. For all of these reasons, we have decided to adopt a multi-chain *inverse kinematics* approach for the generation of whole-body expressive motions.

In this thesis we have defined five articulated chains, one for each limb in the character's body. Figure 5.2 shows the five selected chains and their corresponding end-effector segments (highlighted in red) as well as the number of degrees of freedom associated to all joints within each chain.

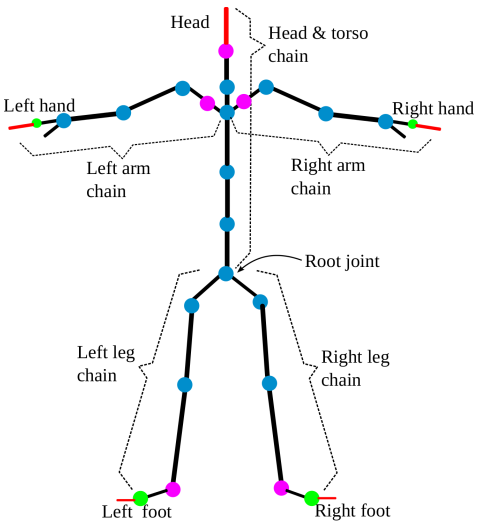


Figure 5.2: Articulated chains and DOF controlled through IK. Red segments indicate end-effector associated to each articulated chain. Degrees of freedom of each joint are indicated as follows: green dots = 1 DOF, pink dots = 2 DOF, and light-blue dots = 3 DOF.

From Figure 5.2, we observe that both right and left leg chains have 12 DOF, left and right arm chains have 15 DOF each, and the head-torso chain has 17 DOF. We find also that there is one common joint (the body's root joint) between the legs and head-torso chains and one common joint (the upper-torso joint) between the three upper-body articulated chains. This

implies that the angular increases computed for these two body joints might differ among the different articulated chains sharing them. To solve this problem, we have adopted an approach based on the weighted sum of the different angular increments.

Given an initial configuration $\mathbf{Q}$ of the character's articulated body, we compute and store the corresponding *inverse kinematics* solution for each chain with respect to this initial configuration. Once angular changes have been computed for all body joints, we average all the angular increments computed for the root and upper-torso joints respectively. We then update the character's body configuration and repeat the same process until all the targets have been attained or new targets have been specified. For the implementation used in all our experiments, we have associated the same weight (w=1) to all articulated chains have a body joint in common. Nonetheless, these weights can eventually be tuned in function of the movements to be generated.

5.2.4 Constraining the IK Solution to the Space of Plausible Human Postures

When using *inverse kinematics*, the representation of the motion range of each joint within the character's body is fundamental to obtain natural and plausible animations for human characters [270].

There are two main approaches for including joint limits into the IK solution process. In the first and most classical approach, a cost function $\mathcal{H}(\mathbf{Q})$ is designed to be minimal when the body joints' orientations are within safe configurations and maximal when these orientations are beyond or at the vicinity of the joint limits [135, 94, 159]. In the second approach, also known as *clamping*, joints' orientations are treated independently. Each orientation is clamped back within its valid range of motion whenever the orientation change computed by IK moves it beyond its validity domain [12, 203].

We have identified two essential differences between these two ways of reinforcing joint limits. The first difference relates to the application domain. Whereas the first approach is mostly used in robotics (e.g., [170]) where articulated bodies have less DOFs, the second one is frequently employed in character animation (e.g., [12]). The second difference lies on their integration within the IK computation process. When the first approach is used, joint limits are considered during the computation of the small variations $\delta\mathbf{Q}$ to be applied on the current chain configuration $\mathbf{Q}_t$. This is done through the projection operator introduced in Equation 5.8. Conversely, the clamping approach is considered (most of the times) as a post-processing step. That is, after the new configuration $\mathbf{Q}_{t+1} = \mathbf{Q}_t + \delta\mathbf{Q}$ has been determined, the orientation $\mathbf{q}_i$ associated to the *i-th* joint is tested against its validity domain, if $\mathbf{q}_i$ violates this domain, we replace it with the closest valid orientation. Both approaches constrain the IK solutions to more natural postures, however the first approach is harder to implement when joints with more than one degree of freedom are considered. As this is the case when working with virtual human characters, we focus hereinafter on *clamping* approaches only.

The implementation of the clamping approach depends on both the number of degrees of freedom associated to each body joint and the rotation parameterization being used [11]. Joints with one or two degrees of freedom are commonly parameterized using Euler angles. That is, each DOF has associated an independent axis of rotation and the range of motion around this axis is determined by a rotation angle θ. These joints are the easiest to constrain, since joint limits can be independently specified for each degree of freedom [11]. It suffices to define a minimum, θ_{min}, and a maximum, θ_{max}, rotation angles. As long as $\theta_{min} \leq \theta \leq \theta_{max}$,

no clamping is necessary. However, if it is not the case, θ is set to the closest limit, i.e., $\theta = \theta_{min}$ or $\theta = \theta_{max}$.

Joints with 3 DOF, also known as ball-and-socket joints, are harder to model and constrain, since we cannot longer assume that all degrees of freedom are independent [11]. Fortunately, for the purpose of defining a range of motion a simple yet intuitive decoupling can be applied. Namely, the orientation of a ball-and-socket joint can be though of as being composed of two motions: (a) a *swing* motion that controls the direction of the limb directly attached to the joint, and (b) a *twist* motion that lets the limb attached to the joint rotate about itself [13]. Each of these two motions can be then independently constrained. On one hand, the *twist* component can be parameterized using Euler angles. Hence, its valid range of motion can be defined and constrained by an $[\theta_{min},\ \theta_{max}]$ interval as done for simple 1 DOF joints. On the other hand, the range of motion of the *swing* component can be expressed as a region in a 3D space [97, 270]. Joint limits are reinforced as follows: we test whether the current *swing* orientation is within the defined region, if it is not the case, we find the closest projection on the 3D valid region. Cylinders [270], ellipses [97], implicit surfaces [115], and reach cones [263] are some of the approaches that have been proposed to approximate the valid region of the *swing* orientation.

Although several authors have argued about the suitability of the *swing-twist* decomposition of ball-and-socket joints [13, 97, 270], we opted for using simple Euler angle constraints for each one of the 15 ball-and-socket joints in our character's articulated body [237, 261] (see Figure 5.2). We did so for two practical reasons. First, in order to model a proper valid region for the *swing* motion of a ball-and-socket joint, it is necessary to record valid motions of maximal amplitude [115]. Unfortunately, at the moment we were recording our MoCap data we did not take this requirement into consideration. Second, we tried to define the *swing* region based on anatomical observations on body joints' range of motions. However, we found that the resulting motions exhibited frequent singularities and discontinuities that could considerably hinder the perception and evaluation of emotion-related content.

Determining Valid Joint Limits from MoCap Data

As it was mentioned, we have decided to model joints limits through simple Euler angles. These limits are obtained from all the MoCap examples used as learning set and ground truth data by the synthesis tasks later described in this chapter. For each one of 56 DOF in our character's body, we defined an $[\theta_{min},\ \theta_{max}]$ interval using the approach introduced in [74] and summarized in Algorithm 4:

Data: $\mathbf{\Theta} = (\theta_1, ..., \theta_n)$: angle values associated to the *i-th* degree of freedom
Result: $[\theta_{min},\ \theta_{max}]$: valid angle interval for the *i-th* degree of freedom
Set $D = []$;
Set $n = |\mathbf{\Theta}|$;
Sort $\mathbf{\Theta}_i$ in ascending order;
for *j from 0 to* $n - 1$ **do**
  Set $k = (j + 1) \bmod n$;
  if $\mathbf{\Theta}[k] > \mathbf{\Theta}[j]$ **then**
    $D[j] = \mathbf{\Theta}[k] - \mathbf{\Theta}[j]$;
  else
    $D[j] = \mathbf{\Theta}[k] + 2\pi - \mathbf{\Theta}[j]$;
  end
end
Set m = index of $\max(D)$;
$\theta_{min} = \mathbf{\Theta}[(m + 1) \bmod n]$;
$\theta_{max} = \mathbf{\Theta}[m]$;

Algorithm 4: Algorithm used to determine the range of motion of each degree of freedom.

Once all valid intervals have been computed with Algorithm 4, after each IK iteration we bound the resulting rotation vector $\mathbf{Q}$, if necessary, using the clamping scheme described for joints with 1 DOF. Although this way of managing joint limits works well most of the times, we are aware that some resulting posture might still look as unusual to human observers. This is due to the lack of mutual constraints between: *i.)* the DOFs that belong to the same body joint [261] or *ii.)* neighboring body joints (e.g., shoulders and elbows) whose movements are not completely independent from each other [9].

An Additional Constraint: Elbow Trajectory

The major difficulty of solving an inverse kinematics problem for human-like figures stems from the excessive number of DOFs associated with this type of model. Whereas our character's body model has 56 DOFs in total, we currently count with only five constraints, i.e., five end-effector target positions, for manipulating the character's articulated body. The inclusion of joint limits helped to considerably reduce the span of the solution space, however there are still some explicit redundancies that required additional constraints.

First mentioned by Korein *et al.* [139], one explicit redundancy in the human model is the "elbow circle" shown in Figure 5.3. Namely, even though the shoulder and the wrist are firmly and correctly positioned, it is still possible to move the elbow along a circle with its axis being defined by the straight line connecting the shoulder and the wrist [153]. In analytical and simpler models of the upper-limbs, this extra degree of freedom is parameterized with the so-called *swivel* angle whose value can be easily determined using standard Euclidean vector operations [127].

However, since our character's arm chains count with more degrees of freedom than the arm models for which analytical solutions can be computed, we have adopted a different approach. We decided to add an additional constraint to both arm chains instead. Namely, at each time step, we provide both a target position for the arm's end-effector (hand) and the elbow joint. We then use the same IK controller associated to the arm articulated chain to

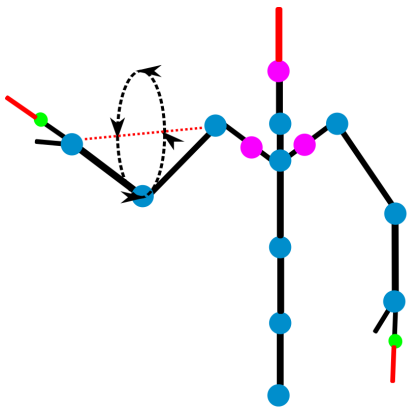


Figure 5.3: Explicit redundancy of the arm linkages

determine the orientation changes $\delta\mathbf{Q}$ that best accomplishes both tasks.

Although we consider that both tasks are complementary and thus have the same priority, we still need to define their relative importance. To do so, we have adopted a weighting strategy as done in [236]. This strategy is frequently used for the animation and positioning of human-like articulated bodies and consists in assigning to each task a weight which defines its importance with respect to the other tasks. In this manner, the IK solver finds a compromise in the satisfaction of both tasks. It has been argued that with this strategy none of the tasks is exactly satisfied [12]. However, we found that the overall patterns of the arm motions obtained through this strategy visually preserve the emotion-related content we are interested in. In our current implementation we have empirically assigned weights of $[0.6, 0.4]$ to hand and elbow targets respectively.

5.3 Trajectory Generation by Re-sampling in Target Space

In the previous two chapters, we showed that both human observers and an automatic affect classifier could recognize the expressive content embedded in the body motions contained in our MoCap database. They did so when presented with either whole-body motions or end-effector trajectories. We also saw that, from the standpoint of automatic classification of affect, features computed from end-effector trajectories preserve and encode most of the content related with the expression of emotions. However, as pointed out by Cowie *et al.* [58], the features used to quantify and estimate expressive behavior might also be affected by the movements being performed. Thus, to further evaluate the suitability (i.e., its robustness and independence from the action being performed) of the proposed motion model and its applicability to the generation of expressive whole-body motions, we need to produce and analyze new samples of expressive end-effector trajectories. These new samples should describe movements sufficiently different from those available in the MoCap database, while preserving all motion patterns related to the emotions and affective phenomena of interest. One way to achieve this goal is to adopt a random walk on the manifold defined by our MoCap database. We argue that trajectories generated in such a manner (i.e., by switching randomly from one observed state to another at a unit time) are, as far as possible, decorrelated from the semantic actions and sequences contained in the database, but hopefully

preserve the motion cues that are indicative of expressed emotions. Hence, we have decided to support the generation of new expressive end-effector trajectories and consequently new whole-body motions on data-driven animation techniques and statistical models (particularly re-sampling schemes). The reasons behind this choice are three fold:

- Statistical methods are powerful tool for encoding and modeling motion specific information such as expressive content. Given a limited number of known observations (i.e., motion examples in a MoCap database), they can efficiently infer information about: *i.)* the statistics of the features salient to body motions expressing a particular emotional state, and *ii.)* the data generation process from which the expressive movements under study were issued.
- Data-driven methods provide us with the means to generate plausible expressive motions since they make use of observations of the corresponding phenomenon in the real world. In other words, the desired expressive behavior has been directly measured on real subjects. These knowledge can then be used to support the generation of equally expressive end-effector trajectories
- The progress made on data-driven character animation during the last years has shown how the combination of data and powerful statistical models can bring realistic results (e.g.,[51, 118, 155]). By learning what are the bodily expressions associated to a particular emotional state it is possible to infer new expressive motions (e.g., [33]).

Among the large spectrum of available statistical models, we have chosen *bootstrap re-sampling* as an efficient and fast way of generating new expressive end-effector trajectories. The particularity of this approach is that since new trajectories are generated in a random manner, we can ensure that: *i.)* they are sufficiently different from the observed trajectories, *ii.)* the semantic significance and dependency associated to known human actions is not longer present, and *iii.)* the underlying motion patterns associated to the expression of emotion and observed in the examples are still present in the novel re-sampled trajectories. In other words, we go beyond the semantic and meaning inherent to human motion and completely focus on what make a movement to be perceived as expressive.

The fundamental principle of bootstrap methods is very simple: generate new samples from the set of known observations (i.e., MoCap end-effector trajectories) such that the statistical information of the underlying population and distribution can be inferred [110] (i.e., the features salient to the expression and/or perception of an emotional state are still present in the new samples). Before introducing the re-sampling approach we have used, we briefly review the challenges of bootstrap methods on time-dependent data such as motion data.

5.3.1 Bootstrap on Time Series

Since their introduction by Effron [70] in 1979 a wide range of bootstrap methods have been proposed. In the context of this thesis, we loosely classify them according to the type of data of interest. If the data is a random sample from an unknown probability distribution, i.e., independent and identically distributed random variables, new bootstrap samples can be simply generated by either sampling the data randomly with replacement or by sampling an approximated parametric model of the data's probability distribution F [108]. Conversely,

when working with dependent data, e.g., time series, normal bootstrap methods as the ones already mentioned are not longer suitable. Methods that take into account the dependence structure of both the data and its generation process are needed instead.

Formally, a time series $(X_1, X_2, ..., X_n)$ can be defined in turn as a realization of an unknown d-dimensional stochastic process $\{\mathcal{X}_t\}_{t\in\mathbb{T}}$. That is, given a probability space $(\Omega, \mathfrak{F}, P)$, so that Ω is the sample space, $\mathfrak{F}$ the σ-algebra of events, and P the probability measure of $\mathfrak{F}$, a stochastic process on $(\Omega, \mathfrak{F}, P)$ with state space $\mathbb{R}^d$ and ordered with respect to a time index set $\mathbb{T}$ corresponds to a collection of random variables $\{\mathcal{X}_t\}_{t\in\mathbb{T}}$ such that $\mathcal{X}_t$ takes values in $\mathbb{R}^d$ for each $t \in \mathbb{T}$. Furthermore, since the end-effector trajectories in our database are equispaced in time, $\{\mathcal{X}_t\}_{t\in\mathbb{T}}$ corresponds to a *discrete-time* stochastic process, i.e., $\mathbb{T} \equiv \mathbb{N}$.

Although we know that there is a temporal dependency on the observed time series and consequently in the stochastic process that generated them, it is possible to assume that this dependency is local in time. That is, the probability of the next state taken by $\mathcal{X}_t$ will be conditional only on the previous states closer in time. In other words, we can assume then that the *discrete-time* stochastic process $\{\mathcal{X}_t\}_{t\in\mathbb{T}}$ follows a general Markovian structure and can be further characterized as a *discrete-time Markov process*.

Formally speaking, a *discrete-time Markov process* is defined as follows: we assume that an integer $p > 0$ exists such that, for all $t \in \mathbb{T}$, the state of the process $\{\mathcal{X}_t\}_{t\in\mathbb{T}}$ at time t depends only on the p previous states, i.e., for every $t \in \mathbb{T}$ and for every Borel set $A \subset \mathbb{R}^d$,

$$P(X_{t+1} \in A | X_j,\ j \leq t) = P(X_t \in A | X_j,\ t - p + 1 \leq j \leq t) \tag{5.9}$$

That is, for any series of observations in discrete time, each future state, given the entire past and the present state of the process, depends only on the p present states [61]. This dependency is captured by the transition probability function listed in Equation (5.9).

When the time series generating process is approximated by a *Markov* process, bootstrap methods can be resumed to the estimation of the Markov transition density (5.9) through non-parametric methods [108]. New bootstrap samples, $(\hat{X}_1, \hat{X}_2, ..., \hat{X}_N)$, are generated by starting at an arbitrary state and then sampling the generating process implied by the estimated transition probability [108, 196].

Among the existing bootstrap methods for *Markov* processes, we have decided to work with the re-sampling procedure proposed by Monbet *et al.* [177] *Local Grid Bootstrap* (LGB). Our choice was motivated by its capacity to: *i*) better capture the time dependence structure of a time series, *ii*) generate sequences whose length may be chosen independently from the length of the observed sequences, and *iii)* produce unobserved states within the new sequences.

5.3.2 Local Grid Bootstrap for Time Series Re-sampling

Given an observed time series $(X_i)_{i\in\{1,\cdots,T\}}$, LGB resampling algorithm generates a new time series $(\hat{X}_i)_{i\in\{1,\cdots,N\}}$ where the length N may be chosen independently from the length T of the observed time series. To do so, it assumes that the *Markov process* $\{\mathcal{X}_t\}_{t\in\mathbb{T}}$ can be further approximated by a strictly stationary p-order *Markov chain* $\{\mathcal{Y}_t\}_{t>p}$ with $Y_t = (X_t, X_{t-1}, ..., X_{t-p+1}) \in \mathbb{R}^{dp}$. This chain, in addition to the transition distribution with con-

tinuous function $F_y(\mathbf{x}) = P(X_{t+1} \leq \mathbf{x}|Y_t = \mathbf{y})$[2], admits also a stationary distribution with continuous function $F(\mathbf{y}) = P(Y_t < \mathbf{y})$[3] which determines the likelihood of reaching the state Y_t at any future time. Since the state space in our case is Euclidean, both the stationary and transition distributions are approximated respectively by kernel estimates density K_d and K_{dp}.

Under the LGB re-sampling scheme each new observation $\hat{X}_{t+1}$ is obtained by assigning probabilities (stationary and transition probabilities) to a finite subset of convenient states and sampling this subset according to these discrete probability masses. The assigned probabilities correspond to the density kernel estimates K_d and K_{dp} computed using a bandwidth parameter h_T previously defined. The finite subset of convenient states from which each new observation $\hat{X}_{t+1}$ is obtained is defined by: *i.)* the successors $(\widehat{VY}_t)^+$ of the observed neighbors, within a ratio of width σ_T, $\widehat{VY}_t$ of the last sampled state $\hat{Y}_t = \{\hat{X}_t, \hat{X}_{t-1}, \cdots, \hat{X}_{t-p+1}\}$, and *ii.)* the points of a local grid with discretization step Δ_g and edge length σ_g around these successors. See Figure 5.4 for a visual representation of the subset we have just described.

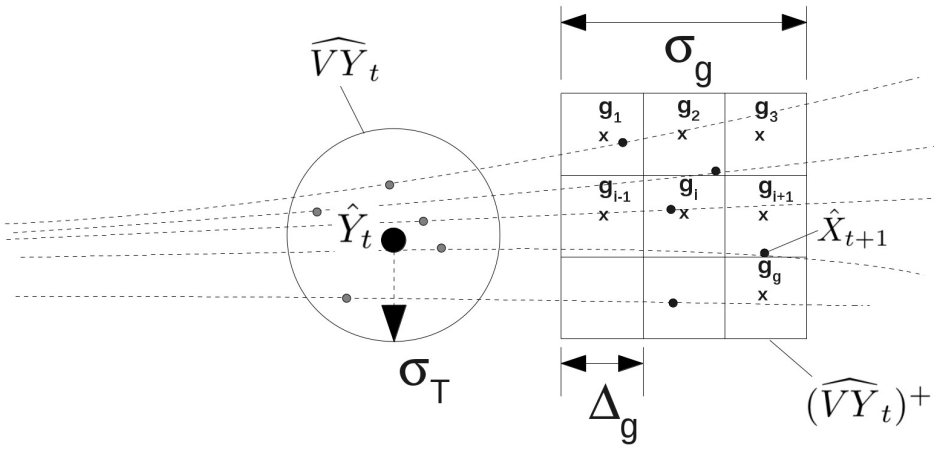


Figure 5.4: Local grid bootstrap procedure

Formally, the LGB re-sampling procedure may be defined as follows:

Initialization step

Let $\hat{Y}_t = \{\hat{X}_t, \hat{X}_{t-1}, ..., \hat{X}_{t-p+1}\}$ denote the state of the generated sequence at time t. An initial state $\hat{Y}_1$, the kernel bandwidth parameter h_T, the width σ_T of the neighborhood of a given state and the grid parameters Δ_g (discretization step) and σ_g = (edge length) are selected.

***Step** t:*

- Let us suppose that the state $\hat{Y}_t$ is already sampled. The neighborhood $\widehat{VY}_t$ of $\hat{Y}_t$ is defined by the set of observed $Y_l \in \{Y|d(Y, \hat{Y}_t) \leq \sigma_T/2\}$. Define $I[\widehat{VY}_t]$ the set of time index such that for all $l \in I[\widehat{VY}_t], Y_l \in \widehat{VY}_t$. Furthermore, the image $(\widehat{VY}_t)^+$ of $\widehat{VY}_t$ is defined by $(\widehat{VY}_t)^+ = \{X_{l+1}, l \in I[\widehat{VY}_t]\} \subset \mathbb{R}^d$.

[2] $X_{t+1} \leq \mathbf{x}$ means $X_{i,t+1} \leq x_i \; \forall_i \in \{1, \cdots, d\}$.
[3] $Y_t < \mathbf{y}$ means $Y_{i,t} \leq y_i \; \forall_i \in \{1, \cdots, dp\}$.

- A grid $G_t = \{g_1^t, \cdots, g_{T_g^t}\}$ is built by discretizing a cube of $\mathbb{R}^d$ with grid step Δ_g and edge length σ_g. The cube is centered on the barycenter of $(\widehat{VY}_t)^+$ and the edge length σ_g is defined such that the cube includes at least all the elements of $(\widehat{VY}_t)^+$. Let us denote $GY_T^{(t)} = (\widehat{VY}_t)^+ \cup G_t$.

- Let J be a discrete random variable taking its values in $I[GY_T^{(t)}] = \{k \in \mathbb{N},\ X_k \in GY_T^{(t)}\}$, with probability mass function given by:

$$P(J = k) = \frac{p(\hat{Y}_t, X_{k+1})}{\sum_{j \in I[GY_T^{(t)}]} p(\hat{Y}_t, X_{j+1})},\ \forall k \in I[GY_T^{(t)}] \tag{5.10}$$

$$p(\hat{Y}_t, X_{k+1}) = \sum_{i \in I[\widehat{VY}_t]} K_d\left(\frac{X_{k+1} - X_{i+1}}{h_T}\right) K_{dp}\left(\frac{\hat{Y}_t - Y_i}{h_T}\right) \tag{5.11}$$

  where K_d and $K_d p$ are the local density kernel estimates of the transition and stationary density probability functions around $GY_T^{(t)}$.

- The sampled state at time $t+1$ is such that $\hat{X}_{t+1} = X_J$; namely $\hat{X}_{t+1}$ is randomly sampled according to the probability mass function (5.11)

The discrete probability mass $p(\hat{Y}_t, X_{k+1})$ may be considered as transition probabilities between $\hat{Y}_t$ and X_J. It depends both on the density of the original sequence around X_J and on the density of the observed state around $\hat{Y}_t$. As the kernels K_d and K_{dp} are continuous on $\mathbb{R}^d$ and $\mathbb{R}^{dp}$ respectively, it is possible to assign probabilities to unobserved points of the grid and consequently to sample observed and unobserved states.

5.3.3 Application of LGB Re-sampling to Motion Data

In Chapter 4 it was mentioned that a motion sequence can be seen as a multidimensional time series which takes its values in the space of human plausible postures. Similarly, since postures evolve smoothly along time, there is a temporal dependence in human motion that must be preserved and accounted for when generating both expressive end-effector trajectories and expressive whole-body motions. Fortunately, as shown by the classification results discussed in Chapter 4, when working with expressive bodily motions, end-effector trajectories provide almost the same temporal information than whole body motion sequences. Hence, as long as a bootstrap method suitable for time series is applied on the known observations of end-effector trajectories, we should be able to obtain temporally coherent whole-body motions. It suffices to apply our IK implementation on the new bootstrap samples. From this, it follows that the observed end-effector trajectories can also be considered as time series and the LGB re-sampling procedure can be applied on them.

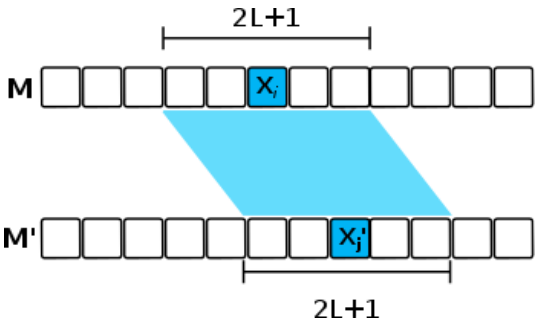
Trajectory Concatenation

Although the end-effector trajectories on which the LGB re-sampling scheme will be applied correspond to several realizations of the same semantic sequence, i.e., any magician trick, it is highly probable that each realization started and ended at a slightly different location

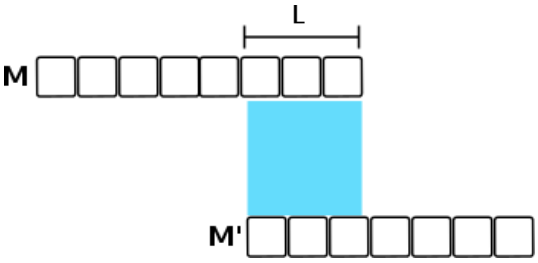
in space. Hence, there may be regions for which not many observations are available and the kernel estimates K_d and K_{dp} are not robust. These regions are thus rarely or even never explored during the re-sampling process.

In order to enhance LGB transition probabilities at such regions, closed and continuous end-effector trajectories are necessary. Such trajectories can be produced by smoothly transitioning between the beginning and end postures issued from each motion in our database. Namely, given all the motion sequences from which end-effector trajectories will be extracted, we automatically generate transition motions between the end of the first motion and the beginning of the second one, the end of the second motion and the beginning of the third one, and so on.

For transitioning from motion $\mathbf{M} = (\mathbf{X}_1, \cdots, \mathbf{X}_T)$ to motion $\mathbf{M}' = (\mathbf{X'}_1, \cdots, \mathbf{X'}_S)$ we have used a simplified version of the procedure proposed by Kovar *et al.* [142]. Instead of automatically determining the best transition point between **M** and **M'** and then blending the $2L + 1$ frames around this point as shown in Figure 5.5a, we have decided to blend the last L frames of **M** with the first L frames of **M'** (see Figure 5.5b).



(a) Blending around best transition point proposed by [142]



(b) Transition between the end and beginning of the two motions **M** and **M'**

Figure 5.5: Transitions procedures between two motions **M** and **M'**. Top figure depicts the original procedure proposed by Kovar *et al.* [142], bottom figure illustrates the simplified version we have adopted.

The transition is then generated as follows:

1. We align motion **M'** with respect to **M** using the rigid 2D transformation $\mathcal{G}_{\theta,x_0,y_0}$ that minimizes the squared distance between $\mathbf{M}'$ and the last L frames in **M**. The $\mathcal{G}_{\theta,x_0,y_0}$ transformation rotates all body joint positions about the vertical z-axis by θ degrees and then translates it by (x_0, y_0).

Given two character postures $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$ and $\mathbf{X}' = (\mathbf{x'}_1, \cdots, \mathbf{x'}_n)$ 3D Cartesian space (i.e., $\mathbf{X}, \mathbf{X}' \in \mathbb{R}^{3\times n}$), with $\mathbf{x}_i = (x_i, y_i, z_i)$ and $\mathbf{x'}_i = (x'_i, y'_i, z'_i)$, the 2D transformations $\mathcal{G}_{\theta,x_0,y_0}$ is computed as follows [142]:

$$\theta = arctan\ \frac{\sum_i w_i(x_i y'_i - x'_i y_i) - \frac{1}{\sum_i w_i}(\bar{x}\bar{y'} - \bar{x'}\bar{y})}{\sum_i w_i(x_i x'_i + y_i y'_i) - \frac{1}{\sum_i w_i}(\bar{x}\bar{x'} + \bar{y}\bar{y'})} \tag{5.12}$$

$$x_0 = \frac{1}{\sum_i w_i}(\bar{x} - \bar{x'}\cos(\theta) - \bar{y'}\sin(\theta)) \tag{5.13}$$

$$y_0 = \frac{1}{\sum_i w_i}(\bar{y} + \bar{x'}\sin(\theta) - \bar{y'}\cos(\theta)) \tag{5.14}$$

where all barred terms are defined in the same way, e.g., $\bar{x} = \sum_i w_i x_i$. In our implementation all n joints positions were attributed the same weight $w_i = 1/n$.

2. For each two corresponding frames $\mathbf{X}_{i\in\{T-L,\cdots,T\}}, \mathbf{X'}_{j\in\{1,\cdots,L\}}$, we apply linear interpolation on root positions and spherical linear interpolation on joint rotations. Interpolation weights correspond to L evenly spaced values between $[0, 1]$.

Once all motion transitions have been generated, the continuous end-effector and pelvis trajectories necessary for LGB re-sampling are computed using the forward kinematics operator $\mathcal{F}$ (Equation 5.2) on each one of the five articulated chains we previously defined (see Figure 5.2).

5.3.4 Overview of the Trajectory Generation Process

From Sections 5.3.2 and 5.3.3 we observe that the generation of expressive end-effector trajectories by re-sampling consists on a two-step process (see Figure 5.6). Given a set of n motions $\mathbb{M} = \{\mathbf{M}_1, \mathbf{M}_2, ..., \mathbf{M}_n\}$ expressing the same emotional state, we first generate continuous and closed trajectories by transitioning between every two contiguous motions, e.g., from $\mathbf{M}_1$ to $\mathbf{M}_2$, from $\mathbf{M}_2$ to $\mathbf{M}_3$, and so on. Second, we apply LGB re-sampling procedure in order to generate new expressive end-effector trajectories of length N. It is important to notice that since we have decomposed the animated character's body into several articulated chains, we can consider the end-effector trajectories of each articulated chain (e.g., head, right arm, or left arm) or a group of articulated chains (e.g., lower-body chains) as the observed sequence $X = (X_i)_{i\in\{1,\cdots,T\}}$ on which LGB re-sampling is applied. Hence, during the synthesis process a LGB re-sampling module is defined and adapted for each articulated chain or group of chains of interest as shown in Figure 5.6.

An important aspect of the synthesis process we have just described is the number of hyper-parameters to set and their influence in the resulting sampled end-effector trajectories. In the one hand, the duration K (on frames) of a transition motion should be such as to ensure a smooth change from one motion sequence to another and to preserve all emotion-related content. In the other hand, although the LGB re-sampling scheme makes few strong assumptions about the data, also referred to as observed sequence, and its generating process, it still requires to define several sensitive hyper-parameters. For instance, since the transition probabilities used to approximate the data-generating process by a *Markov chain*

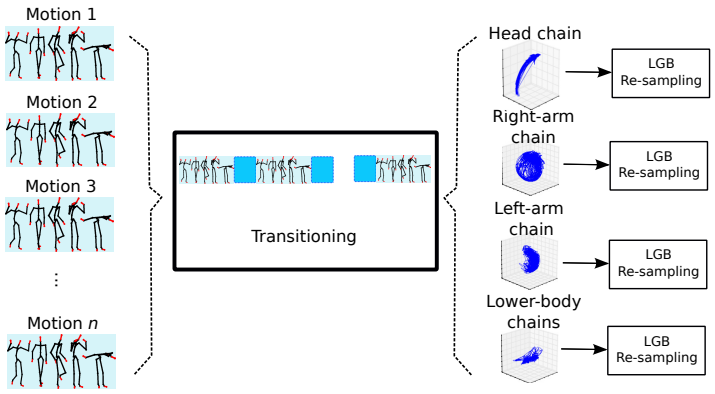


Figure 5.6: Overview of end-effector trajectory synthesis process. Two main steps are involved: *i)* motion transitioning and *ii*) application of LGB re-sampling scheme.

are estimated through kernel-based density estimators, the choice of the kernel bandwidth h_T is very important. Whereas a bandwidth too small will result in an *undersmoothed* estimate, a value too large will produce an *oversmoothed* density estimate [257]. Furthermore, even if a near to optimal choice of kernel bandwidth h_T is made, the density estimates may also be sensitive to the choice of the width of the local neighborhood of the current state σ_T. Hence, Monbet *et al.* [177] suggest to adapt both σ_T and h_T such as to find n_{min} number of neighbors and to consider a minimum of v observations during the averaging process of the kernel estimation.

The grid step Δ_g and the grid edge length σ_g are also sensitive parameters. They both determine the number of unobserved states included at each sampling step. When choosing these two parameters we must consider the following observations made by Monbet *et al.* [177]: *i)* when the ratio $\frac{\sigma_g}{\Delta_g}$ tends to zero, the number of unobserved states tends also to zero and *ii)* if σ_g is large compared to the kernel width h_T, the probability of sampling unobserved states is too small. Both σ_g and Δ_g should be chosen such as both the probability of reaching unobserved states and the number of accessible states at each sample step are sufficiently large [177].

In the following section we detail the tasks used to determine whether both the IK-based motion reconstruction and the proposed trajectory synthesis process generate expressive whole-body motions. We also explain how all aforementioned parameters were defined and which insights guided our choices.

5.4 Synthesis Tasks

We have introduced a synthesis approach based on two main components: expressive end-effector trajectories and an inverse kinematics based mapping from low-dimensional (i.e., end-effector and pelvis trajectories) to high-dimensional space (i.e., full-body motions). Specifically, we have argued that:

i) the end-effector trajectories we have selected, i.e., *head*, *hands*, *feet*, and *pelvis*, encode most of the patterns necessary for conveying affective content through motion, and

ii) given a set of those expressive trajectories, equally expressive bodily motions can be obtained by applying IK controllers on a set of articulated chains. Each chain seeks to follow its respective end-effector's trajectory.

In order to validate those two components, we propose two distinct yet complementary synthesis tasks: *motion reconstruction* and *motions from sampled trajectories*. In the first task, we seek to evaluate whether the proposed IK implementation generates motions that are similar to those from which the end-effector trajectories guiding the reconstruction were extracted. Through the second task, we wish to assess if the whole-body motions obtained from the randomly sampled trajectories exhibit the same affect-related patterns than the motions in our database. Both tasks are considered for each one of the five emotional states we analyzed in this thesis, i.e., *happiness*, *neutral*, *relaxedness*, *sadness*, and *stress*.

The aforementioned tasks are implemented using the data from a single actor. We do so in order to reduce as much as possible any confounding effect that might influence the quantitative and qualitative validation of both tasks and the two components of the proposed synthesis approach. The data we use accounts to 24 motion sequences from the magician scenario: 6 sequences depicting happiness, 6 realizations for the neutral state, 4 sequences for relaxedness, 5 sequences for sadness, and 3 examples for stress.

5.4.1 Motion Reconstruction

Following the results discussed in Chapter 4, we found that end-effector trajectories contain enough information such as to produce classification results similar to those obtained when employing the trajectories of all joints within the actors' body representation. However, so far we have not yet assessed how this information propagates to full-body motions obtained from those trajectories. The *motion reconstruction* task aims to provide some insight about this issue.

Specifically, we seek to evaluate the quality of the whole-body motions reconstructed through the combination of observed end-effector trajectories and IK-based motion controllers. This evaluation is based on the qualitative and quantitative similarity of the emotion-related content of the original motion $\mathbf{M}$ and its reconstructed image $\hat{\mathbf{M}}$.

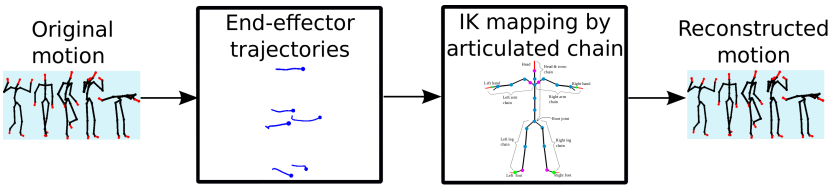


Figure 5.7: Reconstruction task

In this task we proceed as shown in Figure 5.7. For each motion $\mathbf{M}$ we extract the trajectories of the end-effectors associated to each one of the articulated chains enumerated in Section 5.2.3. That is, the head trajectory for the head-torso chain, left and right hand trajectories for the left and right arm chains respectively, left and right foot trajectories for the left and right leg chains respectively, and the root joint trajectory as well. Additionally, as explained in Section 5.2.4, we consider also the elbow trajectories for both arm chains in order to further constraint the space of possible solutions. Each trajectory is expressed with respect

to the coordinate frame system associated to its respective articulated chain; all trajectories are invariant to the position of the character's body in world space.

Once all target trajectories have been defined, we compute the reconstructed motion $\hat{\mathbf{M}}$, posture by posture, for which the articulated chains follow as smoothly and accurately as possible the extracted trajectories. We have then that the motion $\hat{\mathbf{M}}$ is the result of applying the IK controllers associated to all articulated chains for each target state $\mathbf{z_d}$ as listed in Equation (5.4).

5.4.2 Synthesis of New Motions Using Sampled Trajectories

Although the *motion reconstruction* task is a means to measure how expressive are the full-body motions generated from expressive end-effector trajectories, we cannot forget that the trajectories we use are readily obtained from motion sequences whose affective content is known and validated. To fully assess whether expressive end-effector trajectories are indeed a convenient motion model for generating new expressive body movements, it is necessary to produce new trajectories and evaluate the whole-body motions obtained from them. We propose to do so through the *motions from sampled trajectories* task.

In this task, new expressive end-effector trajectories are generated for each target emotional state using the procedure described and depicted respectively in Section 5.3.4 and Figure 5.6 respectively. Once we have obtained target trajectories for each one of the articulated chains listed in Section 5.2.3, we proceed to generate a new motion, posture by posture, using the propose IK-based mapping. The main difference between our two synthesis tasks lies in the origin of the end-effector trajectories guiding the synthesis process. All the steps involved in this synthesis task are illustrated in Figure 5.8.

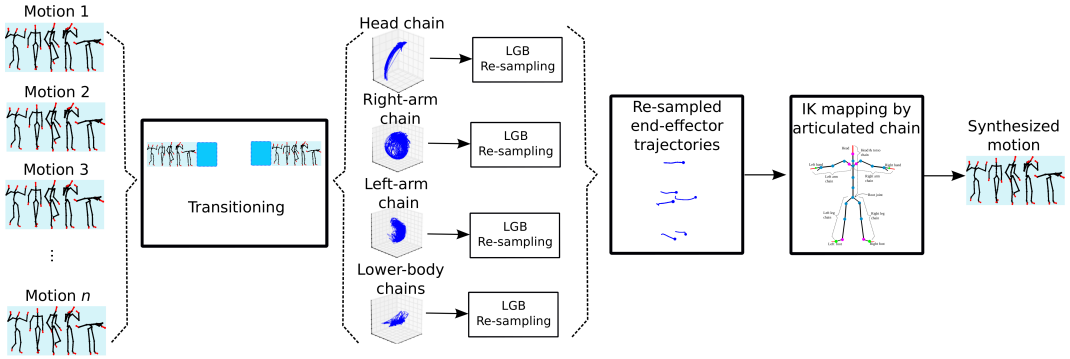


Figure 5.8: Steps involved in the *motions from sampled trajectories* task.

Experimental Setup for the Generation of New End-Effector Trajectories

For each targeted emotional state, transition motions are generated using the last and first $L = 250$ frames of any two contiguous motion sequences $\mathbf{M}$ and $\mathbf{M}'$ respectively. This value was selected such as to ensure that all individual actions within any magician sequence are completed before and after transitioning from one sequence to another. Namely, by using transitions of 250 frames long we are sure of transitioning between the end and beginning

of the bows the magician makes to the public at the introduction and conclusion of each sequence.

In order to produce the trajectories necessaries to control the IK controller associated to each articulated chain, each new time series generated by LGB re-sampling scheme consists of four processes characterized respectively by: the 3D head, the 6D right elbow-hand, the 6D left elbow-hand, the 6D+3D lower body trajectories. Each of these processes is independently re-sampled according to the LGB procedure for which X_t corresponds to the 3D or 6D trajectories. In the case of the leg chains, we have decided to sample them together in order to guarantee the character's stability during motion. Additionally, since the motion of the character's body root joint highly depends on the leg chains displacement and vice-versa, every time a new state is sampled, the corresponding root's position is taken from the observed sequence $X = (X_i)_{i\in\{1,\cdots,T\}}$ associated to the character's lower-body motion.

The order of the Markov process representing each observed sequence has been empirically adjusted to $p = 1$ for the 6D elbow-hand and 6D+3D lower-body trajectories, and to $p = 3$ for the 3D head trajectories. This choice was made based on the density of the area explored by each articulated chain and so that the synthesized trajectories were reliable and smooth enough after considering the number of points and their dimensionality by emotional state.

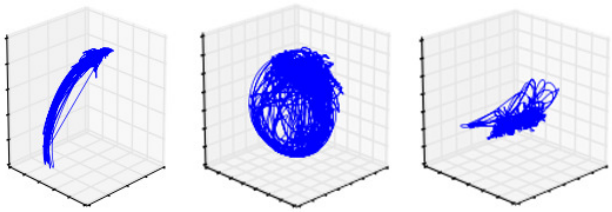
For each process the grid parameter σ_g is locally adjusted so as to maintain between 50 and 100 observations inside the neighborhood $\widehat{VY}_t$. The parameter h_T is set up to $\sigma_g/3$, which seems to nicely fit with the implementation of a grid (hypercube covering the image $\widehat{VY}_t)^+$) that has 3 subdivisions along each of the 3 or 6 dimensions that are considered. A Kd-tree is used to index $p.d$ dimensional samples collected along the observed end-effector trajectories. The search for the neighborhood of the current $\hat{Y}_t$, which conditions the algorithmic complexity of the method, is near logarithmic with the size of the data as far as $p.d$ is sufficiently small, basically below 20.

For each emotional state re-sampled end-effector trajectories of $N = 30000$ *frames* are generated. Figure 5.9 shows for the emotional state *happiness* the observed and re-sampled 3D submanifolds obtained from the head, right hand and right foot trajectories.

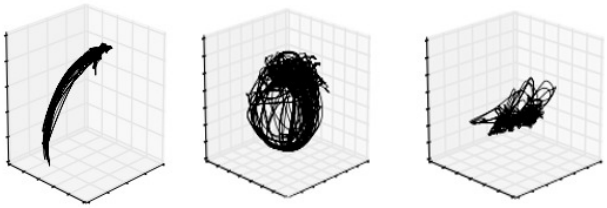
5.5 Quantitative Evaluation

In the character animation domain, in particular for data-driven approaches, quantifying the resemblance or similarity between two or more motions is of great importance. Not only is a key step for popular methods and applications such as *motion graphs* [113, 142, 154] and *motion retrieval* [128, 141], but it is also required for evaluating the performance and quality of the motions produced by a determined method [199, 204].

Similarity metrics can also be useful to represent the characteristics inherent to a particular set of motion sequences, e.g., different ways of performing the same logical action or different logical actions expressing the same affective content, and to consequently determine if a given motion exhibits such characteristics [189, 249]. In this thesis we are interested in this latter application of similarity metrics. Specifically, we wish to objectively evaluate whether the motion sequences obtained from the two aforementioned synthesis tasks share the same patterns associated with the expression of affect we found in the ground truth data.



(a) Observed trajectories



(b) Sampled trajectories

Figure 5.9: Observed (top) and sampled (bottom) 3D submanifolds obtained from the trajectories, from left to right, of: head, right hand and right foot.

To measure the resemblance between two or more motions requires to define: *i.)* a way of encoding the attributes upon which the similarity will be measured, and *ii.)* a metric that specifies how the comparison will be made [144, 249]. Furthermore, the selected encoding should not only emphasize the important motion components on which we are interested, e.g., emotion, but also abstract out any other element that is not significant for the motion comparison, e.g., action or actor dependent information. In the same manner, the chosen metric or distance function should also account for all the aspects we expect to find are common to the two motion sequences being compared. For instance, common metric functions such as the Euclidean distance on joint positions and/or angles can successfully capture the structural similarity between two motions, but fail to determine if both motion are logically or contextually analogous [128]. As it has been mentioned all along this thesis, our aim is to generate emotionally expressive body movements. Hence, both the encoding and metric functions employed for assessing the motions obtained from the aforementioned synthesis tasks must account for the variations due to the expression of emotions and affect.

The results presented and discussed in Chapter 4 showed that the motion features (*velocity, acceleration, jerk,* and *curvature*) we used to represent joint trajectories and consequently motion sequences successfully encoded most of the underlying affective content. Thus, we have decided to use the same encoding when comparing the ground truth and synthesized motion sequences. Similarly, we have decided to use automatic affect classification and an information theory divergence measure as our metric functions. The classification-based metric will helps us to determine whether the synthesized motions belong to the affective space defined by the ground truth data. The divergence similarity metric will quantify how different are the affect-related variations observed in the ground truth data from those found in the synthesize motions.

5.5.1 Classification Similarity Measure

This metric function aims to determine whether the bodily movements generated by the *motion reconstruction* and *motions from sampled trajectories* synthesis tasks are recognized by an automatic classifier trained on the training dataset, i.e., the sequences from which the end-effector trajectories used on both tasks were extracted. Furthermore, by analyzing the recognition rates obtained from the classifier we will not only evaluate the quality of the generated motions, but also provide an initial answer to the following questions:

i.) how well does the expressive content encoded in the end-effector trajectories propagate through the whole character's body?,

ii.) how expressive are the end-effector trajectories generated by the re-sampling scheme previously described?, and

iii.) how expressive are the motions generated through the combination of end-effector trajectories re-sampling and IK reconstruction?

Evaluation Design and Setup

As we have already mentioned, end-effector trajectories are independently re-sampled from each other in the *motions from sampled trajectories* synthesis task. While this choice provides us with a more detailed and efficient control of the character's articulated body and a richer vocabulary of possible motions to synthesize [122, 123], it also complexifies the use of automatic affect classification as a quantitative evaluation metric. On the one hand, the co-articulation and synergy between articulated chains observed in the ground truth data is not longer present in the resulting motions (since the re-sampling is achieved chain by chain). On the other hand, the features' relationships implicitly learned and employed by the classifier when trained on feature vectors containing information for all body joints are also missing in the synthesized motions. Thus, it is likely that the classifier's performance will be influenced by this structural difference between the ground truth data and the generated whole body movements.

For this reason we have decided to employ a different classifier for each synthesis task. For the *motion reconstruction* task, which preserves all synergies and dependencies between articulated chains, a RF classifier with default parameters ($n\_tress = 500$, $m\_try = \sqrt{p}$, with p being the number of features used to represent each observation) is used. For the *motions from sampled trajectories* synthesis task, a combined classifier model is employed instead. Both classifiers follow the same classification procedure described in Chapter 4. That is, *motion chunks* and their respective feature vectors are computed for each motion sequence in the training and testings datasets, and sequences labels are estimated by averaging the probabilistic predictions provided by the classifier across all *motion chunks* belonging to the same motion. Default *motion chunk* parameters are used, i.e., *window_size*= 500 frames and *overlap*= 250 frames.

The combined classifier model extends the principle of independent articulated chains into the classification procedure we have just summarized. Instead of generating a unique feature vector for *motion chunk*, we associate a feature vector to each one of the chains for which end-effector trajectories are sampled. Only the body joints that belong to a given

articulated chain are considered when computing the chain's feature vector for the current *motion chunk*. We train then a RF classifier with default parameters for each controlled chain (i.e., head, right and left arms) or group of chains (i.e., lower-body). The final label predicted for a determined *motion chunk* is computed using the same majority vote scheme we employ for sequence label prediction (see Section 4.5). Different weights could be used for each individual classifier. For instance the classifier associated to the right arm chain might be more informative for right-handed subjects than for left-handed ones.

Below we present and discuss the results obtained when using the MoCap data and synthesize motions as training and testing sets respectively.

Classification Results: Motion Reconstruction Task

Table 5.1 (bottom) shows the confusion matrices obtained when using the motions generated by the *motion reconstruction* task as testing set. Left matrix correspond to recognition rates obtained on *motion chunks*, while right matrix shows the rates reported by the classifier on whole motion sequences. Results were obtained using features computed for all body joints (i.e., 216 features in total) and stratified 3-fold cross-validation. For each fold we excluded during training the corresponding original motion of each reconstructed sequence in the testing set. We have also included the results obtained when performing stratified 3-fold cross-validation on MoCap data only (top matrices in Table 5.1).

0.63	0.03	0.13	0.05	0.16
0.03	**0.90**	0.0	0.0	0.07
0.14	0.01	**0.83**	0.01	0.01
0.35	0.06	0.25	**0.24**	0.10
0.30	0.17	0.07	0.03	**0.43**

average rate: 0.61

(a) Confusion matrix (%) for *motion chunks* extracted from ground truth data.

0.93	0.0	0.0	0.0	0.07
0.0	**1.0**	0.0	0.0	0.0
0.0	0.0	**1.0**	0.0	0.0
0.60	0.0	0.23	**0.17**	0.0
0.13	0.0	0.0	0.0	**0.87**

average rate: 0.79

(b) Confusion matrix (%) for ground truth motion sequences.

0.63	0.02	0.17	0.04	0.14
0.03	**0.76**	0.0	0.01	0.20
0.10	0.01	**0.85**	0.04	0.0
0.36	0.08	0.28	**0.20**	0.08
0.28	0.11	0.10	0.02	**0.48**

average rate: 0.58

(c) Confusion matrix (%) for *motion chunks* extracted from generated motions (*motion reconstruction* task).

1.0	0.0	0.0	0.0	0.0
0.0	**1.0**	0.0	0.0	0.0
0.0	0.0	**1.0**	0.0	0.0
0.67	0.0	0.33	**0.0**	0.0
0.0	0.25	0.0	0.0	**0.75**

average rate: 0.75

(d) Confusion matrix (%) for generated motions (*motion reconstruction* task).

Table 5.1: Confusion matrices (%) for ground truth data (top) and movements generated by *motion reconstruction* synthesis task (bottom). Emotions are listed in the order *neutral*, *sadness*, *happiness*, *stress*, and *relaxedness*.

At a large scale, we observe that for four of the target emotional states, i.e, *neutral*, *sadness*, *happiness*, and *relaxedness*, the synthesized motion sequences were recognized above

chance level with an 75% average recognition rate (see Table 5.1d). The reconstructed motions conveying *stress* were instead labeled as either *neutral* or *happy* motions. However, from Table 5.1b we find that the classifier was also unsuccessful to accurately recognize ground truth motions conveying *stress* above chance level. It is probable that, in both cases, the classifier's difficulty to distinguish *stress* from the other emotional states was due to the limited number of examples available within the learning dataset; we could only use three examples in total. We observe also that the differences between synthesized and original motions change according to the target emotional state. For instance, while the accuracy rates of sequences conveying *happiness* increased for the synthesized motions (100%) with respect to the ground truth data (93%), the opposite happened for the *relaxedness* state. Namely, the accuracy rates registered by the classifier for the synthesized motions (75%) decreased with respect to the original motions (87%). No difference was found on the accuracy rates on synthesized and original motion sequences conveying *sadness*.

Although the average recognition rate across classes decreased of 3% and 4% for *motion chunks* and sequences respectively, we find that the automatic recognition rate of the generated movements is still high and demonstrates that end-effector trajectories encoded sufficient expressive information. Similarly, this first result proves that, from a quantitative perspective, a simple IK-reconstruct suffices to propagate affective content through the animated character's body and thus generate emotionally expressive bodily motions. We believe that recognition rates for the reconstructed motions can be increased with the improvement of our IK reconstruction procedure.

Classification Results: Motions from Sampled Trajectories Task

Since a unique re-sampled end-effector trajectory was generated for each articulated chain controlled during the *motion from sampled trajectories* task, we present only recognition rates registered for *motion chunks*. The combined classifier model we previously described was trained using all motion sequences belonging to the ground truth data. Once again, we used feature vectors containing information from all body joints (i.e., 27 joints, 216 features in total). Recognition rates listed in Table 5.2b were obtained by testing the classifier on *motion chunks* extracted from the 30000-*frames* long generated sequences. For making the comparison with MoCap data possible, accuracy rates on the ground truth data were obtained using the same combined classifier and a stratified 3-fold cross-validation procedure. Results obtained for *motion chunks* are listed in Table 5.2a.

From Table 5.2b we observe that *motion chunks* from all target emotional states were recognized above chance level at an average recognition rate of 60%. After comparison with the results obtained on ground truth data only and reported in Table 5.1a, we find that: *i.)* the average recognition rate across all class decreased only by 3% and *ii.)* misclassification patterns between the two confusion matrices are considerably different. For instance, although *relaxedness* is recognized above chance level (39%) for the *motion chunks* obtained from the generated motions, it is also labeled as *neutral* for 50% of the cases with respect to the 24% of the cases reported by the ground truth data. Similarly, *motion chunks* conveying the *neutral* state are often labeled as *happiness* rather than *relaxedness* as it was observed on ground truth data. These changes can be due to either differences in the re-sampled end-effector trajectories or noise introduced by the IK reconstruction procedure.

To gather more information and a better understanding of the reasons behind the

0.68	0.04	0.10	0.02	0.16
0.02	**0.88**	0.0	0.0	0.10
0.13	0.01	**0.84**	0.02	0.0
0.34	0.07	0.26	**0.28**	0.05
0.24	0.17	0.07	0.03	**0.49**

average rate: 0.63

(a) Confusion matrix (%) for *motion chunks* extracted from ground truth data.

0.54	0.0	0.32	0.04	0.10
0.10	**0.73**	0.06	0.03	0.08
0.0	0.0	**1.0**	0.0	0.0
0.33	0.0	0.30	**0.34**	0.03
0.50	0.0	0.10	0.01	**0.39**

average rate: 0.60

(b) Confusion matrix (%) for *motion chunks* extracted from generated motions (*motion from sampled trajectories* task).

Table 5.2: Confusion matrices (%) for ground truth data (left) and movements generated by *motions from sampled trajectories* synthesis task (right). Results were obtained at *motion chunk* level with combined classifier. Emotions are listed in the order *neutral*, *sadness*, *happiness*, *stress*, and *relaxedness*.

changes we previously mentioned, we trained a RF with default parameters on end-effector trajectories only and tested it against the re-sampled end-effector trajectories. A total of 48 features were used to encode all *motion chunks*, i.e., features were computed from *head*, *hands*, *feet*, and *pelvis* joints. We also computed recognition rates on ground truth data only using anew stratified 3-fold cross-validation. Confusion matrices obtained for both classification scenarios are listed below:

0.65	0.03	0.14	0.04	0.14
0.03	**0.91**	0.0	0.01	0.05
0.11	0.01	**0.84**	0.03	0.01
0.40	0.08	0.26	**0.21**	0.05
0.25	0.16	0.07	0.01	**0.51**

average rate: 0.62

(a) Confusion matrix (%) for *motion chunks* extracted from ground truth end-effector trajectories.

0.58	0.0	0.28	0.06	0.08
0.09	**0.73**	0.05	0.05	0.08
0.0	0.0	**0.99**	0.0	0.01
0.25	0.0	0.26	**0.46**	0.03
0.38	0.0	0.10	0.02	**0.50**

average rate: 0.65

(b) Confusion matrix (%) for *motion chunks* extracted from re-sampled end-effector trajectories.

Table 5.3: Confusion matrices (%) for ground truth data (top) and movements generated by *motion reconstruction* synthesis task (bottom). Emotions are listed in the order *neutral*, *sadness*, *happiness*, *stress*, and *relaxedness*.

A careful comparison between the confusion matrices listed in Table 5.3 indicates that whereas the changes on the classification patterns of *motion chunks* conveying *relaxedness* may be due to noise or artifacts introduced by the IK reconstruction procedure, this is not the case for *motion chunks* belonging to the *neutral* state. It seems that re-sampled trajectories for the latter emotional state are frequently considered as conveying *happiness*, which explains why 32% of the whole-body *motion chunks* belonging to the *neutral* synthesized sequences were labeled as conveying *happiness*. Nevertheless, the results listed in Table 5.2b and Table 5.3 indicate that re-sampled trajectories for most of the target emotional states are as expressive as the ground truth data. Furthermore, whole-body motions generated from re-sampled end-effector trajectories are as expressive as the motions found in the ground truth data.

5.5.2 A Divergence-Based Similarity Measure

Although our classification results indicate that both synthesis tasks generate motions that are congruent with the target emotional states contained in our database, we wished to further measure the similarity between the affective content encoded in the generated sequences and the ground truth data. Inspired by recent work done in motion retrieval [226] in which motions are compared based on their content and variations rather than their numerical proximity, we have a similarity measure based on information theory metrics.

Statistically speaking, the expressive content of a motion can be described and quantified by estimating the probability distributions of the features considered as the most salient during emotion and affect discrimination. By comparing the estimates of the ground-truth data to those of the synthesized motions, it is possible to have a quantitative measure of their similarity.

When measuring the similarity between two probability distributions P and Q, relative entropy measures such as the *Kullback-Leibler* (KL) divergence are widely used. However, KL divergence is non-symmetric, i.e., $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, and has no upper-bound, which makes it harder to interpret. For these reasons, the *Jensen-Shannon* (JS) divergence measure has been used instead. Let $\mathcal{P}$ denote the space of all probability distributions defined over a discrete set of events Ω. The JS divergence is a function $\mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}$ defined by:

$$D_{JS}(P,Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M) \tag{5.15}$$

where $D_{KL}(P,Q) = \frac{1}{2}\sum_{x\in\Omega} P(x)\log\frac{P(x)}{Q(x)}$ and $M(x) = \frac{1}{2}P(x) + \frac{1}{2}Q(x)$ is the mean distribution. The JS divergence is equal to zero (0) only if the distributions are identical, and therefore are indistinguishable, and reaches its maximum value of one (1) if the two distributions do not overlap, thus showing that there is not similarity between them. Thus, the closer the JS score is to zero (0), the more the affective content that is conveyed by both the ground truth data and the synthesized motions is similar.

In our case, the probability distributions P and Q are associated to two distinct movement generation sources. For each salient feature f_i^*, we compute the distribution P using the motions from the first generation source and the distribution Q using the motions from the second source. Both distributions are empirically estimated through 1-dimensional histograms. We then proceed to compute $D_{JS}(P,Q)$.

Selection of the Most Salient Features

One of our main goals has been to show that end-effector trajectories are a sufficient low-dimensional representation that can be used as control signal for the synthesis of expressive motions. Hence, it is important that the comparison between MoCap and synthesized motions is made on a subset of features considered as the most salient when characterizing the expression of affect and emotions. This will show that the affective content encoded in the target trajectories is also present in the synthesized motions and exhibits the same patterns than the ground truth data.

Supported by the results presented in Chapter 4 which indicate that a sufficiently small set features can characterize the emotional content in human movement, we have decided

to use only a limited number of such features when computing the JS divergence score between the ground-truth data and the synthesized motions. This set can be defined through an automatic feature selection procedure. The procedure – similar to the one employed in Chapter 4 – uses a RF model as the base classifier of a wrapper algorithm in which a recursive backward elimination strategy selects the most salient features.

Data: X: dataset, α: set of possible hyper-parameters, d: number of least important features to remove, n_rep: number of runs for each ranking
Result: tuple (α^*, p^*)
Divide dataset X into K-folds;
for *I from 1 to K* **do**
  Define set L as dataset X without the I-th fold;
  Define set T as the I-th fold of the dataset X;
  for α_m *from 1 to* $|\alpha|$ **do**
    selected = all features in dataset X;
    $n = |selected|$;
    while $n>0$ **do**
      Define L' as set L with selected features on it;
      Define T' as set T with selected features on it;
      for *i from 0 to n_rep* **do**
        Build statistical model $f' = f(L', \alpha_m)$;
        Store ranking from f';
        Apply f' on T' and store test error;
      end
      Calculate average ranking and test error;
      Remove from *selected* the d least important features according to average ranking;
      $n = |selected|$;
    end
  end
  Calculate average test error for each point (n, α_m) across K folds;
  Define the pair (p^*, α_m) with minimal test average error as the optimal test error model;
end

Algorithm 5: Cross-validation procedure for parameter selection.

The selection procedure described below is based on the work done by [228] and has been extended to deal with the selection of RF hyper-parameters. The essential idea of this procedure is to redefine the problem of feature selection and hyper-parameter fitting as a problem of parameter selection. It seeks the tuple $(p^*, n\_trees^*, m\_try^*)$, where the number of trees, n_trees^*, and the fraction of variables to consider during a split, m_try^*, define the optimal RF model, and p^* corresponds to the smallest number of features to consider when quantitatively comparing ground truth and synthesized motions.

The optimal $tuple^*$ is determined by cross-validation, and when a final set of features to use for motion comparison is required, one trains the RF classifier with hyper-parameters n_trees^* and m_try^* on all the data, ranks the features, and applies the same recursive elim-

ination procedure until only the most important $p*$ features are left. The optimal tuple is selected based on the smallest average test error across folds. The specific steps of the cross-validation procedure are listed in Algorithm 5.

Since the final goal of this procedure is to select the smallest possible feature subset, we have decided to also consider other tuples for which the error rates are within the one standard error rule (1 S.E) during the selection of the optimal parameters $tuple^*$. If possible, a substitute for the optimal tuple with a smaller number of features is selected such that it does not lead to a significant increase in the error rate. The feature selection procedure starts with the initial 216 features and uses *motion chunks* defined by a sliding window of $500-$*frames* length and with an overlap between any two contiguous *motion chunks* of 250-*frames*. A total of 20 different RF models are tested; a model for each possible combination of the following hyper-parameters: $n\_trees = \{400, 800, 1200, 1600, 2000\}$ and $m\_try = \sqrt{p}, \log_2(p), 0.1875 \times p, 0.4 \times p$, with p being the current dimensionality of the feature vectors computed from *motion chunks*. At each iteration the number of features to be considered decreases by 4. Results depicted in Figure 5.10 have been obtained through stratified 5-fold cross-validation. The optimal tuple and the substitute model selected using the 1 S.E are $(n\_trees = 1200, m\_try = log_2(p), p = 180)$ and $(n\_trees = 1200, m\_try = \sqrt{p}, p = 164)$ respectively.

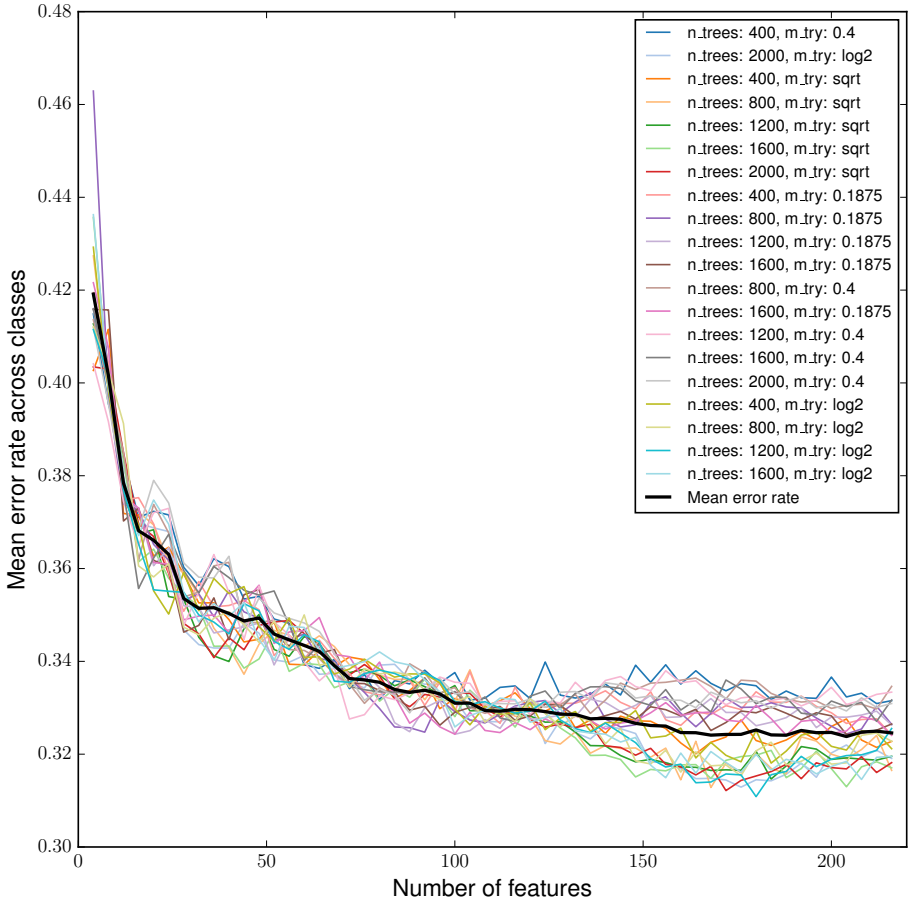


Figure 5.10: Error rates obtained during parameter selection procedure. Each colored line indicates the changes on the error rate of different RF model as the number of features decreased. n_trees = number of trees, m_try = fraction of total features to be considered during a split.

JS Divergence Scores for Generated Motions

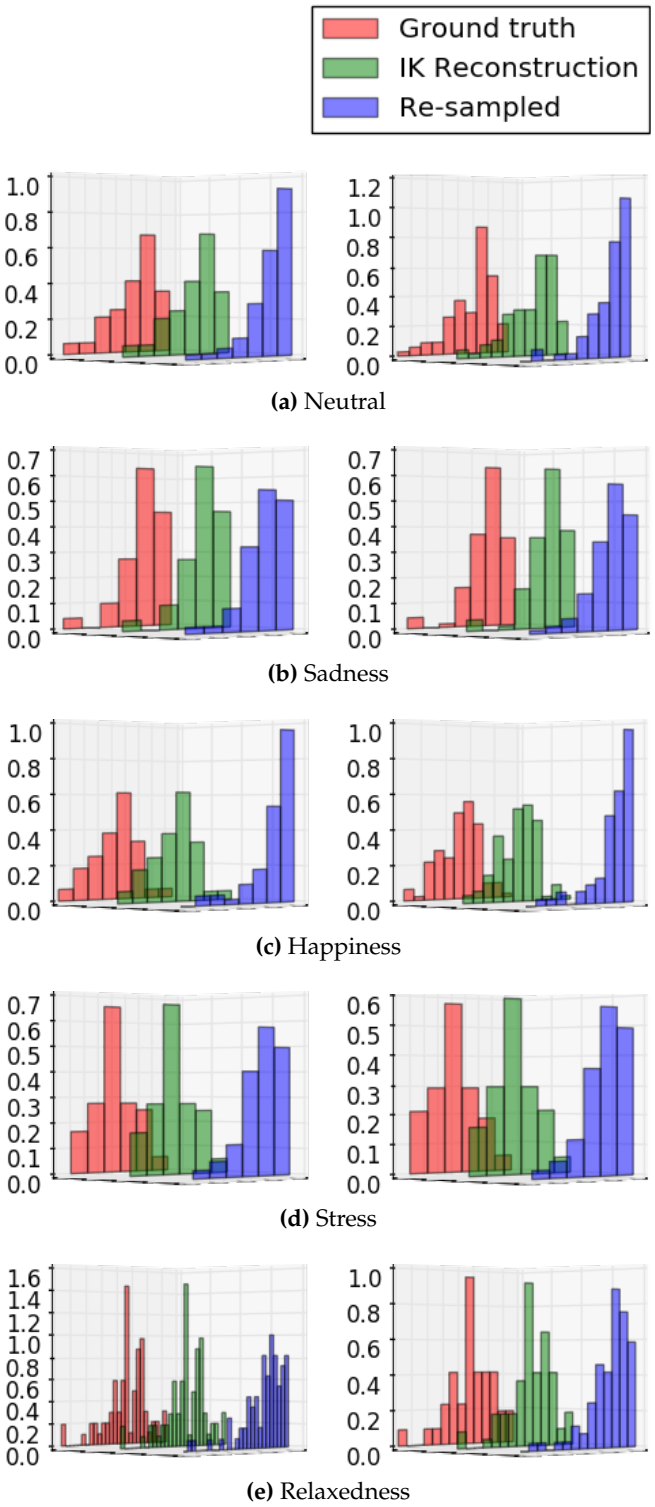
One-dimensional histograms are used to empirically approximate the probability distributions of each one of the $p* = 164$ features selected as the most important. Figure 5.11 depicts one-dimensional histograms computed for the two most salient features for each emotional state. The number of bins used by each histograms is automatically determined using the procedure described in [223]. Using JS divergence measure, histograms are compared bin-by-bin, i.e., only pairs of bins with the same index are compared. The weighted average of the 164 divergence measures has been defined as the final similarity score between the two motions. The weights associated to the individual JS divergence scores correspond to the importance feature measures computed by the RF classifier and associated to the features for which the JS scores are computed. In this way, the contribution of all features into the final similarity score is directly proportional to their importance.

JS divergence measures were computed for three evaluation scenarios:

- **Similarity base-line**: Before comparing the MoCap ground truth data and the motions generated by each synthesis task, it is necessary to define an initial baseline score by emotional state. These scores will provide us with a reference against which compare the divergence scores obtained for each synthesis task. A leave-one-out cross-validation procedure has been used for that purpose. For each target emotional state, $n_i - 1$ sequences are used to define the empirical probability distribution P while the remaining sequence represents the distribution Q; n_i indicates the number of sequences belonging to the i-th emotional state. Both distributions are computed for each one of the 164 most salient features. JS divergence measure is applied on P and Q for all possible $n_i - 1$ combinations. The similarity score for each affective state corresponds to the average JS measure across all folds.

- **Ground truth data (MoCap) vs. motions generated by the *motion reconstruction* task (Task 1)**: Since there is a one-to-one correspondence between ground truth sequences and generated motions – end-effector trajectories from MoCap sequences guided the IK reconstruction of each generated motion –, it is possible that the JS scores obtained by comparing all MoCap motions against all generated motions are overoptimistic. For this reason, we decided to employ the same leave-one-out cross-validation procedure than the previous scenario. However, the remaining sequence used to approximate the distribution Q is taken from the generated motions rather than from the MoCap data. The similarity score for each emotional state corresponds to the average JS measure across all fold.

- **Ground-truth data (MoCap) vs. motions generated from re-sampled trajectories (Task 2):** Sequences are separated by emotional state. In this third scenario the probability distributions P and Q are computed from MoCap data and the motions generated by the synthesized trajectories and the IK controllers respectively. It is possible that the divergence scores obtained for this scenario are slightly smaller than those obtained from the other two evaluation scenarios. This is due to the number of sequences used to empirically approximate the distribution Q in each scenario. While in the first and second scenario only one motion sequence (a unique realization of a magician trick) with an average duration of 4000 *frames* is used, in the third and last scenario we are using an 30000-*frames* long sequence.



(a) Neutral

(b) Sadness

(c) Happiness

(d) Stress

(e) Relaxedness

Figure 5.11: Histograms generated for the two most important features: pelvis and lower-torso joints' average velocity for each emotional state. Histograms were computed for each movement generation source: ground truth data (red), *motion reconstruction* task (green), and *motions from re-sampled trajectories* task (blue).

Scenario	*Neutral*	*Sadness*	*Happiness*	*Stress*	*Relaxedness*
Base line	0.144	0.134	0.132	0.117	0.131
MoCap vs. Task 1	0.157	0.158	0.153	0.123	0.152
MoCap vs. Task 2	0.084	0.0978	0.0840	0.0905	0.160

Table 5.4: *Jensen-Shannon* divergence measures for each emotional state.

Table 5.4 lists the results we obtained for the three evaluation scenarios we just described. We observed that for both types of synthesized motions the divergence measures are at the same time close to the base line and to the lower-bound of the *Jensen-Shannon* measure. Thus, as measured by the 164 features used to characterize the affective content within the different motions, the generated whole body movements are significantly similar to the ground truth data from which end-effector trajectories were extracted. From the first and second row in Table 5.4, we notice also that our IK-based mapping from end-effector trajectories to whole body motions (second row) resulted only in an average increase of 0.017 with respect to the divergence scores of ground truth data (first row). This indicates that although some noise might have been introduced during the mapping, most of the variations related to the expression of affective content through bodily motions are successfully encoded by the end-effector trajectories and propagated by the IK mapping. We can conclude then that from a quantitative standpoint both the proposed motion model, i.e., expressive end-effector trajectories and the designed synthesis approach generate emotionally expressive body motions.

5.6 Qualitative Evaluation: User Study

Human interpretation of emotionally expressive body movements although intuitive can also be highly subjective. Because of this, in addition to the quantitative evaluation of both synthesis tasks, subjective measures are needed. These measures assess whether the perception of affect and emotions changes in the light of the source from which the motions to evaluate are obtained.

Specifically we have conducted a user study in which participants rated the emotional content and expressiveness of: *i)* the motions performed by an actor and used both as learning dataset and ground truth data, *ii)* the IK reconstruction generated for each sequence in the learning dataset (first synthesis task), and *iii)* the motions obtained from re-sampled expressive end-effector trajectories (second synthesis task). The results obtained from the first set of motions, i.e., learning set, provide us with a base-line of perceived expressiveness against which we can compare the perception rates obtained for the synthetically generated motions. The design and results observed for this subjective evaluation are discussed below.

5.6.1 User Study Design

For this new user study, we have employed the same questionnaire used for the evaluation of our MoCap database extension (see Section 3.7.2). In this questionnaire – composed of four different questions – participants are asked to assess the expressiveness and affect-related content of a group of motions. For each motion to be evaluated, participants are first asked to select, among five options, which emotion is being displayed by the animated character;

second, they are asked to rate from 1 to 7 the intensity with which the emotion is being conveyed; third, using the same scale, the participants rate the current motion along the valence-arousal axes; and finally, we asked them to assess, from 1 to 7, the difficulty of evaluating such motion.

For this study 10 video clips at 30 fps were created for each movement generation source, i.e., *MoCap*, IK reconstruction or re-sampled end-effector trajectories. For the *MoCap* source, 2 realizations by emotional state were randomly selected among the 24 available sequences. In the case of the stimuli belonging to the IK reconstruction source, video clips were generated by applying the already described IK controllers on the the end-effector trajectories of each one of the 10 realizations representing the *MoCap* source.

The stimuli associated to the last generation source, i.e., re-sampled end-effector trajectories, were obtained through a two-step process. First, for each emotional state, a group of motion candidates (approximately 4 candidates by emotion) was generated by selecting several motion segments from the 30000 *frames* long synthesized motion sequences. These segments were considered to be good examples of each target emotional state. Second, two annotators were asked to choose the two most expressive motion segments by emotional state. Videos for which there was a complete agreement between the annotators were automatically selected. If there was no agreement, we randomly selected one or two videos from the lists provided by the annotators for each emotional state. We found that there was no agreement for only the second stimuli for the *neutral* state.

To correctly assess the effect of movement generation source on the perception of emotional states, each source was evaluated separately. More precisely, participants were randomly assigned to either the *MoCap*, IK reconstruction, or re-sampled end-effector trajectories condition. In this way, we sought to discard any possible carry-over effect between generation sources and guaranteed that participants would remain naive to the main purposes of the study. Our user study has intended emotion as a within-subject factor, whereas movement generation source is considered as a between-subject factor instead.

In our previous studies, participants often remarked that assessing emotional states from the motions we presented to them was harder because they had no base line against to which make comparisons/judgments. It seems that although a training stage was added to show the spectrum of possible expressive movements, this information was not sufficient. For this reason, we have introduced a slight modification into the dynamics of the user study design we proposed in Chapter 3. We eliminated the training stage and instead of showing each video clip once, participants were asked to rate the same video twice. Video clips were presented in random order, however we made sure that all video clips belonging to one condition were rated once before showing any video a second time. Through this modification, we expect to provide enough information to make the task less complex while avoiding, at the same time, any possible lingering learning effect.

A total of 72 participants took part in this new user study. In total, we had 35 women and 37 men, ranging in age from 22 and 69 years old. Since we have three movement generation sources to be evaluated separately, 24 participants were randomly assigned to each source; the same participant could not be appointed to more than one source. Once again, we used Amazon Mechanical Turk (MTurk) [55] as an intermediate platform for recruiting participants and conducting our study. Since each video clip was rated twice, each participant was presented with 20 videos in total. A consent form was electronically signed by each participant prior to the start of the survey. Participants took in average approximately

25 minutes to answers all questions. As done in Chapter 3, participants considered as outliers were detected using the procedure already detailed in Section 3.7.3. That is, participants whose agreement score within their group was outside a determined interval were marked as outliers. Among the initial 72 participants, only 2 outliers were detected; one belonged to the IK reconstruction group, the other to the re-sampled end-effector trajectories group.

5.6.2 The Effect of Movement Generation Source and Intended Emotion

Two-way repeated measures ANOVA was used to evaluate the main and interaction effects of intended emotion and movement generation source on the perception of emotionally expressive body movements. In this analysis, emotion and generation source where modeled as within-subject and between-subject factors respectively. Five ANOVAs tests were performed on the average accuracy, intensity, valence, arousal, and difficulty ratings across participants. Following the same notation used in Chapter 3, we list below the null hypotheses evaluated in this study:

> $H_0(1, i)$: *The means of the participants' ratings of i for the different intended emotions are equal.*
>
> $H_0(2, i)$: *The means of the participants' ratings of i for the different movement generation sources are equal.*
>
> $H_0(3, i)$: *Movement generation source and intended emotion are independent factors and no interaction between the two is present on the participants' ratings of i.*

With $i = \{$*accuracy, intensity, valence, arousal, difficulty*$\}$. Table 5.5 list the resulting F-statistics, p-values and effect sizes (η^2). Effects and interactions were evaluated at a significant level of $\alpha = 0.05$. P-values indicating a significant difference were highlighted.

In agreement with our previous user studies and as it is listed in Table 5.5, we found that the intended emotion has a significant effect in all cases ($p < 0.0001$); hence we reject the null hypotheses $H_0(1, i)$ for $i \in \{$*accuracy, intensity, valence, arousal, difficulty*$\}$. This effect is large in size ($\eta^2 > 0.16$), which suggests that the variance observed on participants' ratings is mostly due to the affective content conveyed through the character's motion. Furthermore, this detected main effect indicates that both synthesis tasks successfully encoded the differences between the intended emotions in the generated whole-body motions. Thus, from a perceptual point of view, it seems that as long as expressive end-effector trajectories guide the synthesis process, affective content is still successfully conveyed and perceived in the synthesized motions.

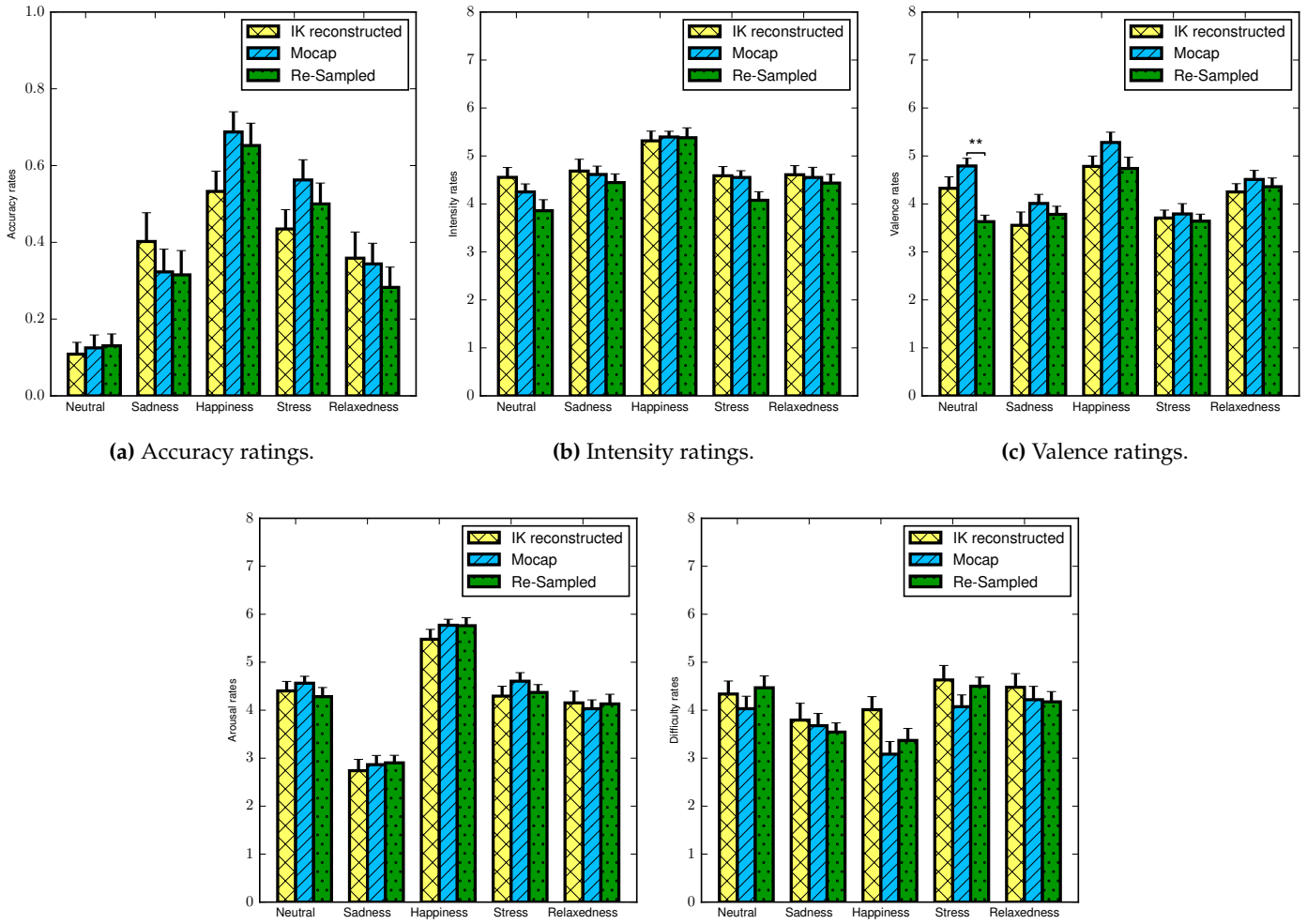
The effect of the movement generation source and the paired differences in the participants' ratings of accuracy, intensity, arousal and difficulty with respect to this factor (see Figures 5.12a, 5.12b, 5.12d, and 5.12e) were found no significant since the corresponding p-values are greater than $\alpha = 0.05$. Therefore, we retain the four null hypotheses $H_0(2, i)$ for $i \in \{$*accuracy, intensity, arousal, difficulty*$\}$, which state that the average ratings of movements from different generation sources are equal. However, since a significant effect of generation source on valence ratings was found ($p < 0.05$), we reject $H_0(2, \textit{valence})$ in favor of the alternative hypothesis.

	Ratings(i):	*Accuracy*	*Intensity*	*Valence*	*Arousal*	*Difficulty*
Intended emotion $H_0(1,i)$	F(4, 268) =	40.158	26.159	22.484	146.168	19.141
	p-value =	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**
	η^2 =	0.775	0.639	0.586	0.874	0.477
Generation source $H_0(2,i)$	F(2, 67) =	0.627	1.392	4.424	0.319	1.016
	p-value =	0.538	0.256	**0.015**	0.728	0.368
	η^2 =	0.018	0.040	0.117	0.009	0.029
Intended emotion × generation source $H_0(3,i)$	F(8, 268) =	1.146	1.087	1.749	0.663	1.554
	p-value =	0.332	0.372	0.099	0.684	0.139
	η^2 =	0.054	0.071	0.098	0.044	0.084

Table 5.5: F-statistics[4], p-values and effect size (η^2) results from two-way repeated measures ANOVA's for main effect of intended emotion and representation. Greenhouse-Geisser correction was used when sphericity assumption was violated. P-values indicating a significant difference at level of $\alpha = 0.05$ are highlighted.

The effects of generation source are of small size ($0.01 \leq \eta^2 < 0.06$) for accuracy, intensity, arousal and difficulty ratings, and of medium size ($0.06 \leq \eta^2 < 0.14$) for valence ratings. This suggests although the movement generation source is responsible for a very limited portion of the variability observed in the participants' recognition of the intended emotions and in their perception of intensity, difficulty and activation movement qualities, some differences observed on valence ratings might still be due to the generation source factor. Follow-up post-hoc paired Tukey-HSD tests with Bonferroni corrections on valence ratings over intended emotion and generation source reported that the most significant effect of generation source is due to the differences in the average valence ratings of the neutral state ($p < 0.01$). From Figure 5.12c, we observe indeed that the average valence rating for the pair (*neutral*, *re-sampled trajectories*) is significantly smaller than the value reported by the pair (*neutral*, *MoCap*). However, since there is no difference between the recognition rates obtained for the *neutral* state for both *MoCap* and *re-sampled trajectories* sources – both sources registered an average recognition rate of 13% for this emotional state as shown in Table 5.6 –, it seems that the differences on valence ratings had little impact on the perception and recognition of emotionally expressive bodily motions belonging to this intended emotional state.

[4]Knowing that the answers of 70 subjects were analyzed, the degrees of freedom of these F-statistics are: *a.)* generation source is a between-subjects factor with 3 levels (i.e., MoCap, reconstructed motions, motions from re-samples trajectories): $df_{source} = 3 - 1 = 2$ and $df_{error1} = 70 - 3 = 67$, *b.)* intended emotion is a within-subjects factor with 5 levels (i.e., *neutral*, *happiness*, etc.): $df_{emotion} = 5 - 1 = 4$ and $df_{error2} = df_{emotion} \times df_{error1} = 268$, and *c.)* the interaction between these two factors: $df_{interaction} = df_{source} \times df_{emotions} = 8$ and $df_{error2} = 268$.



(a) Accuracy ratings.

(b) Intensity ratings.

(c) Valence ratings.

(d) Arousal ratings.

(e) Difficulty ratings.

Figure 5.12: Average participants' ratings (mean ± standard error) for the three movement generation sources: *MoCap*, *IK reconstruction*, and *re-sampled end-effector trajectories*. Each bar shows the average rating obtained across all examples conveying the same intended emotion. Significant pair-wise differences between the levels of movement generation source are indicated by "∗" and follow the convention introduced in Table 3.9.

Since we retained four of the five null hypothesis on the effects of generation source and found that the other paired differences between valence ratings obtained for the same intended emotion across the different generation sources reported p-values greater than $\alpha = 0.05$ (see *happiness, sadness, stress* and *relaxedness* ratings on Figure 5.12c), we conclude that the movements generated with the proposed synthesis tasks are perceived very similarly to the original motions executed by the human actor. Consequently, it seems that the small differences in the recognition of the intended emotions and perception of intensity, arousal, difficulty, and valence of the original motions versus the synthesized movements are likely due to chance.

Finally, regarding our third set of null hypotheses $H_0(3, i)$ on the interaction of movement generation source and intended emotion, we found no significant effect ($p > 0.05$) on participants' accuracy, intensity, valence, arousal, and difficulty ratings. Hence, we retain the five null hypotheses $H_0(3, i)$ for $i \in \{$*accuracy, intensity, valence, arousal, difficulty*$\}$, which state that intended emotion and movement generation source factors are independent and that when they are combined they have no effect on the mean participants' ratings.

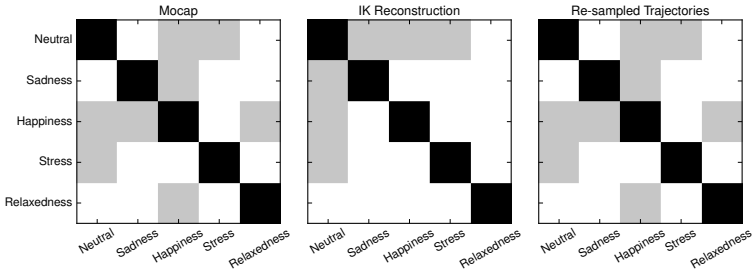
5.6.3 A Closer Look at the Effect of Intended Emotion

The significant effects detected by the ANOVA tests indicate that intended emotion has the strongest impact on participants' perception of expressive movements for all dependent variables, i.e., ratings. Although this result demonstrates that the proposed synthesis tasks are capable of generating movements that seem to be as expressive as the motions executed by a human performer, we still wished to analyze the overall behavior of participants' ratings for each one of the generation sources of interest and intended emotions. By doing so, we seek to determine if there are emotions or movement qualities on which the proposed synthesis approach has a stronger effect.

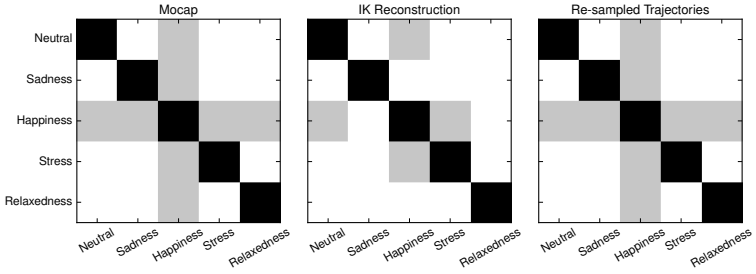
Figure 5.13 summarizes the results obtained from the follow-up post-hoc analysis with respect to the intended emotions for each movement generation source. Significant pairwise differences ($p < 0.05$) between the participants' ratings for each intended emotion are indicated by grey boxes. Similarly, the white boxes indicate no significant differences and the black boxes represent the matrices diagonal.

Regarding the accuracy rates obtained for each intended emotion, we found that only four of the five target emotional states were recognized above chance level (20%). Independently of the movement source from which the visual stimuli were obtained, the *neutral* state was not perceived as intended, that is as conveying no particular expressive content. From Table 5.6, we observe that the *neutral* state was often labeled as either *happiness*, *relaxedness*, or *stress*. Similarly, for all movement generation sources, *happiness* and *stress* registered the highest recognition rates, followed by *sadness* and *relaxedness*. This indicates that movements with higher activation were better recognized than those with lower activation levels, i.e., *sadness* and *relaxedness*. A careful analysis of the confusion matrices registered by all movement generation sources shows that participants often misclassified stimuli within the same activation level but with opposite valence. For instance *stress* and *happiness* were frequently mixed up and *sadness* was often labeled as *relaxedness*.

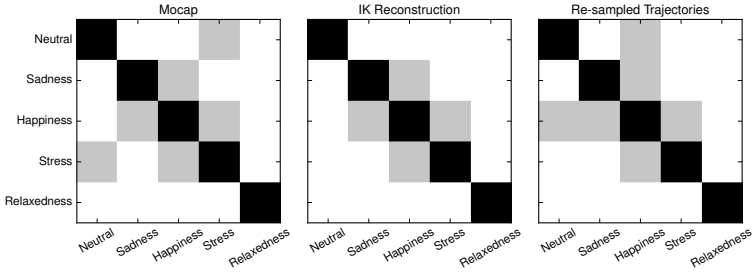
The pairwise differences on intensity ratings with respect to intended emotion were found to be only significant ($p < 0.05$) for *happiness* for both *MoCap* and *re-sampled trajectories* sources, and for the (*happiness, neutral*) and (*happiness, stress*) pairs for the *IK reconstructed*



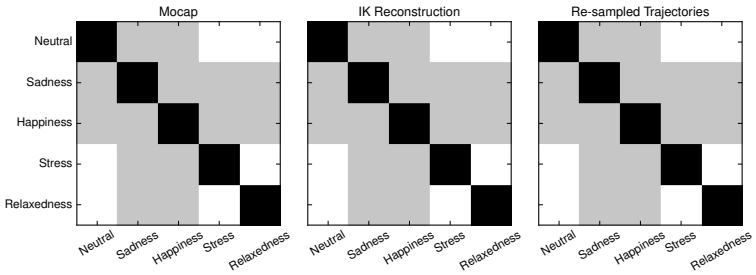
(a) Accuracy ratings for all movement sources.



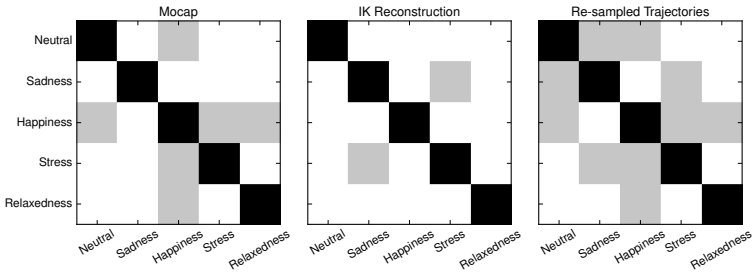
(b) Intensity ratings for all movement sources.



(c) Valence ratings for all movement sources.



(d) Arousal ratings for all movement sources.



(e) Difficult ratings for all movement sources.

Figure 5.13: Significant pair-wise differences between emotions' accuracy, intensity, valence, arousal, and difficulty ratings for each movement generation source. The grey boxes indicate significant differences between emotions' ratings at $p < 0.05$ and the white boxes indicate that there is no significant difference, i.e., $p > 0.05$.

0.13	0.05	0.38	0.15	0.29
0.15	**0.32**	0.03	0.09	0.41
0.02	0.02	**0.69**	0.18	0.09
0.11	0.04	0.18	**0.56**	0.11
0.11	0.07	0.30	0.18	**0.34**
average rate: 0.41				

(a) MoCap source.

0.11	0.04	0.39	0.26	0.20
0.14	**0.40**	0.08	0.05	0.33
0.03	0.02	**0.53**	0.33	0.09
0.11	0.10	0.22	**0.43**	0.14
0.10	0.13	0.23	0.18	**0.36**
average rate: 0.37				

(b) IK reconstruction source.

0.13	0.11	0.16	0.36	0.24
0.14	**0.32**	0.03	0.09	0.42
0.03	0.05	**0.65**	0.26	0.01
0.08	0.04	0.25	**0.50**	0.13
0.07	0.14	0.34	0.17	**0.28**
average rate: 0.38				

(c) Re-sampled end-effector trajectories source.

Table 5.6: Confusion matrices from user study on the effect of movement generation source on emotion perception. Emotions are listed in the order *neutral*, *sadness*, *happiness*, *stress*, and *relaxedness*.

motions. A closer look at the average intensity ratings obtained for each intended emotion (see Figure 5.13b) shows that the highest intensity rates were indeed assigned to the stimuli belonging to the most easily recognized emotional state (*happiness*) independently of the generation source. This results is in agreement with the patterns we observed in our previous perceptual studies (see Section 3.8.2).

When analyzing the significant pairwise differences of valence ratings with respect to intended emotion, we found two patterns common to all movement generation sources: *i.)* valence differences were easier to identify among stimuli with high activation as indicated by the p-values ($p < 0.05$) obtained from the paired-t-test between *happiness* and *stress*, and *ii.)* the significant difference between *happiness* and *sadness* valence rates we observed in our previous perceptual studies was still present on the movements issued from the three generation sources. We observed also an interesting relationship between the valence ratings of motions issued from the *MoCap* and *re-sampled trajectories* sources and the misclassification of stimuli belonging to the *neutral* emotional state. A closer look to both Figure 5.13c (left matrix) and Table 5.6a, shows that a significant pairwise difference between the valence ratings obtained for the *neutral* and *stress* emotional states for the *MoCap* source resulted in a higher misclassification of neutral stimuli as either *happiness* or *relaxedness*. Conversely, the significant difference between the *neutral* and *happiness* valence ratings for the *re-sampled trajectories* source (see right matrix on Figure 5.13c) produced the opposite misclassification pattern, i.e., neutral stimuli were instead labeled as either *stress* or *relaxedness*. This claim is further supported by the confusion matrix associated to *IK reconstructed* motions. Since no significant difference was found with respect to the *neutral* target emotion, stimuli belonging to this category were equally misclassified as *happiness, stress, or relaxedness*. This indicates that the combination of both expressive end-effector trajectories and IK whole-body reconstruction preserve partially the cues participants employ when rating the different valence

levels (i.e., low and high). Consequently the small differences observed on the perception of the intended emotional states for the IK reconstructed and re-sampled trajectories generation sources with respect to MoCap generation source might be explained by a loss of information in this motion quality.

Post-hoc analysis of arousal ratings with respect to the intended emotion for each generation source showed the same set of statistically significant differences for all sources. That is, *happiness* and *sadness* activation ratings were consistently found higher and lower with respect to the other intended emotions (see Figures 5.12d and 5.13d). This indicates that both synthesis tasks provided the same kinematic information observed in the motions obtained from the *MoCap* generation source.

Finally, when analyzing any possible significant difference of difficulty ratings with respect to intended emotion we found that: *i.)* for both *MoCap* and *re-sampled trajectories* sources, *happiness* was consistently labeled as the easiest emotional state to analyze and rate, *ii.)* for motions issued from *IK reconstruction* source only the difficulty ratings between *sadness* and *stress* were significantly different, and *iii.)* stimuli generated from *re-sampled trajectories* that convey *sadness* were rated as less difficult than stimuli expressing either *stress* or *neutral* state. However, although *sadness* was often rated as less difficult to perceive than *stress*, Table 5.6 shows that the latter was better recognized than the former across all movement generation sources.

5.7 Summary and General Discussion

A novel motion synthesis approach for the generation of emotionally expressive body motions was proposed in this chapter. This approach focuses in the use of expressive end-effector trajectories and comprises two main elements: *i.)* an inverse kinematics implementation that maps end-effector trajectories to whole body motions, and *ii.)* a re-sampling scheme that generates novel expressive end-effector trajectories while preserving all the kinematic patterns associated with the expression of affect and emotions.

After surveying the work done by the affective computing and computer animation domains on the generation of expressive and/or stylistic body motions, two main approaches were identified: rule-based and example-based methods. We then argued that the motion synthesis method we propose benefits from the strengths and advantages of these two approaches, while addressing their respective shortcomings. In the one hand, by using end-effector trajectories issued from a labeled and validated MoCap database as control signals, we simultaneously preserve most of the visual appeal intrinsic to the exampled-based methods and the much better and intuitive control offered by the rule-based approach. On the other hand, our use of a purely procedural method, i.e., inverse kinematics, to map end-effector trajectories to whole body motions provide us with the flexibility inherent to the rule-based approach while decreasing the dependency of the example-based approach on the MoCap database.

The chapter then presented a detailed description and discussion of the two fundamental ideas behind our IK implementation: *i.)* the division of the character's articulated body into five independently controlled joint groups, herein called articulated chains and *ii.)* the combination of joint limits and multiple tasks for the arm chains as a means of constraining the space of possible solutions to more natural and humanly plausible postures. We then

proceeded to review the challenges behind the use of re-sampling schemes on temporally dependent data such as motion data and the reasons behind the choice of LGB as our re-sampling schema.

The performance of the proposed synthesis approach was tested using two complementary tasks. Whereas the first task, herein called *motion reconstruction*, was useful to assess the capacity of our IK-based mapping and how well emotion-related cues propagated through the character's body, the second task, referred to as *motions from sampled trajectories*, allowed us to evaluate the re-sampling scheme and the suitability of end-effector trajectories as a motion model for the generation of expressive body motions. Automatic affect classification and an information theory divergence measure were used to quantitatively compare the generated motions against the ground truth data. Similarly, a user study was conducted to determine whether the source from which expressive body motions were obtained, i.e., MoCap data, IK reconstruction, or re-sampled trajectories, had an effect on the perception of the intended emotional states.

For the quantitative evaluations we found that: *i.)* the accuracy rates obtained for the synthesized motions are very close to the rates observed for the ground truth data. In particular we observed that the synthesized motions depicting *happiness*, *sadness* and the *neutral* state were the best recognized. *ii.)* The divergence measure between the ground truth data and the synthesized motions are close to zero, which indicates that the features used to quantify the underlying affective content present the same statistics (distributions) for the three movement generation sources. We found anew that depictions of *happiness*, *sadness* and the *neutral* state reported the best divergence scores. *iii.)* The kinematic features (*velocity*, *acceleration*, *jerk*, and *curvature*) we used to characterized expressive content seem to be robust and, to some extent, independent from the action (i.e., semantic content) being analyzed.

For the qualitative evaluation we found that movement generation source factor had no significant effect in 4 of the 5 ratings we analyzed (i.e., accuracy, intensity, arousal, and difficulty). Hence, synthesize motions were perceived as expressive as the ground truth data. In the case of valence ratings, we found that the pleasantness rates attributed to stimuli conveying the *neutral* state were lower for the motions generated from re-sampled trajectories, than for the other generations sources. For the other emotional states, no significant effect of generation source was found. An interesting finding of this perceptual study is the fact that whole body motions with no apparent semantic meaning (i.e., motions generated from randomly re-sampled trajectories) were still perceived as being emotionally expressive. This indicates, that independently of the motion class being studied, end-effector trajectories are sufficient to describe affective content. Hence, they can be used as control signal for the generation of expressive whole-body motions.

Chapter 6

Conclusions

Contents

In this thesis, a low-dimensional motion parameterization consisting of end-effectors (i.e., head, hands, and feet) and pelvis, also referred to as root joint, trajectories was proposed. Based on results from three different research areas: *perception of emotion and biological motion*, *automatic recognition of affect and emotions*, and *computer character animation*, we argued that the proposed representation was suitable and sufficient for the analysis and generation of expressive bodily motions. A computational and perceptual analysis of expressive MoCap movements was carried out in order to assess the validity and convenience of our main hypothesis and arguments. The principle behind our analysis and the work presented in this thesis was the following:

"If the proposed low-dimensional representation is indeed suitable for the study and generation of expressive bodily motions, the computational and perceptual evaluation of such representation should provide results close (or as good as) the ones obtained when information of the entire body is available and used."

In the following, we summarize our main contributions as well as some interesting directions for the continuation of the work presented in this thesis.

6.1 Contributions

There are three main contributions arising from this work. Firstly, a new motion capture database was specifically designed for the purpose of this thesis. This database contains ex-

amples from different motion classes (i.e., periodic movements, functional behaviors, spontaneous motions, and theater-inspired motion sequences) performed by several actors. This diversity was notably useful to determine the determine the generalization capabilities of the proposed low-dimensional motion representation. Secondly, a user study and automatic classification framework were designed and used in order to perceptually and quantitatively assess the amount of emotion-related information still conveyed and encoded in the proposed representation. We observed that although slightly differences in performance were found with respect to the cases in which the entire body was used, our proposed representation preserves most of the motion cues salient to the expression of affect and emotions. Lastly and most importantly, we have proposed a simple motion synthesis system able to: *i.)* reconstruct whole-body motions from the proposed low-dimensional representation and *ii.)* produce novel end-effector and pelvis expressive trajectories. A quantitative and qualitative evaluation of the generated whole body motions shows that these motions are as emotionally expressive as the movements recorded from human actors. These main contributions are discussed in more detail below.

6.1.1 A New Motion Capture Database Designed Using Principles from Physical Theater Theory *(Chapter 3)*

In the context of this thesis, we required a database suitable for both the analysis and synthesis of emotionally expressive bodily motions. That is, a database that comprises among other technical and theoretical requirements: several subjects, different types of movements in which all body limbs are employed, several emotional states and various repetitions for each possible combination of motor behavior and emotional state.

After reviewing the existing and publicly available databases, we determined that none of them was fully appropriate for our work. Mainly because they were designed with a different purpose in mind. However, after studying their design principles, we found two common elements that latter served us in the definition and construction of the MoCap database we proposed: *i.)* the use of mood induction procedures and skilled actors [14] and *ii.)* the interest on theater-inspired scenarios and theater actor training theory [38, 174].

Inspired by these two elements, we designed and recorded a MoCap expressive database that borrowed some theoretical precepts and ideas from a particular way of doing theater known as *physical theater*. Specifically, we devised a mime-magician scenario in which skilled actors were asked to only use their bodies to channel meaning and express emotions. In this scenario, each actor embodied a magician during performance. Three magician tricks: *the disappearing box*, *pulling a rabbit from a hat*, and *taking scarves from an empty jacket*, were to be performed under one of the following emotional states: *happiness*, *sadness*, *stress*, *relaxedness*, and *neutral*. A combined mood induction procedure (story-based and imagination-based MIP) was used in order to facilitate the enacting of the selected emotional states.

The mime-magician scenario was further extended with two more sequences: a locomotion example and an freely chosen improvisation sketch. By doing so, we enlarged the number of movement classes available in our database and provided a means to assess whether theater inspired scenarios are better recognized and hence more suitable for the analysis and generation of expressive bodily motions.

The proposed database includes in total: 7 actors (3 women and 4 men), 5 emotional states, 3 repetitions for each magician trick by emotional state and actor, 1 repetition of the

locomotion example and improvisation sketch by emotional state (only 5 actors performed these sequences), and 5 repetitions of a selected number of actions within the magician scenario for each emotional state (these recordings are available for 2 actors only).

Two user studies were carried out to evaluate the human perception of the proposed database. We found that the perception of the recorded expressive motions was significantly influenced by the emotional states we asked the actors to convey via their body movements and that all emotional states were recognized largely above chance level.

6.1.2 A Qualitative and Quantitative Evaluation of the Proposed Low-Dimensional Motion Parameterization *(Chapters 3 and 4)*

By definition low-dimensional representations hidden in high-dimensional data are often selected so as to highlight and preserve most of the relevant information embedded in the original data [26]. This implies that by measuring how much information is lost with respect to the initial high-dimensional representation, we can obtain an estimate of the quality of the selected low-dimensional representation.

Our main interest was to define a low-dimensional representation that preserves most of the emotion-related content encoded in the initial bodily motion examples and thus could be used for generating new expressive bodily motions. In order to assess if the proposed low-dimensional representation met this criteria, we designed two evaluation procedures to determine how much expressive information is preserved by the end-effector and pelvis trajectories in comparison with the entire body. Data from our database was used for both evaluations.

The first evaluation corresponded to a user study in which observers were asked to rate and recognize the emotional content conveyed by two distinct visual displays of the same body movements. One display showed the entire body using a stick-figure representation, whereas the other one depicted the 3-dimensional trajectories of hands, head, feet and pelvis. Both displays were evaluated separately. This user study provided us with: *i.)* a qualitative estimate of the quality and suitability of the selected low-dimensional representation and *ii.)* a base line for the automatic affect classifier used in our quantitative evaluation. In this study, motion examples of 5 actors and all movement classes were used.

The second evaluation was of a quantitative nature. An automatic affect classifier (Random Forest) was tested using two different groups of features. The first group consisted of features computed using the proposed low-dimensional representation, whereas the second group comprised features from the entire body and feature subsets automatically determined via feature selection techniques. In order to carry out this evaluation procedure, we proposed have a systematic approach for transforming variable-length and temporal dependent motion data to feature-based and fixed-length vectorial representations.

Overall the results of the perceptual study indicated that although observers' accuracy was impaired when presented with end-effectors and pelvis trajectories only, they still recognized 4 of the 5 emotional states above-chance (20 % recognition rate). Nonetheless, as shown by contribution 6.1.3 (Chapter 5), this impairment was not longer present when whole-body motions were generated from the observed end-effectors and pelvis trajectories via the proposed inverse mapping.

From the quantitative evaluation we found that the chosen class of classifier (i.e., Ran-

dom Forest) exhibited the same behavior independently of the subset of features used to summarize the motion data. In other words, features subsets computed from the proposed low-dimensional representation seem to provide the same amount of information about the expressive content of different motion classes for different subjects than features computed either from the entire body or automatically determined via feature selection techniques.

The results obtained from both evaluation studies allowed us to conclude that, in the context of the work done in this thesis, the selected low-dimensional representation provides sufficient emotion-related information for the perception and automatic recognition of emotion states from bodily motions. Furthermore, as reported by Krüger and colleagues [145], we observed that richer and more complex representations give little or no advantage over the use of end-effector and pelvis trajectories.

6.1.3 A Validated Motion Synthesis Approach for the Generation of Novel Bodily Expressive Motions *(Chapter 5)*

Contribution 6.1.2 showed own end-effectors and pelvis trajectories encode most of the motion cues salient to the expression and recognition of emotions and affect. However, since our primary objective was the generation of expressive bodily motions, a motion synthesis approach was proposed in order to assess how expressive are the bodily motions generated either from observed or synthesized low-dimensional trajectories.

The motion synthesis approach we have proposed consists of two main components:

- A set of independent inverse kinematics controllers that map the low-dimensional representation to high-dimensional full body motions. A controller is associated to each limb, also called articulated chain, in the character's body. Joint limits and elbow trajectories are used during reconstruction in order to constrain the generated motions to the space of human plausible body postures.
- A re-sampling scheme that generates new random end-effector trajectories while preserving the underlying emotional content. The synthesized trajectories are both spatio-temporally and semantically different than those observed in the training MoCap database. Hence, they can be used to further evaluate the suitability of the proposed low-dimensional representation for the generation of bodily motions belonging to different movement classes.

The expressiveness of motion generated via the proposed synthesis approach was verified both qualitatively and quantitatively. A user study was used as qualitative validation. Observers were asked to rate the emotional content conveyed by motions issued from three different sources: MoCap database, motions reconstructed from observed end-effectors and pelvis trajectories and motions generated from synthesized trajectories. We found that the movement generation source had no statistically significant effect on the perception of emotion.

The same classification schema used in contribution 6.1.2 and a similarity measure based on information-theory literature were used as quantitative evaluation. First, we compared the recognition rates of the RF classifier when tested on motions generated from the three sources we listed above. Second, we measured the statistical similarity between: a.) MoCap

database and motions reconstructed from observed end-effector trajectories and b.) MoCap database and motions generated using re-sampled trajectories. We observed no significant difference in the classifiers performance as well as a consequential similarity between the synthesized and observed (MoCap) bodily motions.

Both qualitative and quantitative results led us to conclude that end-effectors are pelvis trajectories are certainly an interesting low-dimensional representation of expressive bodily motions. Furthermore, it can be used for both the study and the generation of such motions.

6.1.4 Concluding Remarks

In a recent survey, Gunes and colleagues [105] highlighted the fact that there is an unlimited vocabulary of body postures and motions with combinations of movements of various body parts that can be employed to communicate and express emotions. In order to categorize, analyze and understand what are the bodily motion cues permeating all different movement depictions of the same emotional state will require to collect and study a considerable amount of motion data.

Through our work, we have proven that is possible to focus on low-dimensional, yet meaningful, motion representation that will facilitate such task. This representation will not only reduce the amount of data required to characterized the space of expressive human poses and motions, but it will also simplify the acquisition of such data. Since we are interested in a reduced set of 3-dimensional trajectories, low-cost and less invasive technologies such as the Kinect [175] sensor can be used. A further advantage of the proposed representation is the definition of a compact space suitable for the recognition and generation of novel expressive bodily motions. Furthermore, our low-dimensional representation shows that the number of critical informative body joints for emotion recognition is quite low and that better engineered features can be computed only from these joints (i.e., end-effectors and pelvis joint).

6.2 Perspectives and Future Work

In this section, we propose several directions for the continuation of the work presented in this thesis. The improvements and potential research topics herein listed concern the three research domains in which most of our work is based: *emotion perception*, *automatic affect recognition*, and *character animation*.

6.2.1 Application to Other Motion Classes and Databases

One of the main barriers to the development of reliable models for the analysis and generation of expressive bodily motions is that most of the existing approaches are thought, built and tested using a particular dataset of motion data. Furthermore, most of the times these dataset contains examples of one single motor behavior such as human gait and a limited number of emotional states. Thus, it is hard to determine how well these models generalize to other motion classes, emotions, and applications.

Since the proposed low-dimensional representation consists of 3-dimensional trajectories frequently found in the existing motion datasets, it is quite straightforward to test its

suitability on larger sets of motor behaviors and emotions. Both the proposed automatic classification scheme and motion synthesis approach can be easily applied on datasets such as Emilya [87] or the dance examples used by [4].

6.2.2 A Further Evaluation of the Expressive Quality and Believability of the Synthesized Motions

The perceptual studies conducted in the context of this thesis aimed to determine whether the different motion generation sources had any significant effect on the perception of the selected 5 emotional states: *neutral*, *happiness*, *sadness*, *stress* and *relaxedness*. In order words, we wished to evaluate how expressive were the bodily motions generated using end-effector trajectories. Although our results show that no significant difference was found, we believe that a further evaluation is necessary in order to determine the believability of the generated motions and thus of the animated character. Such evaluation must be conducted on both the original and synthesized motions, since it is possible that the proposed motion parameterization and synthesis approach might induce a change on the perception of such quality.

An additional issue raised by the results obtained in these perceptual studies is the nature of the motion data to be used when synthesizing expressive bodily motions. Although we found that the emotional states conveyed by the both original and synthesized motions were distinguished above chance level, the recognition rates remained low in comparison to studies in which much more exaggerated motions were used [7, 187]. This might suggest that whereas expressive bodily motions obtained under more naturalistic settings are desirable for application such as automatic affect recognition [136], more elaborated depictions might be needed when animating believable and easily understood virtual characters. A perceptual study in which naturalistic bodily motions are compared to more exaggerated movements might provide some insight into which kind of stimuli must be used when animating believable virtual characters.

6.2.3 Improved Mapping from the Proposed Low-Dimensional Representation to High-Dimensional Space

The function mapping from end-effector trajectories to whole-body motions is one of the key elements of the motion synthesis approach proposed in this thesis. This mapping must: *i.)* preserve and correctly propagate the expressive content encoded in the end-effector trajectories used as control signals, and *ii.)* ensure that the resulting bodily motions lie within the space of humanly plausible and biological movements.

Although the inverse kinematics controllers we used to approximate this mapping provides an efficient and flexible control over the resulting motions, additional constraints (joint limits and elbow trajectories) were needed in order to enhance the solutions provided by this mapping. Nonetheless, we observed that the generated motions might still suffer from visual artifacts inherent to both the redundancy of the articulated chains being controlled and the use of purely procedural synthesis techniques.

We believe that deep learning techniques represent an interesting direction for the improvement of this mapping function. Specifically, convolutional networks have been shown to successfully encode within the hidden units the bio-mechanical constraints governing human motion [118]. By combining this type of network and a generation scheme as the one

recently proposed by the deep generative network WaveNet [66], we believe it is possible to generate biologically correct and visually appealing whole-body motions that closely follow the control signal and exhibit the intended emotional content.

6.2.4 Controlled Generation of Expressive End-Effector Trajectories

In Chapter 5, we showed that whole-body motions generated from expressive end-effectors trajectories were perceptually and quantitatively rated as expressive as MoCap data. Furthermore, we used a re-sampling scheme in order to generate novel expressive end-effectors trajectories. However, the user had no control over the motion path described by these trajectories.

Real-world applications for character animation must offer some control over the possible motion outcomes [250] to the user. Fortunately, end-effector trajectories represent an intuitive parameterization of the motions to be generated. Based on this, it is possible to extend the synthesis approach we proposed by combining two existing computer animation techniques: style transfer and performance animation.

First, the user is asked to specify the end-effector trajectories that best characterize the desired whole-body motion via low-cost sensors as done in previous performance animation methods [51, 132, 230]. These trajectories do not contain emotion-related content. Second, the user determines the emotional state to be conveyed by the resulting expressive bodily motions. Third, using an adapted implementation of the existing style transfer approaches [119, 266], the input (non expressive) end-effector trajectories are modified according to the spatio-temporal patterns associated with the desired emotional state. Finally, the resulting expressive bodily motion is obtained by applying our inverse kinematic mapping controllers to the transformed end-effector trajectories.

6.2.5 Perceptually Guided Generation of End-Effectors Trajectories

As mentioned earlier, there are two main approaches commonly employed to determine and replicate the motion features salient to the expression of emotions via bodily motions: computational data-driven and perceptual guided studies. Whereas the former are often complex, computational expensive and highly dependent on the data used during the training/configuration stage, the latter are based on subjective and empirical observations not easily mapped and applied to new data [78]. However, as shown throughout this thesis and in the recent work done in [77], it is possible to combine the main principles behind these two approaches and produce simple, intuitive, yet powerful models.

Based on this idea, the generation of new expressive end-effectors could be enhanced through an iterative application in which expert users modify the spatio-temporal characteristics (e.g., form, timing, symmetry, etc.) of initially non expressive trajectories so as to produce a whole body motion that conveys certain emotional state. Through this application and a assessment based on a user study, it is possible to determine the perceptual significance and impact of different transformations as well as a sufficient number of motion features that can be generalized to across different motion classes.

Appendix A

Scenarios for Story-Based Mood Induction Procedure

During the second set of recordings, five short stories related to the magician scenario were created. Each story contextualized and described each one of the following intended emotional states: *neutral*, *happiness*, *sadness*, *stress* and *relaxedness*.

Neutral State

After a pleasant night of sleep and a good breakfast, you are ready to start your daily training routine. You start taking all you need (your hat, jacket, wand, scarfs...). It is a sunny day and you are thinking about what have to do today. You are ready to practice the first magic trick of your show.

Happiness

It is your last performance of the day and you feel great. You have been in already two shows and everything went as you planned. Both audiences were so overly enthusiastic and joyed with your performance, that the theater's owner proposed to hire you for the main spectacle of the next session. Moreover, you received the phone call you were expecting for the last three weeks: you have been nominated to the best magician award of the year. You hear your name been called, you are excited and more than ready for your show.

Stress

You are taking part in the European Annual Magician Competition of this year. You are competing for your place at the semifinal. You have seen the performance of your concurrent and it seems he has been given a very good score. Plus, he seemed to master the magic trick

you have been struggling with for the last couple of months and specially prepared for the competition. You hear your name been called, it's your turn to perform in front of the jury and a crowded audience. You feel the pressure on you as this is your last chance to make it to the semifinal.

Sadness

You are getting ready for your show, the one that you used to perform with your partner on stage. But this one (your best friend) died recently in an accident. You can't stop thinking about him, how much you miss him and how hard will be to perform without him.

Relaxedness

You are in the last part of you show. The audience is enthusiastic with your performance. Your friends are waiting for you backstage, you are all going to have dinner in your favorite restaurant. You prepare yourself for the last trick, the first you learned and the one you master the most.

Appendix B

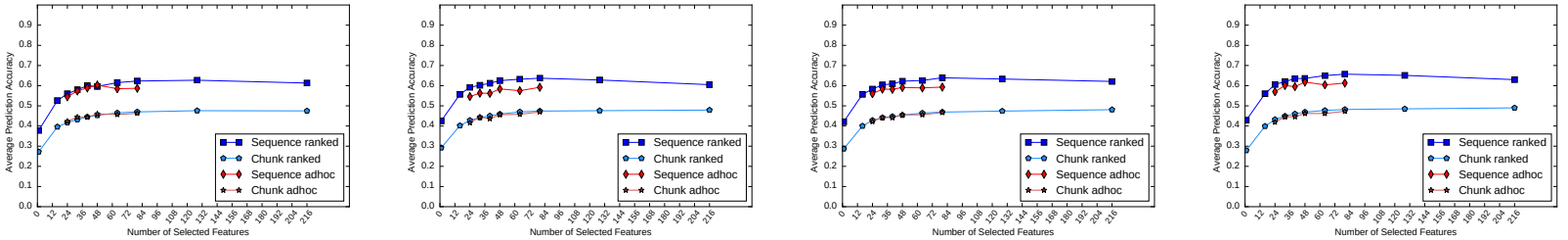
Effects of Sliding Window Parameters on Classification Results

One of the main elements in our classification pipeline is the manner in which *motion chunks* and, consequently, feature vectors are defined. The former are computed using a overlapped sliding window approach and the latter are the result of averaging the kinematic features computed along this window. Thus, the window parameters, i.e., its length and the overlap percentage between two contiguous windows, might have a significant impact on the classification results obtained for each task and in the validation of our hypothesis.

In order to study the effect of these two parameters, we have followed the same procedure described in Algorithm 2 for all combinations of three possible window's lengths: $\{250, 500, 750\}$ frames, and four possible overlap percentages: $\{0.0, 0.25, 0.50, 0.75\}$. As in the previous section, we fixed the RF model hyper-parameters to the default values $ntrees = 500$ and $mtry = \sqrt{p}$. Our analysis is centered on the results obtained for between subject and within subject *one sequence out* classification tasks, since these are the tasks in which both the generalization of the RF classifier and the features computed from end-effector trajectories are truly assessed.

FiguresB.1 and B.2 show the results obtained for between subject and within subject *one sequence out* tasks respectively. At first glance, we observe that there is no significant difference in chunk accuracy for all parameter combinations and for both ***adhoc*** and ***ranked feature groups***. It is in the sequences accuracy's rates where changes due to window parameters are observed. Hence, our analysis is centered on sequence predictions only.

For both classification tasks similar tendencies are observed. In one hand, we find that, besides its impact on the resulting number of samples in the learning dataset, the percentage of overlap between contiguous *motion chunks* seems to have no significant effect on the recognition rates registered for both types of feature subsets. On the other hand, we observe that the relative difference between the recognition rates of the ***adhoc group*** and those of the ***ranked group***, although small, systematically increases with the sliding window's length.

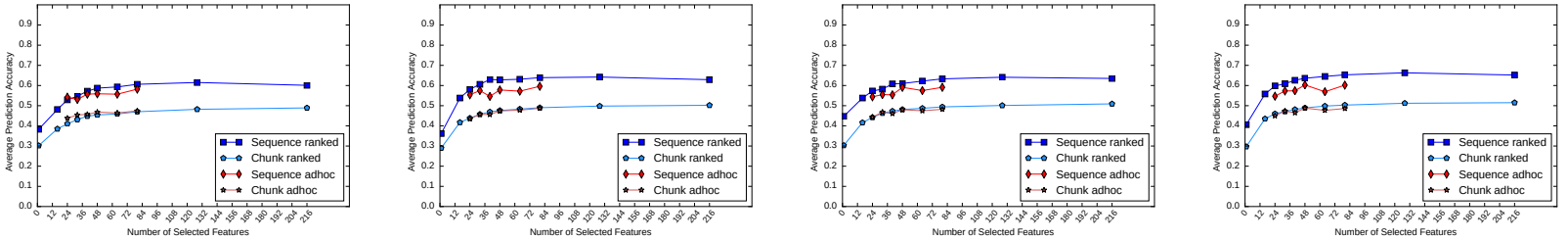


(a) *Motion chunks* of 250 frames with zero overlap

(b) *Motion chunks* of 250 frames with 25% overlap

(c) *Motion chunks* of 250 frames with 50% overlap

(d) *Motion chunks* of 250 frames with 75% overlap

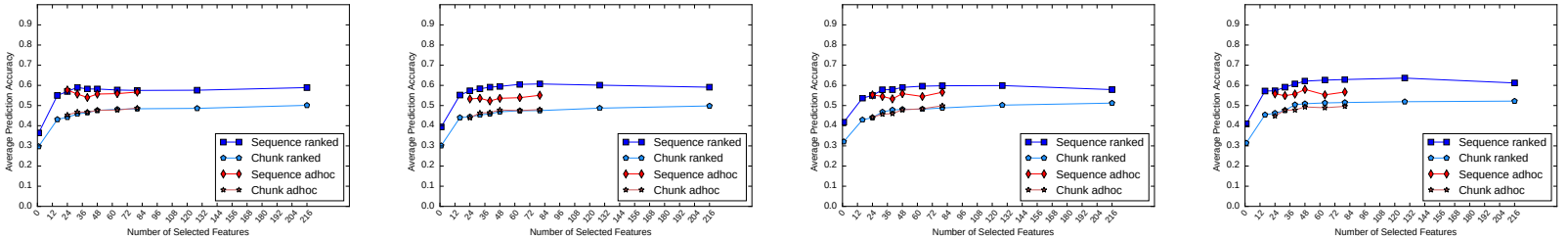


(e) *Motion chunks* of 500 frames with zero overlap

(f) *Motion chunks* of 500 frames with 25% overlap

(g) *Motion chunks* of 500 frames with 50% overlap

(h) *Motion chunks* of 500 frames with 75% overlap



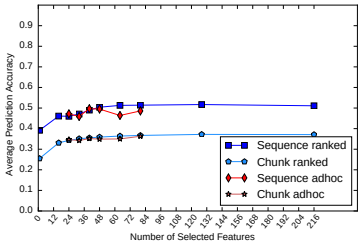
(i) *Motion chunks* of 750 frames with zero overlap

(j) *Motion chunks* of 750 frames with 25% overlap

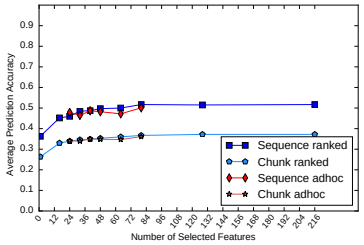
(k) *Motion chunks* of 750 frames with 50% overlap

(l) *Motion chunks* of 750 frames with 75% overlap

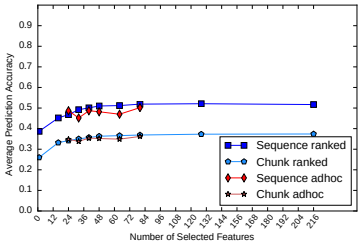
Figure B.1: Behavior of accuracy curves for both *motion chunks* (yellow and light blue) and sequences (red and dark blue) as the sliding window parameters change. Results were obtained from within subject *one sequence out* task and depict both ***adhoc*** (red and yellow curves) and ***ranked*** (light and dark blue curves) feature groups.



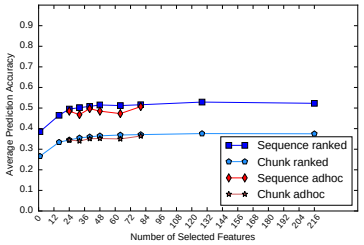
(a) *Motion chunks* of 250 frames with zero overlap



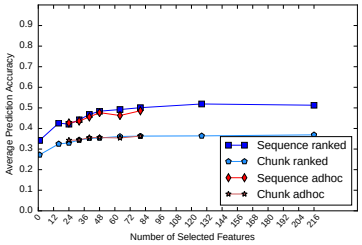
(b) *Motion chunks* of 250 frames with 25% overlap



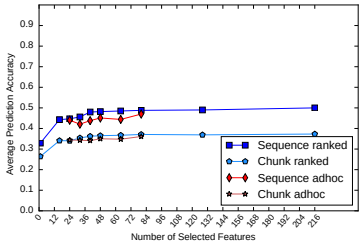
(c) *Motion chunks* of 250 frames with 50% overlap



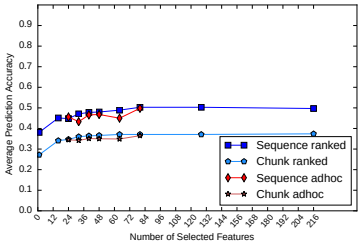
(d) *Motion chunks* of 250 frames with 75% overlap



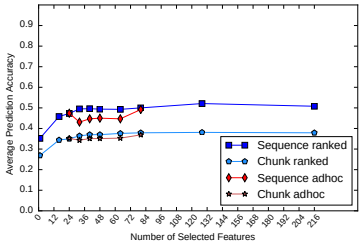
(e) *Motion chunks* of 500 frames with zero overlap



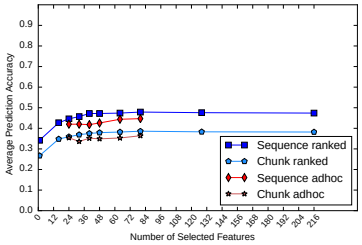
(f) *Motion chunks* of 500 frames with 25% overlap



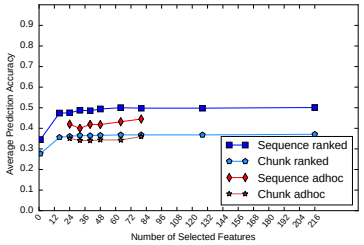
(g) *Motion chunks* of 500 frames with 50% overlap



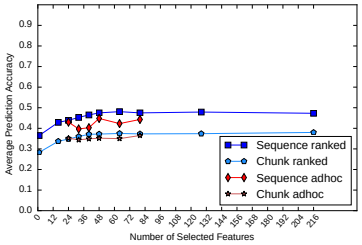
(h) *Motion chunks* of 500 frames with 75% overlap



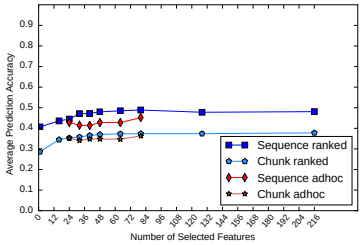
(i) *Motion chunks* of 750 frames with zero overlap



(j) *Motion chunks* of 750 frames with 25% overlap



(k) *Motion chunks* of 750 frames with 50% overlap



(l) *Motion chunks* of 750 frames with 75% overlap

Figure B.2: Behavior of accuracy curves for both *motion chunks* (yellow and light blue) and sequences (red and dark blue) as the sliding window parameters change. Results were obtained from between subject task and depict both ***adhoc*** (red and yellow curves) and ***ranked*** (light and dark blue curves) feature groups.

That is, as the number of frames to summarize into a feature vector increases, it seems that the features computed from end-effector trajectories only do not preserve as much affect-related content as we first expected. If we look carefully to the accuracy curves of the ***ranked group*** (dark blue) across all possible window lengths, we find the same tendency on the classifier's accuracy. However, it seems that the automatic feature selection procedure attenuates much better this effect than end-effector trajectories.

The reduction on accuracy as the window length increases can be explained by how features vectors are defined. Recall that after computing *velocity*, *acceleration*, *jerk* and *curvature* for all frames contained in a window (i.e., *motion chunk*), we proceed then to compute the average and standard deviation of their norms. By averaging across long durations (e.g., 750 frames $\approx$ 3.8 s), we are considerable smoothing out all those small variations related to the expression of emotion. Thus, the resulting feature vectors contain less affective information, which makes their classification much more harder. Finally, we observe that the results obtained from 250-frames and 500-frames windows, independently of the overlap percentage, are considerable close. This suggests that the optimal window length value may be contained in this interval.

Appendix C

Effect of Random Forest Hyper-Parameters on Classification Results

An automatic classification model can be seen as a generic function that has been tuned to the particularities of each problem on which it is applied. This tuning process involves two main elements: the training observations and the model's hyper-parameters. The training dataset shows to the classifier what kind of information we want to recognize. The hyper-parameters help the model to determine class boundaries that will generalize well to unseen observations.

When working with a Random Forest classifier, there are two hyper-parameters to which the performance of the model may be sensible at: the number of trees to grow in the forest, often referred to as *mtrees*, and the number of randomly selected features to analyze during the search of a node's optimal split, often referred to as *mtry*. These two parameters may also influence the variable importance measures internally computed by the RF model. Hence, before using RF for feature selection, classification, or both, it is widely common to do a non-exhaustive search of the best set of hyper-parameters. However, this additional step would have supposed to further subdivide our already limited dataset into training, validation and testing sets, or to use computationally expensive cross-validation nested procedures as those describe in [143].

In the original paper on RF written by L. Breiman [34], the author proposed to use the OOB estimate during the hyper-parameters search. Nevertheless, due to the temporal correlation between feature vectors coming from the same sequence, our oob estimates are overly optimistic and could lead us to a set of hyper-parameters that might produce a model with poor prediction at the test observations and that yields subsets of non-informative features during feature selection. We adopted a different approach instead. We have used the default model's hyper-parameters, $mtrees = 500$ and $mtry = \sqrt{p}$, for all our experiments. We have

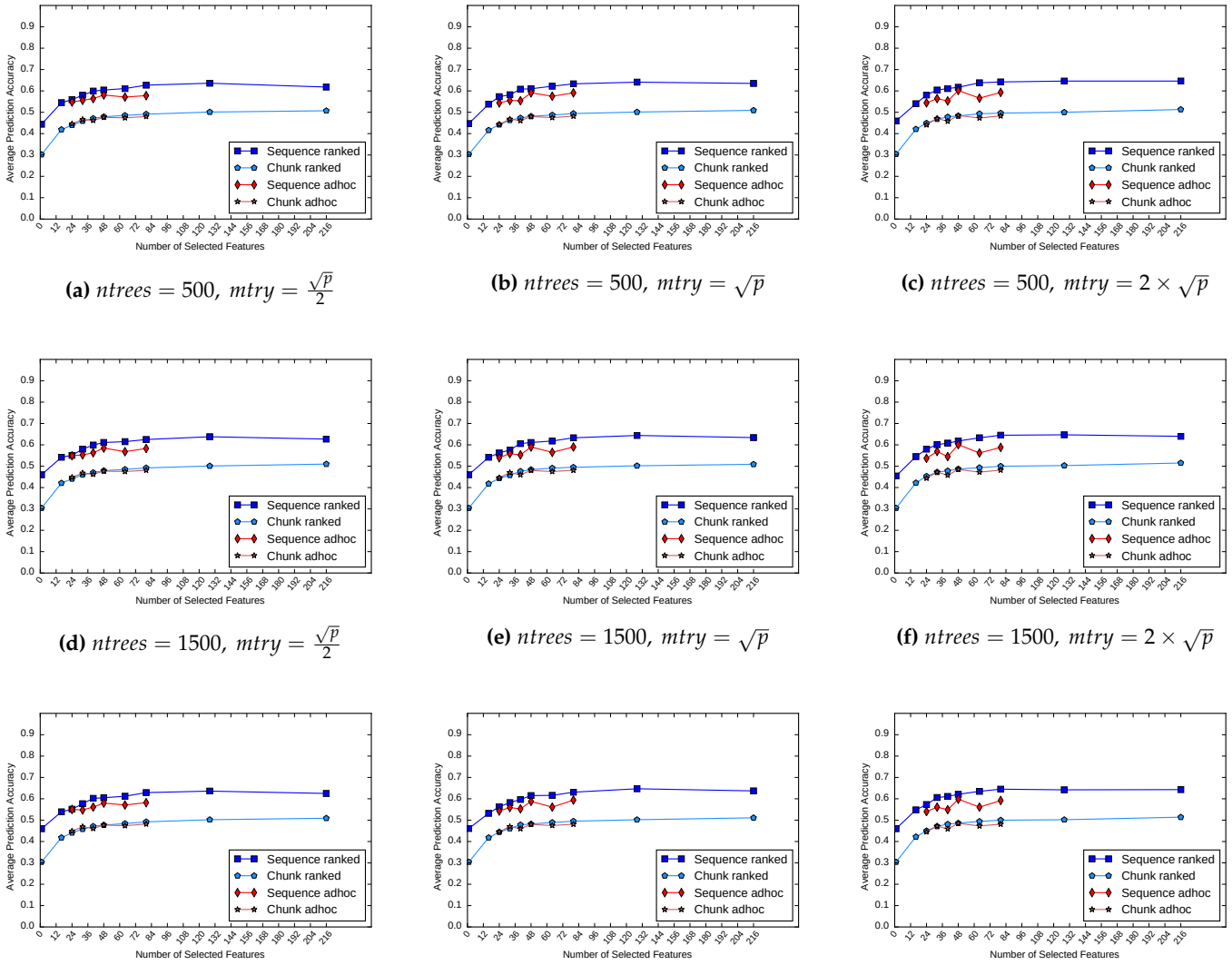
then run all our classification tasks using different combinations of hyper-parameters. We have tested three possible number of trees $\{500, 1500, 2500\}$ and the three potential choices of *mtry* suggested by Breiman [34] $\{\frac{\sqrt{p}}{2}, \sqrt{p}, \sqrt{p} \times 2\}$.

We present and analyze the results obtained for both within subject *one sequence out* (see Figure C.1) and between subject (see Figure C.2) classification tasks. RF Classifier has been trained and tested on feature vectors computed using 500-frames long *motion chunks* with 50% overlap percentage between them. It is important to notice that for each set of hyper-parameters, the same RF model configuration was used both for the definition of features subsets in ***ranked group*** and for the posterior recognition of emotional states. That is, the same set of parameters was used each time the classier was trained and tested on each one of the feature subsets of interest.

Several authors [23, 30] suggest that increasing the number of trees will lead to more accurate and reliable predictions. However, both Figures C.1 and C.2 show that the number of trees grown in the forest has negligible impact on the classifier's accuracy for both feature groups. This suggests that the performance of feature subsets based on end-effector trajectories is invariant to this hyper-parameter. Furthermore, both *motion chunks* and sequences predictions show the same tendencies and average recognition rates as the number of trees increases. Thus, it seems that our RF classification model has already reached its best behavior with the default $ntrees = 500$ value.

When working with a RF classifier it has been shown that its performance depends on the correlation between any two trees in the forest and the strength of each individual tree. These two aspects are completely determined by the choice of the *mtry* hyper-parameter. As stated by Breiman in [34]: "*reducing its value reduces both the correlation and the strength. Increasing its value increases both*". Similarly, it has been also reported that the usefulness of RF as a feature ranking method is also strongly influenced by this hyper-parameter [96]. For example, in the scenario in which many informative predictors are available, a large value can potentially ignore important by weak predictors. In the contrary, a small value might give them a higher chance to be selected. Hence, *mtry* should be carefully defined.

After a thorough inspection of both Figure C.1 and Figure C.2, we notice indeed a slight but systematic increase on the accuracy of sequence predictions for both feature groups as *mtry* changes. All the same, the overall effect is so small that we can considerate it as negligible. Even if we tune this parameter's value, it seems highly probable that the global accuracy of RF as a **classification model** will not change. This fact confirms our choice of $mtry = \sqrt{p}$ as the value to be used when assesing the difference of accuracy between any two feature subsets. Nevertheless, Figure C.1 also indicates that the relative difference between the ***ad-hoc*** and the ***ranked group*** systematically changes as *mtry* increases. This can be explained by a change in the ranking of features and consequently in the feature subsets defined in the ***ranked group***.

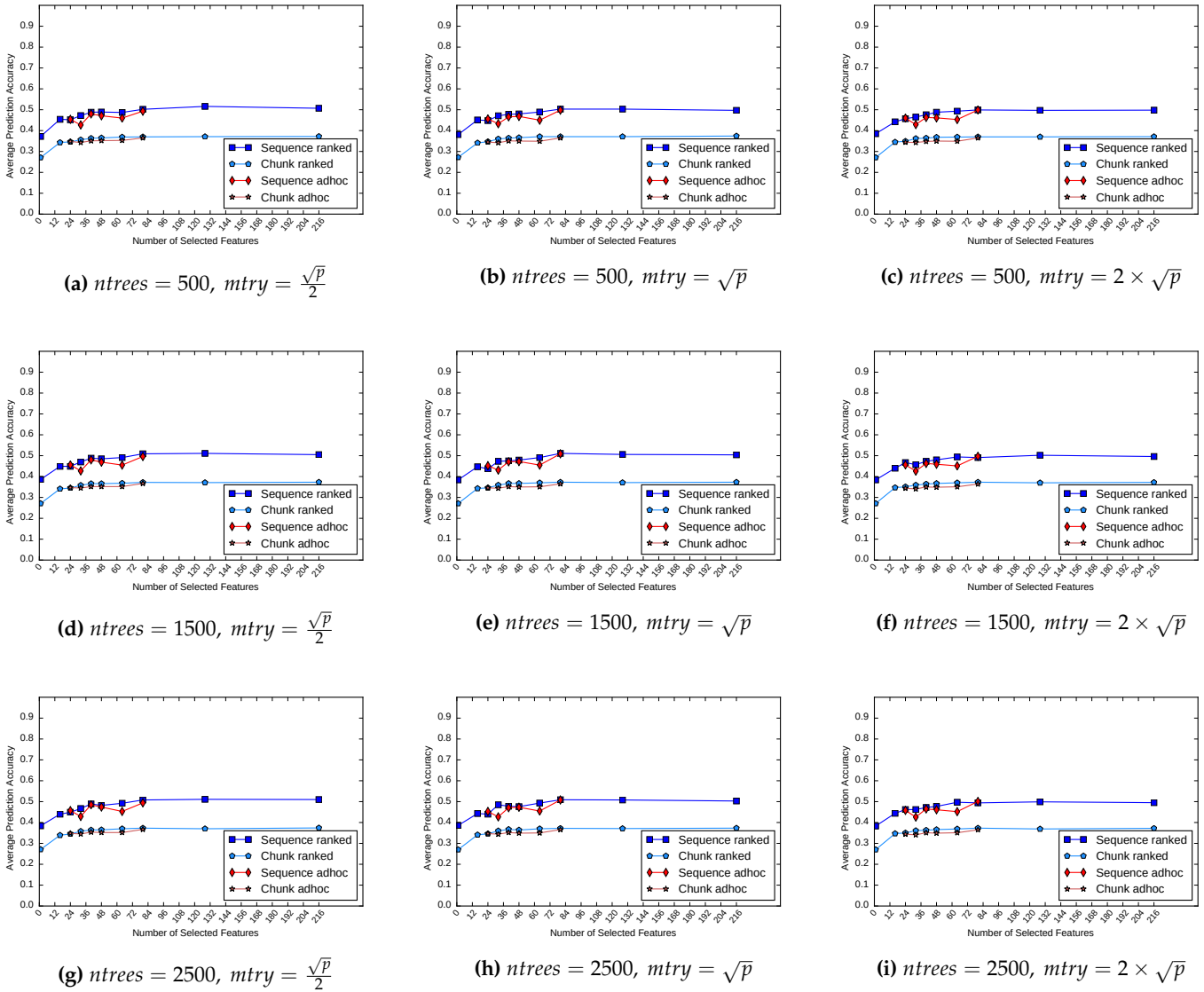


(a) $ntrees = 500,\ mtry = \frac{\sqrt{p}}{2}$ **(b)** $ntrees = 500,\ mtry = \sqrt{p}$ **(c)** $ntrees = 500,\ mtry = 2 \times \sqrt{p}$

(d) $ntrees = 1500,\ mtry = \frac{\sqrt{p}}{2}$ **(e)** $ntrees = 1500,\ mtry = \sqrt{p}$ **(f)** $ntrees = 1500,\ mtry = 2 \times \sqrt{p}$

(g) $ntrees = 2500,\ mtry = \frac{\sqrt{p}}{2}$ **(h)** $ntrees = 2500,\ mtry = \sqrt{p}$ **(i)** $ntrees = 2500,\ mtry = 2 \times \sqrt{p}$

Figure C.1: Behavior of accuracy curves for both *motion chunks* (yellow and light blue curves) and sequences (red and dark blue curves) for different combinations of RF hyper-parameters. Results were obtained from within subject *one sequence out* task and depict both ***adhoc*** (red and yellow curves) and ***ranked*** (light and dark blue curves) feature groups.



(a) $ntrees = 500,\ mtry = \frac{\sqrt{p}}{2}$ **(b)** $ntrees = 500,\ mtry = \sqrt{p}$ **(c)** $ntrees = 500,\ mtry = 2 \times \sqrt{p}$

(d) $ntrees = 1500,\ mtry = \frac{\sqrt{p}}{2}$ **(e)** $ntrees = 1500,\ mtry = \sqrt{p}$ **(f)** $ntrees = 1500,\ mtry = 2 \times \sqrt{p}$

(g) $ntrees = 2500,\ mtry = \frac{\sqrt{p}}{2}$ **(h)** $ntrees = 2500,\ mtry = \sqrt{p}$ **(i)** $ntrees = 2500,\ mtry = 2 \times \sqrt{p}$

Figure C.2: Behavior of accuracy curves for both *motion chunks* (yellow and light blue curves) and sequences (red and dark blue curves) for different combinations of RF hyper-parameters. Results were obtained from between subject task and depict both ***adhoc*** (red and yellow curves) and ***ranked*** (light and dark blue curves) feature groups.

The changes on the global ranking of features may be induced by: (a) noisy features that are deemed as important when small values of *mtry* are used, (b) an important correlation between features, or (c) an increase in the number of weak features considered as good candidates when a larger *mtry* value is used and that combined seem to produce better predictions than stronger features alone [30]. Ultimately, the results depicted in Figure C.1 and Figure C.2 show that our choice of default parameter for the RF classifier was appropriate in the context our of application. More importantly, accuracy rates obtained for the *adhoc group*, i.e., features of subsets extracted from end-effector trajectories, are robust to the choice of RF hyper-parameters.

Bibliography

[1] Sarah Fdili Alaoui, Frédéric Bevilacqua, Bertha Bermudez Pascual, and Christian Jacquemin. "Dance interaction with physical model visuals based on movement qualities". In: *International Journal of Arts and Technology* 6.4 (2013), pp. 357–387.

[2] Sarah Fdili Alaoui, Baptiste Caramiaux, Marcos Serrano, and Frédéric Bevilacqua. "Movement qualities as interaction modality". In: *Proceedings of the Designing Interactive Systems Conference*. ACM. 2012, pp. 761–769.

[3] Kenji Amaya, Armin Bruderlin, and Tom Calvert. "Emotion from motion". In: *Graphics interface*. Vol. 96. Toronto, Canada. 1996, pp. 222–229.

[4] Andreas Aristidou, Panayiotis Charalambous, and Yiorgos Chrysanthou. "Emotion Analysis and Classification: Understanding the Performers' Emotions Using the LMA Entities". In: *Computer Graphics Forum* 34.6 (2015), pp. 262–276. ISSN: 1467-8659. DOI: 10.1111/cgf.12598. URL: http://dx.doi.org/10.1111/cgf.12598.

[5] Andreas Aristidou and Yiorgos Chrysanthou. "Feature extraction for human motion indexing of acted dance performances". In: *Computer Graphics Theory and Applications (GRAPP), 2014 International Conference on*. IEEE. 2014, pp. 1–11.

[6] Anthony P Atkinson, Mary L Tunstall, and Winand H Dittrich. "Evidence for distinct contributions of form and motion information to the recognition of emotions from body gestures". In: *Cognition* 104.1 (2007), pp. 59–72.

[7] Anthony P Atkinson, Winand H Dittrich, Andrew J Gemmell, and Andrew W Young. "Emotion perception from dynamic and static body expressions in point-light and full-light displays". In: *Perception* 33 (2004), pp. 717–746.

[8] AP Atkinson. "Bodily expressions of emotion: visual cues and neural mechanisms". In: *The Cambridge handbook of human affective neuroscience* (2013), pp. 198–222.

[9] Matthieu Aubry, Frédéric Julliard, and Sylvie Gibet. "Modeling joint synergies to synthesize realistic movements". In: *International Gesture Workshop*. Springer. 2009, pp. 231–242.

[10] Norman I Badler, Cary B Phillips, and Bonnie Lynn Webber. *Simulating humans: computer graphics animation and control*. Oxford University Press, 1993.

[11] Paolo Baerlocher. "Inverse kinematics techniques for the interactive posture control of articulated figures". PhD thesis. Ecole Polytechnique Federale de Lausanne - EPFL, 2001.

[12] Paolo Baerlocher and Ronan Boulic. "An inverse kinematics architecture enforcing an arbitrary number of strict priority levels". In: *The visual computer* 20.6 (2004), pp. 402–417.

[13] Paolo Baerlocher and Ronan Boulic. "Parametrization and range of motion of the ball-and-socket joint". In: *Deformable avatars*. Springer, 2001, pp. 180–190.

[14] Tanja Bänziger, Marcello Mortillaro, and Klaus R Scherer. "Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception." In: *Emotion* 12.5 (2012), p. 1161.

[15] Tanja Bänziger and Klaus R Scherer. "Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus". In: *Affective computing and intelligent interaction*. Springer, 2007, pp. 476–487.

[16] Avi Barliya, Lars Omlor, Martin A Giese, Alain Berthoz, and Tamar Flash. "Expression of emotion in the kinematics of locomotion". In: *Experimental brain research* 225.2 (2013), pp. 159–176.

[17] Joseph Bates. "The Role of Emotion in Believable Agents". In: *Communications of the ACM* 37.7 (1994), pp. 122–125.

[18] Aryel Beck, Brett Stevens, Kim A Bard, and Lola Cañamero. "Emotional body language displayed by artificial agents". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2.1 (2012), p. 2.

[19] Asa Ben-Hur and Jason Weston. "A user's guide to support vector machines". In: *Data mining techniques for the life sciences* (2010), pp. 223–239.

[20] Daniel Bernhardt. "Emotion inference from human body motion". PhD thesis. University of Cambridge, 2010.

[21] Daniel Bernhardt and Peter Robinson. "Detecting affect from non-stylised body motions". In: *Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 59–70.

[22] Elisabetta Bevacqua, Igor Stanković, Ayoub Maatallaoui, Alexis Nédélec, and Pierre De Loor. "Effects of coupling in human-virtual agent body interaction". In: *International Conference on Intelligent Virtual Agents*. Springer. 2014, pp. 54–63.

[23] Gérard Biau and Erwan Scornet. "A random forest guided tour". In: *TEST* (2016), pp. 1–31. ISSN: 1863-8260. DOI: 10.1007/s11749-016-0481-7. URL: http://dx.doi.org/10.1007/s11749-016-0481-7.

[24] Rama Bindiganavale and Norman I Badler. "Motion abstraction and mapping with spatial constraints". In: *Modelling and Motion Capture Techniques for Virtual Environments*. Springer, 1998, pp. 70–82.

[25] Leslie Bishko. "Nonverbal Communication in Virtual Worlds". In: ed. by Joshua Tanenbaum, Magy Seif El-Nasr, and Michael Nixon. Pittsburgh, PA, USA: ETC Press, 2014. Chap. Our Empathic Experience of Believable Characters, pp. 47–59.

[26] Sebastian Bitzer. "Nonlinear dimensionality reduction for motion synthesis and control". PhD thesis. The University of Edinburgh, 2011.

[27] Randolph Blake and Maggie Shiffrar. "Perception of human motion". In: *Annu. Rev. Psychol.* 58 (2007), pp. 47–73.

[28] Alexandre Bouënard, Sylvie Gibet, and Marcelo M Wanderley. "Hybrid inverse motion control for virtual characters interacting with sound synthesis". In: *The Visual Computer* 28.4 (2012), pp. 357–370.

[29] Alexandre Bouënard, Marcelo MM Wanderley, and Sylvie Gibet. "Gesture control of sound synthesis: Analysis and classification of percussion gestures". In: *Acta Acustica united with Acustica* 96.4 (2010), pp. 668–677.

[30] Anne-Laure Boulesteix, Silke Janitza, Jochen Kruppa, and Inke R König. "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.6 (2012), pp. 493–507.

[31] Ronan Boulic, Ramon Mas, and Daniel Thalmann. "A robust approach for the control of the center of mass with inverse kinetics". In: *Computers & Graphics* 20.5 (1996), pp. 693–701.

[32] Jeroen JA van Boxtel and Hongjing Lu. "Joints and their relations as critical features in action discrimination: Evidence from a classification image method". In: *Journal of vision* 15.1 (2015), p. 20.

[33] Matthew Brand and Aaron Hertzmann. "Style machines". In: *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co. 2000, pp. 183–192.

[34] L. Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[35] Leo Breiman. "Bagging predictors". In: *Machine learning* 24.2 (1996), pp. 123–140.

[36] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[37] Samuel R Buss. "Introduction to inverse kinematics with jacobian transpose, pseudoinverse and damped least squares methods". In: *IEEE Journal of Robotics and Automation* 17.1-19 (2004), p. 16.

[38] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. "IEMOCAP: Interactive emotional dyadic motion capture database". In: *Language resources and evaluation* 42.4 (2008), pp. 335–359.

[39] Dumphna Callery. *Through the body: a pratical guide to physical theatre*. Nick Hern Books, 2001.

[40] Antonio Camurri, Ingrid Lagerlöf, and Gualtiero Volpe. "Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques". In: *International journal of human-computer studies* 59.1 (2003), pp. 213–225.

[41] Antonio Camurri, Barbara Mazzarino, Matteo Ricchetti, Renee Timmers, and Gualtiero Volpe. "Gesture-Based Communication in Human-Computer Interaction: 5th International Gesture Workshop, GW 2003, Genova, Italy, April 15-17, 2003, Selected Revised Papers". In: ed. by Antonio Camurri and Gualtiero Volpe. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. Chap. Multimodal Analysis of Expressive Gesture in Music and Dance Performances, pp. 20–39.

[42] Carnegie Mellon University. *Motion Capture Database*. 2003. URL: http://mocap.cs.cmu.edu/ (visited on 04/10/2014).

[47] Rich Caruana, Nikos Karampatziakis, and Ainur Yessenalina. "An empirical evaluation of supervised learning in high dimensions". In: *Proceedings of the 25th international conference on Machine learning*. ACM. 2008, pp. 96–103.

[48] Schubert R Carvalho, Ronan Boulic, Creto A Vidal, and Daniel Thalmann. "Latent motion spaces for full-body motion editing". In: *The Visual Computer* 29.3 (2013), pp. 171–188.

[49] J. Chai and J.K. Hodgins. "Performance animation from low-dimensional control signals". In: *ACM Transactions on Graphics (TOG)*. Vol. 24. 3. ACM. 2005, pp. 686–696.

[50] Jinxiang Chai and Jessica K Hodgins. "Constraint-based motion optimization using a statistical dynamic model". In: *ACM Transactions on Graphics (TOG)* 26.3 (2007), p. 8.

[51] Jinxiang Chai and Jessica K Hodgins. "Performance animation from low-dimensional control signals". In: *ACM Transactions on Graphics (TOG)*. Vol. 24. 3. ACM. 2005, pp. 686–696.

[52] Diane Chi, Monica Costa, Liwei Zhao, and Norman Badler. "The EMOTE model for effort and shape". In: *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co. 2000, pp. 173–182.

[53] Kwang-Jin Choi and Hyeong-Seok Ko. "On-line motion retargetting". In: *Computer Graphics and Applications, 1999. Proceedings. Seventh Pacific Conference on*. IEEE. 1999, pp. 32–42.

[54] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences. 2nd edn. Hillsdale, New Jersey: L*. 1988.

[55] Amazon.com company. *Amazon Mechanical Turk*. URL: https://www.mturk.com/mturk/welcome (visited on 06/27/2016).

[56] Mark Coulson. "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence". In: *Journal of nonverbal behavior* 28.2 (2004), pp. 117–139.

[57] Nicolas Courty. "Contributions to Analysis/Synthesis Schemes in Computer Animation". Habilitation à diriger des recherches. Université de Bretagne Sud, 2013.

[58] Roddy Cowie, Ellen Douglas-Cowie, and Cate Cox. "Beyond emotion archetypes: Databases for emotion modelling using neural networks". In: *Neural networks* 18.4 (2005), pp. 371–388.

[59] E. Crane and M. Gross. "Motion Capture and Emotion: Affect Detection in Whole Body Movement". In: vol. 4738. Affective Computing and Intelligent Interaction, ACII, LNCS. In Proc. of ACII, Springer Verlag, 2007, pp. 95–101.

[60] Nele Dael, Marcello Mortillaro, and Klaus R Scherer. "Emotion expression in body action and posture." In: *Emotion* 12.5 (2012), p. 1085.

[61] Anirban DasGupta. *Probability for statistics and machine learning: fundamentals and advanced topics*. Springer Science & Business Media, 2011.

[62] Marco De Meijer. "The contribution of general features of body movement to the attribution of emotions". In: *Journal of Nonverbal behavior* 13.4 (1989), pp. 247–268.

[63] P. Ravindra De Silva and Nadia Bianchi-Berthouze. "Modeling human affective postures: an information theoretic characterization of posture features". In: *Computer Animation and Virtual Worlds* 15.3-4 (2004), pp. 269–276.

[64] Janez Demšar. "Statistical comparisons of classifiers over multiple data sets". In: *The Journal of Machine Learning Research* 7 (2006), pp. 1–30.

[66] Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. "WaveNet: A Generative Model for Raw Audio". In: *arXiv preprint arXiv:1609.03499* (2016).

[67] Winand H Dittrich. "Action categories and the perception of biological motion". In: *Perception* 22.1 (1993), pp. 15–22.

[68] Winand H. Dittrich and Anthony P. Atkinson. "The perception of bodily expressions of emotion and the implications for computing". In: *Affective computing, focus on emotion expression, synthesis and recognition*. Ed. by Jimmy Or. I-Tech education and publishing, 2008. Chap. 9, pp. 157–184.

[69] Włodzisław Duch. "Feature Extraction: Foundations and Applications". In: ed. by Isabelle Guyon, Masoud Nikravesh, Steve Gunn, and Lotfi A. Zadeh. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. Chap. Filter Methods, pp. 89–117. ISBN: 978-3-540-35488-8. DOI: 10.1007/978-3-540-35488-8_4.

[70] Bradley Efron. "Bootstrap Methods: Another Look at the Jackknife". In: *Breakthroughs in Statistics: Methodology and Distribution*. Ed. by Samuel Kotz and Norman L. Johnson. New York, NY: Springer New York, 1992, pp. 569–593. ISBN: 978-1-4612-4380-9. DOI: 10.1007/978-1-4612-4380-9_41. URL: http://dx.doi.org/10.1007/978-1-4612-4380-9_41.

[71] P Ekman. "Are there basic emotions?" In: *Psychological review* 99.3 (1992), pp. 550–553.

[72] Paul Ekman and Wallace V Friesen. "Head and body cues in the judgment of emotion: A reformulation". In: *Perceptual and motor skills* 24.3 PT 1 (1967), pp. 711–724.

[73] Ahmed Elgammal and Chan-Su Lee. "The Role of Manifold Learning in Human Motion Analysis". In: *Human Motion: Understanding, Modelling, Capture, and Animation*. Ed. by Bodo Rosenhahn, Reinhard Klette, and Dimitris Metaxas. Dordrecht: Springer Netherlands, 2008, pp. 25–56. ISBN: 978-1-4020-6693-1. DOI: 10.1007/978-1-4020-6693-1_2.

[74] Morten Engell-Nørregård and Kenny Erleben. "Estimation of Joint types and Joint Limits from Motion capture data". In: *17-th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*. IEEE. 2009, pp. 9–16.

[75] Cathy Ennis, Ludovic Hoyet, Arjan Egges, and Rachel McDonnell. "Emotion Capture: Emotionally Expressive Characters for Games". In: *Proceedings of Motion on Games*. MIG '13. 2013, 31:53–31:60.

[76] Frank Enos and Julia Hirschberg. "A framework for eliciting emotional speech: Capitalizing on the actor's process". In: *First international workshop on emotion: Corpora for research on emotion and affect (international conference on language resources and evaluation (LREC 2006))*. 2006, pp. 6–10.

[77] S. A. Etemad and A. Arya. "Expert-Driven Perceptual Features for Modeling Style and Affect in Human Motion". In: *IEEE Transactions on Human-Machine Systems* 46.4 (2016), pp. 534–545. ISSN: 2168-2291. DOI: 10.1109/THMS.2016.2537760.

[78] S. Ali Etemad. "Perceptually Guided Processing of Style and Affect in Human Motion for Multimedia Applications". PhD thesis. Carleton University, 2014.

[79] S Ali Etemad and Ali Arya. "Classification and translation of style and affect in human motion using RBF neural networks". In: *Neurocomputing* 129 (2014), pp. 585–595.

[80] S Ali Etemad, Ali Arya, and Avi Parush. "Additivity in perception of affect from limb motion". In: *Neuroscience letters* 558 (2014), pp. 132–136.

[81] Donghui Feng, Sveva Besana, and Remi Zajac. "Acquiring high quality non-expert knowledge from on-demand workforce". In: *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*. Association for Computational Linguistics. 2009, pp. 51–56.

[82] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. "Do we need hundreds of classifiers to solve real world classification problems". In: *J. Mach. Learn. Res* 15.1 (2014), pp. 3133–3181.

[83] Tamar Flash and Neville Hogan. "The coordination of arm movements: an experimentally confirmed mathematical model". In: *The journal of Neuroscience* 5.7 (1985), pp. 1688–1703.

[84] Klaus Förger and Tapio Takala. "Animating with style: defining expressive semantics of motion". In: *The Visual Computer* 32.2 (2016), pp. 191–203.

[85] N. Fourati and C. Pelachaud. "Perception of emotions and body movement in the Emilya database". In: *IEEE Transactions on Affective Computing* PP.99 (2016), pp. 1–14. ISSN: 1949-3045. DOI: 10.1109/TAFFC.2016.2591039.

[86] Nesrine Fourati and Catherine Pelachaud. "Collection and Characterization of Emotional Body Behaviors". In: *Proceedings of the 2014 International Workshop on Movement and Computing*. MOCO '14. Paris, France: ACM, 2014, 49:49–49:54. ISBN: 978-1-4503-2814-2. DOI: 10.1145/2617995.2618004. URL: http://doi.acm.org/10.1145/2617995.2618004.

[87] Nesrine Fourati and Catherine Pelachaud. "Emilya: Emotional body expression in daily actions database." In: *LREC*. 2014, pp. 3486–3493.

[88] Nesrine Fourati and Catherine Pelachaud. "Relevant body cues for the classification of emotional body expression in daily actions". In: *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE. 2015, pp. 267–273.

[89] Jules Françoise, Agnès Roby-Brami, Natasha Riboud, and Frédéric Bevilacqua. "Movement Sequence Analysis Using Hidden Markov Models: A Case Study in Tai Chi Performance". In: *Proceedings of the 2Nd International Workshop on Movement and Computing*. MOCO '15. Vancouver, British Columbia, Canada: ACM, 2015, pp. 29–36. ISBN: 978-1-4503-3457-0. DOI: 10.1145/2790994.2791006. URL: http://doi.acm.org/10.1145/2790994.2791006.

[90] Tak-chung Fu. "A review on time series data mining". In: *Engineering Applications of Artificial Intelligence* 24.1 (2011), pp. 164–181.

[91] Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. "Variable selection using random forests". In: *Pattern Recognition Letters* 31.14 (2010), pp. 2225–2236.

[93] AL Gilet. "Procédures d'induction d'humeurs en laboratoire: une revue critique [Mood induction procedures: A critical review]". In: *L'encéphale* 34 (2008), pp. 233–239.

[94] Michael Girard and Anthony A Maciejewski. "Computational modeling for the computer animation of legged figures". In: *ACM SIGGRAPH Computer Graphics*. Vol. 19. 3. ACM. 1985, pp. 263–270.

[95] Donald Glowinski, Nele Dael, Antonio Camurri, Gualtiero Volpe, Marcello Mortillaro, and Klaus Scherer. "Toward a minimal representation of affective gestures". In: *Affective Computing, IEEE Transactions on* 2.2 (2011), pp. 106–118.

[96] Benjamin A Goldstein, Eric C Polley, and Farren Briggs. "Random forests for genetic association studies". In: *Statistical applications in genetics and molecular biology* 10.1 (2011).

[97] F Sebastian Grassia. "Practical parameterization of rotations using the exponential map". In: *Journal of graphics tools* 3.3 (1998), pp. 29–48.

[98] Baptiste Gregorutti, Bertrand Michel, and Philippe Saint-Pierre. "Correlation and variable importance in random forests". In: *arXiv preprint arXiv:1310.5726* (2013).

[99] J. Grezes, S. Pichon, and B. de Gelder. "Perceiving fear in dynamic body expressions." In: *NeuroImage* 35 (2007), pp. 959–967.

[100] Keith Grochow, Steven L Martin, Aaron Hertzmann, and Zoran Popović. "Style-based inverse kinematics". In: *ACM transactions on graphics (TOG)*. Vol. 23. 3. ACM. 2004, pp. 522–531.

[101] M Melissa Gross, Elizabeth A Crane, and Barbara L Fredrickson. "Methodology for assessing bodily expression of emotion". In: *Journal of Nonverbal Behavior* 34.4 (2010), pp. 223–248.

[102] Gutemberg Guerra-Filho and Arnab Biswas. "The human motion database: A cognitive and parametric sampling of human motion". In: *Image and Vision Computing* 30.3 (2012), pp. 251–261.

[103] Hatice Gunes and Maja Pantic. "Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners". In: *International conference on intelligent virtual agents*. Springer. 2010, pp. 371–377.

[104] Hatice Gunes and Massimo Piccardi. "A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior". In: *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*. Vol. 1. IEEE, pp. 1148–1153.

[105] Hatice Gunes, Caifeng Shan, Shizhi Chen, and YingLi Tian. "Bodily expression for automatic affect recognition". In: *Emotion Recognition: A Pattern Analysis Approach.(2015)* (2015).

[106] Shihui Guo, Richard Southern, Jian Chang, David Greer, and Jian Jun Zhang. "Adaptive motion synthesis for virtual characters: a survey". In: *The Visual Computer* 31.5 (2015), pp. 497–512.

[107] Isabelle Guyon and André Elisseeff. "An introduction to variable and feature selection". In: *The Journal of Machine Learning Research* 3 (2003), pp. 1157–1182.

[108] Wolfgang Härdle, Joel Horowitz, and Jens-Peter Kreiss. "Bootstrap methods for time series". In: *International Statistical Review* 71.2 (2003), pp. 435–459.

[109] Björn Hartmann, Maurizio Mancini, and Catherine Pelachaud. "Implementing Expressive Gesture Synthesis for Embodied Conversational Agents". In: *Gesture in Human-Computer Interaction and Simulation: 6th International Gesture Workshop, GW 2005, Berder Island, France, May 18-20, 2005, Revised Selected Papers*. Ed. by Sylvie Gibet, Nicolas Courty, and Jean-François Kamp. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 188–199. ISBN: 978-3-540-32625-0. DOI: 10.1007/11678816_22. URL: http://dx.doi.org/10.1007/11678816_22.

[110] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Vol. 1. Springer series in statistics Springer, Berlin, 2009.

[111] David J Hauser and Norbert Schwarz. "It's a trap! Instructional Manipulation checks prompt systematic thinking on "tricky" tasks". In: *SAGE Open* 5.2 (2015), p. 2158244015584617.

[112] Zhiying He, Xiaohui Liang, Jian Wang, Qinping Zhao, and Chengyu Guo. "Flexible editing of human motion by three-way decomposition". In: *Computer Animation and Virtual Worlds* 25.1 (2014), pp. 57–68.

[113] Rachel Heck and Michael Gleicher. "Parametric motion graphs". In: *Proceedings of the 2007 symposium on Interactive 3D graphics and games*. ACM. 2007, pp. 129–136.

[114] Alexis Heloir, Nicolas Courty, Sylvie Gibet, and Franck Multon. "Temporal alignment of communicative gesture sequences". In: *Computer animation and virtual worlds* 17.3-4 (2006), pp. 347–357.

[115] Lorna Herda, Raquel Urtasun, Pascal Fua, and Andrew Hanson. "Automatic determination of shoulder joint limits using quaternion field boundaries". In: *The International Journal of Robotics Research* 22.6 (2003), pp. 419–436.

[116] Jari K Hietanen, Jukka M Leppänen, and Ulla Lehtonen. "Perception of emotions in the hand movement quality of Finnish sign language". In: *Journal of nonverbal behavior* 28.1 (2004), pp. 53–64.

[117] Masahiro Hirai and Kazuo Hiraki. "The relative importance of spatial versus temporal structure in the perception of biological motion: An event-related potential study". In: *Cognition* 99.1 (2006), B15 –B29.

[118] Daniel Holden, Jun Saito, and Taku Komura. "A Deep Learning Framework for Character Motion Synthesis and Editing". In: *ACM Trans. Graph.* 35.4 (July 2016), 138:1–138:11. ISSN: 0730-0301. DOI: 10.1145/2897824.2925975. URL: http://doi.acm.org/10.1145/2897824.2925975.

[119] Eugene Hsu, Kari Pulli, and Jovan Popović. "Style translation for human motion". In: *ACM Transactions on Graphics (TOG)*. Vol. 24. 3. ACM. 2005, pp. 1082–1089.

[120] Eva Hudlicka. "Affective game engines: motivation and requirements". In: *Proceedings of the 4th international conference on foundations of digital games*. ACM. 2009, pp. 299–306.

[121] Bon-Woo Hwang, Sungmin Kim, and Seong-Whan Lee. "A full-body gesture database for automatic gesture recognition". In: *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. 2006, pp. 243–248. DOI: 10.1109/FGR.2006.8.

[122] Leslie Ikemoto and David A Forsyth. "Enriching a motion collection by transplanting limbs". In: *Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation*. Eurographics Association. 2004, pp. 99–108.

[123] Won-Seob Jang, Won-Kyu Lee, In-Kwon Lee, and Jehee Lee. "Enriching a motion database by analogous combination of partial human motions". In: *The Visual Computer* 24.4 (2008), pp. 271–280.

[124] Gunnar Johansson. "Visual perception of biological motion and a model for its analysis". In: *Perception & psychophysics* 14.2 (1973), pp. 201–211.

[125] Graham M. Jones. *Trade of tricks: inside the magician's craft*. University of California Press, 2011.

[126] Céline Jost, Pierre De Loor, Lexis Nédélec, Elisabetta Bevacqua, and Igor Stanković. "Real-time gesture recognition based on motion quality analysis". In: *Intelligent Technologies for Interactive Entertainment (INTETAIN), 2015 7th International Conference on*. IEEE. 2015, pp. 47–56.

[127] Marcelo Kallmann. "Analytical inverse kinematics with body posture control". In: *Computer Animation and Virtual Worlds* 19.2 (2008), pp. 79–91.

[128] Mubbasir Kapadia, I-kao Chiang, Tiju Thomas, Norman I Badler, Joseph T Kider Jr, et al. "Efficient motion retrieval in large motion databases". In: *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*. ACM. 2013, pp. 19–28.

[129] Asha Kapur, Ajay Kapur, Naznin Virji-Babul, George Tzanetakis, and Peter F. Driessen. "Gesture-Based Affective Computing on Motion Capture Data". In: *ACII'05*. 2005, pp. 1–7.

[130] Michelle Karg, Kolja Kühnlenz, and Martin Buss. "Recognition of affect based on gait patterns". In: *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 40.4 (2010), pp. 1050–1061.

[131] Michelle Karg, Ali-Akbar Samadani, Rob Gorbet, Kolja Kühnlenz, Jesse Hoey, and Dana Kulić. "Body movements for affective expression: A survey of automatic recognition and generation". In: *IEEE Transactions on Affective Computing* 4.4 (2013), pp. 341–359.

[132] Jongmin Kim, Yeongho Seol, and Jehee Lee. "Human motion reconstruction from sparse 3D motion sensors using kernel CCA-based regression". In: *Computer Animation and Virtual Worlds* 24.6 (2013), pp. 565–576. ISSN: 1546-427X. DOI: 10.1002/cav.1557. URL: http://dx.doi.org/10.1002/cav.1557.

[133] Yejin Kim and Michael Neff. "Component-based locomotion composition". In: *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. Eurographics Association. 2012, pp. 165–173.

[134] *KIT Whole-Body Human Motion Database*. URL: https://motion-database.humanoids.kit.edu/.

[135] Charles A Klein and Ching-Hsiang Huang. "Review of pseudoinverse control for use with kinematically redundant manipulators". In: *IEEE Transactions on Systems, Man, and Cybernetics* 2 (1983), pp. 245–250.

[136] Andrea Kleinsmith and Nadia Bianchi-Berthouze. "Affective body expression perception and recognition: A survey". In: *IEEE Transactions on Affective Computing* 4.1 (2013), pp. 15–33.

[137] Andrea Kleinsmith, Nadia Bianchi-Berthouze, and Anthony Steed. "Automatic recognition of non-acted affective postures". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41.4 (2011), pp. 1027–1038.

[138] Andrea Kleinsmith, P Ravindra De Silva, and Nadia Bianchi-Berthouze. "Cross-cultural differences in recognizing affect from body posture". In: *Interacting with Computers* 18.6 (2006), pp. 1371–1389.

[139] J. U. Korein and N. I. Badler. "Techniques for Generating the Goal-Directed Motion of Articulated Structures". In: *IEEE Comput. Graph. Appl.* 2.9 (Sept. 1982), pp. 71–81. ISSN: 0272-1716. DOI: 10.1109/MCG.1982.1674498. URL: http://dx.doi.org/10.1109/MCG.1982.1674498.

[140] Joe W Kotrlik and Heather A Williams. "The incorporation of effect size in information technology, learning, and performance research". In: *Information Technology, Learning, and Performance Journal* 21.1 (2003), p. 1.

[141] Lucas Kovar and Michael Gleicher. "Automated extraction and parameterization of motions in large data sets". In: *ACM Transactions on Graphics (TOG)*. Vol. 23. 3. ACM. 2004, pp. 559–568.

[142] Lucas Kovar, Michael Gleicher, and Frédéric Pighin. "Motion Graphs". In: *ACM Trans. Graph.* 21.3 (July 2002), pp. 473–482. ISSN: 0730-0301. DOI: 10.1145/566654.566605. URL: http://doi.acm.org/10.1145/566654.566605.

[143] Damjan Krstajic, Ljubomir J Buturovic, David E Leahy, and Simon Thomas. "Cross-validation pitfalls when selecting and assessing regression and classification models". In: *Journal of cheminformatics* 6.1 (2014), pp. 1–15.

[144] Björn Krüger, Jan Baumann, Mohammad Abdallah, and Andreas Weber. "A Study On Perceptual Similarity of Human Motions." In: *VRIPHYS*. 2011, pp. 65–72.

[145] Björn Krüger, Jochen Tautges, Andreas Weber, and Arno Zinke. "Fast local and global similarity searches in large motion capture databases". In: *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. Eurographics Association. 2010, pp. 1–10.

[146] R Kulpa, F Multon, and Bruno Arnaldi. "Morphology-independent representation of motions for interactive human-like animation". In: *Computer Graphics Forum*. Vol. 24. 3. Wiley Online Library. 2005, pp. 343–351.

[147] Rudolf Laban and Lisa Ullmann. "The mastery of movement." In: (1971).

[148] Thomas Navin Lal, Olivier Chapelle, Jason Weston, and André Elisseeff. "Feature Extraction: Foundations and Applications". In: ed. by Isabelle Guyon, Masoud Nikravesh, Steve Gunn, and Lotfi A. Zadeh. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. Chap. Embedded Methods, pp. 137–165. ISBN: 978-3-540-35488-8. DOI: 10.1007/978-3-540-35488-8_6.

[149] Joachim Lange and Markus Lappe. "The role of spatial and temporal information in biological motion perception". In: *Advances in Cognitive Psychology* 3.4 (2007), pp. 419–428.

[150] Benoît Le Callennec and Ronan Boulic. "Interactive motion deformation with prioritized constraints". In: *Graphical Models* 68.2 (2006), pp. 175–193.

[151] J. Lecoq, J.G. Carasso, J.C. Lallias, and D. Bradby. *The Moving Body (Le Corps Poetique): Teaching Creative Theatre*. Bloomsbury Academic, 2009.

[152] Jacques Lecoq. *Theater of movement and gesture*. Ed. by Bradby David. Taylor & Francis, 2006.

[153] Jehee Lee and Sung Yong Shin. "A hierarchical approach to interactive motion editing for human-like figures". In: *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co. 1999, pp. 39–48.

[154] Jehee Lee, Jinxiang Chai, Paul SA Reitsma, Jessica K Hodgins, and Nancy S Pollard. "Interactive control of avatars animated with human motion data". In: *ACM Transactions on Graphics (TOG)*. Vol. 21. 3. ACM. 2002, pp. 491–500.

[155] Sergey Levine, Jack M Wang, Alexis Haraux, Zoran Popović, and Vladlen Koltun. "Continuous character control with low-dimensional embeddings". In: *ACM Transactions on Graphics (TOG)* 31.4 (2012), p. 28.

[156] Yan Li, Tianshu Wang, and Heung-Yeung Shum. "Motion texture: a two-level statistical model for character motion synthesis". In: *ACM Transactions on Graphics (ToG)*. Vol. 21. 3. ACM. 2002, pp. 465–472.

[157] T Warren Liao. "Clustering of time series data—a survey". In: *Pattern recognition* 38.11 (2005), pp. 1857–1874.

[158] Andy Liaw and Matthew Wiener. "Classification and regression by randomForest". In: *R news* 2.3 (2002), pp. 18–22.

[159] Alain Liegeois. "Automatic supervisory control of the configuration and behavior of multibody mechanisms". In: *IEEE transactions on systems, man, and cybernetics* 7.12 (1977), pp. 868–871.

[160] Gengdai Liu, Zuoyan Lin, and Zhigeng Pan. "Style subspaces for character animation". In: *Computer Animation and Virtual Worlds* 19.3-4 (2008), pp. 199–209.

[161] H. Liu, X. Wei, J. Chai, I. Ha, and T. Rhee. "Realtime human motion control with a small number of inertial sensors". In: *Symposium on Interactive 3D Graphics and Games*. ACM. 2011, pp. 133–140.

[162] Gilles Louppe. "Understanding Random Forests: From Theory To Practice". PhD thesis. University of Liège, 2014.

[163] A Bryan Loyall. "Believable agents: building interactive personalities". PhD thesis. Mitsubishi Electric Research Laboratories, 1997.

[164] Wanli Ma, Shihong Xia, Jessica K Hodgins, Xiao Yang, Chunpeng Li, and Zhaoqi Wang. "Modeling style and variation in human motion". In: *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. Eurographics Association. 2010, pp. 21–30.

[165] Yingliang Ma, Helena M Paterson, and Frank E Pollick. "A motion capture library for the study of identity, gender, and emotion perception from biological motion". In: *Behavior research methods* 38.1 (2006), pp. 134–141.

[166] Anthony A Maciejewski. "Dealing with the ill-conditioned equations of motion for articulated figures". In: *IEEE Computer Graphics and Applications* 10.3 (1990), pp. 63–71.

[167] Stephen L Macknik, Mac King, James Randi, Apollo Robbins, John Thompson, Susana Martinez-Conde, et al. "Attention and awareness in stage magic: turning tricks into research". In: *Nature Reviews Neuroscience* 9.11 (2008), pp. 871–879.

[168] Nadia Magnenat-Thalmann, HyungSeok Kim, Arjan Egges, and Stephane Garchery. "Believability and interaction in virtual worlds". In: *null*. IEEE. 2005, pp. 2–9.

[169] Maurizio Mancini and Ginevra Castellano. "Real-time analysis and synthesis of emotional gesture expressivity". In: *Proc. of the Doctoral Consortium of Intl. Conf. on Affective Computing and Intelligent Interaction*. Citeseer. 2007.

[170] Mohammed Marey and François Chaumette. "New strategies for avoiding robot joint limits: Application to visual servoing using a large projection operator". In: *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. IEEE. 2010, pp. 6222–6227.

[171] Michael Mateas. "Interactive Drama, Art and Artificial Intelligence". PhD thesis. Pittsburgh, PA, USA: Carnegie Mellon University, 2002.

[172] Rachel McDonnell, Sophie Jörg, Joanna McHugh, Fiona N Newell, and Carol O'Sullivan. "Investigating the role of body shape on the perception of emotion". In: *ACM TAP* 6.3 (2009), p. 14.

[173] Albert Mehrabian. "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament". In: *Current Psychology* 14.4 (1996), pp. 261–292.

[174] Angeliki Metallinou, Zhaojun Yang, Chi-chun Lee, Carlos Busso, Sharon Carnicke, and Shrikanth Narayanan. "The USC CreativeIT database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations". In: *Language resources and evaluation* (2015), pp. 1–25.

[175] Microsoft Corporation. *Kinect Sensor*. 2012. URL: https://msdn.microsoft.com/en-us/library/hh438998.aspx (visited on 09/20/2016).

[176] Jianyuan Min and Jinxiang Chai. "Motion graphs++: a compact generative model for semantic motion analysis and synthesis". In: *ACM Transactions on Graphics (TOG)* 31.6 (2012), p. 153.

[177] V. Monbet and P.F. Marteau. "Non parametric resampling for stationary Markov processes: The local grid bootstrap approach". In: *Journal of Statistical Planning and Inference* 136.10 (2006), pp. 3319 –3338.

[178] Joann M Montepare, Sabra B Goldstein, and Annmarie Clausen. "The identification of emotions from gait information". In: *Journal of Nonverbal Behavior* 11.1 (1987), pp. 33–42.

[179] R Mukundan. "Quaternions: From classical mechanics to computer graphics, and beyond". In: *Proceedings of the 7th Asian Technology conference in Mathematics*. 2002, pp. 97–105.

[180] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. *Documentation Mocap Database HDM05*. Tech. rep. Universität Bonn, 2007.

[181] S. Murray and J. Keefe. *Physical Theatres: A Critical Introduction*. Taylor & Francis, 2007.

[182] Thibaut Le Naour. "Utilisation des relations spatiales pour lánalyse et lédition de mouvement". PhD thesis. Université de Bretagne Sud, 2012.

[183] Michael Neff. "Nonverbal Communication in Virtual Worlds: Understanding and Designing Expressive Characters". In: ed. by Joshua Tanenbaum, Magy Seif El-Nasr, and Michael Nixon. Pittsburgh, PA, USA: ETC Press, 2014. Chap. Lessons From The Arts: What The Performing Arts Literature Can Teach Us About Creating Expressive Character Movement, pp. 123–146.

[184] Michael Neff and Eugene Fiume. "AER: aesthetic exploration and refinement for expressive character animation". In: *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*. ACM. 2005, pp. 161–170.

[185] Michael Neff and Eugene Fiume. "From performance theory to character animation tools". In: *Human Motion*. Springer, 2008, pp. 597–629.

[186] Radoslaw Niewiadomski, Maurizio Mancini, and Stefano Piana. "Human and virtual agent expressive gesture quality analysis and synthesis". In: *Coverbal Synchrony in Human-Machine Interaction* (2013), pp. 269–292.

[187] Aline Normoyle, Fannie Liu, Mubbasir Kapadia, Norman I Badler, and Sophie Jörg. "The effect of posture and dynamics on the perception of emotion". In: *Proceedings of the ACM Symposium on Applied Perception*. ACM. 2013, pp. 91–98.

[188] *NUS MOCAP*. URL: http://animation.comp.nus.edu.sg/nusmocap.html.

[189] Kensuke Onuma, Christos Faloutsos, and Jessica K Hodgins. "FMDistance: A fast and effective distance function for motion capture data". In: *Short Papers Proceedings of EUROGRAPHICS* 2 (2008).

[190] Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. "Instructional manipulation checks: Detecting satisficing to increase statistical power". In: *Journal of Experimental Social Psychology* 45.4 (2009), pp. 867–872.

[191] Jim R Parker. "Theater as virtual reality". In: *Nonverbal Communication in Virtual Worlds*. ETC Press. 2014, pp. 151–174.

[192] Catherine Pelachaud. "Studies on gesture expressivity for a virtual agent". In: *Speech Communication* 51.7 (2009), pp. 630–639.

[193] *Performing arts*. URL: https://en.wikipedia.org/wiki/Theatre.

[194] Ken Perlin and Athomas Goldberg. "Improv: A System for Scripting Interactive Actors in Virtual Worlds". In: *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '96. New York, NY, USA: ACM, 1996, pp. 205–216. ISBN: 0-89791-746-4. DOI: 10.1145/237170.237258. URL: http://doi.acm.org/10.1145/237170.237258.

[195] Lankoski Petri and Björk Staffan. "Gameplay Design Patterns for Believable Non-Player Characters". In: *Situated Play: Proceedings of the 2007 Digital Games Research Association International Conference*. The University of Tokyo, 2007, pp. 416–423.

[196] Dimitris N Politis et al. "The impact of bootstrap methods on time series analysis". In: *Statistical Science* 18.2 (2003), pp. 219–230.

[197] F.E. Pollick, H.M. Paterson, . A Bruderlin, and A.J Sanford. "Perceiving affect from arm movement". In: *Cognition* 82.2 (2001), B51–61.

[198] Frank E. Pollick and HM Paterson. "Movement style, movement features, and the recognition of affect from human movement". In: *Understanding events: From perception to action* (2008), pp. 286–308.

[199] M Pražák, R McDonnell, L Kavan, and C O'Sullivan. "A perception based metric for comparing human locomotion". In: *Eurographics Ireland* (2009).

[200] *Qualysis, Motion Capture System*. URL: www.qualysis.com.

[201] Carl Edward Rasmussen. "Gaussian processes for machine learning". In: (2006).

[202] Daniel Raunhardt and Ronan Boulic. "Motion constraint". In: *The Visual Computer* 25.5-7 (2009), pp. 509–518.

[203] Daniel Raunhardt and Ronan Boulic. "Progressive clamping". In: *Proceedings 2007 IEEE International Conference on Robotics and Automation*. IEEE. 2007, pp. 4414–4419.

[204] Liu Ren, Alton Patrick, Alexei A Efros, Jessica K Hodgins, and James M Rehg. "A data-driven approach to quantifying natural human motion". In: *ACM Transactions on Graphics (TOG)*. Vol. 24. 3. ACM. 2005, pp. 1090–1097.

[205] Claire L Roether, Lars Omlor, Andrea Christensen, and Martin A Giese. "Critical features for the perception of emotion from gait". In: *Journal of Vision* 9.6 (2009), p. 15.

[206] Charles Rose, Michael F Cohen, and Bobby Bodenheimer. "Verbs and adverbs: Multidimensional motion interpolation". In: *IEEE Computer Graphics and Applications* 18.5 (1998), pp. 32–40.

[207] James A Russell. "A circumplex model of affect." In: *Journal of personality and social psychology* 39.6 (1980), p. 1161.

[208] Martin Saerbeck and Christoph Bartneck. "Perception of affect elicited by robot motion". In: *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*. IEEE Press. 2010, pp. 53–60.

[209] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. "A review of feature selection techniques in bioinformatics". In: *bioinformatics* 23.19 (2007), pp. 2507–2517.

[210] Alla Safonova, Jessica K Hodgins, and Nancy S Pollard. "Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces". In: *ACM Transactions on Graphics (TOG)* 23.3 (2004), pp. 514–521.

[211] Ali-Akbar Samadani. "Automatic Recognition and Generation of Affective Movements". PhD thesis. University of Waterloo, 2014.

[212] Ali-Akbar Samadani, Ali Ghodsi, and Dana Kulić. "Discriminative functional analysis of human movements". In: *Pattern Recognition Letters* 34.15 (2013), pp. 1829–1839.

[213] Ali-Akbar Samadani, Rob Gorbet, and Dana Kulić. "Affective movement recognition based on generative and discriminative stochastic dynamic models". In: *IEEE Transactions on Human-Machine Systems* 44.4 (2014), pp. 454–467.

[214] David Sander. "Models of Emotion: The Affective Neuroscience Approach". In: *The Cambridge handbook of human affective neuroscience* (2013), pp. 5–53.

[215] Nikolaos Savva and Nadia Bianchi-Berthouze. "Automatic recognition of affective body movement in a video game scenario". In: *International Conference on Intelligent Technologies for interactive entertainment*. Springer. 2011, pp. 149–159.

[216] Misako Sawada, Kazuhiro Suda, and Motonobu Ishii. "Expression of emotions in dance: Relation between arm movement characteristics and emotion". In: *Perceptual and motor skills* 97.3 (2003), pp. 697–708.

[217] Klaus R Scherer. "Emotion and emotional competence: conceptual and theoretical issues for modelling agents". In: *Blueprint for Affective Computing* (2010), pp. 3–20.

[218] Marc Schroder, Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark Ter Maat, Gary McKeown, Sathish Pammi, Maja Pantic, et al. "Building autonomous sensitive artificial listeners". In: *IEEE Transactions on Affective Computing* 3.2 (2012), pp. 165–183.

[219] *scikit-learn*. URL: http://scikit-learn.org/stable/.

[220] Simon Senecal, Louis Cuel, Andreas Aristidou, and Nadia Magnenat-Thalmann. "Continuous body emotion recognition system during theater performances". In: *Computer Animation and Virtual Worlds* 27.3-4 (2016), pp. 311–320.

[221] Ari Shapiro, Yong Cao, and Petros Faloutsos. "Style components". In: *Proceedings of Graphics Interface 2006*. Canadian Information Processing Society. 2006, pp. 33–39.

[222] Ari Shapiro, Marcelo Kallmann, and Petros Faloutsos. "Interactive motion correction and object manipulation". In: *Proceedings of the 2007 symposium on Interactive 3D graphics and games*. ACM. 2007, pp. 137–144.

[223] Hideaki Shimazaki and Shigeru Shinomoto. "A method for selecting the bin size of a time histogram". In: *Neural computation* 19.6 (2007), pp. 1503–1527.

[224] Hyun Joon Shin, Jehee Lee, Sung Yong Shin, and Michael Gleicher. "Computer puppetry: An importance-based approach". In: *ACM Transactions on Graphics (TOG)* 20.2 (2001), pp. 67–94.

[225] Vin D Silva and Joshua B Tenenbaum. "Global versus local methods in nonlinear dimensionality reduction". In: *Advances in neural information processing systems*. 2002, pp. 705–712.

[226] C.K.F. So and G. Baciu. "Entropy-based motion extraction for motion capture animation: Motion Capture and Retrieval". In: *Computer Animation and Virtual Worlds* 16.3-4 (2005), pp. 225–235.

[227] AA Sokolov, S Krüger, P Enck, I Krägeloh-Mann, and MA Pavlova. "Gender affects body language reading." In: *Frontiers in psychology* 2 (2011), pp. 2–16.

[228] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, and B.P. Feuston. "Random forest: a classification and regression tool for compound classification and QSAR modeling". In: *Journal of chemical information and computer sciences* 43.6 (2003), pp. 1947–1958.

[229] Jiliang Tang, Salem Alelyani, and Huan Liu. "Feature selection for classification: A review". In: *Data Classification: Algorithms and Applications* (2014), p. 37.

[230] J. Tautges, A. Zinke, B. Krüger, J. Baumann, A. Weber, T. Helten, M. Müller, H.P. Seidel, and B. Eberhardt. "Motion reconstruction using sparse accelerometer data". In: *ACM Transactions on Graphics (TOG)* 30.3 (2011), p. 18.

[231] Graham W Taylor and Geoffrey E Hinton. "Factored conditional restricted Boltzmann machines for modeling motion style". In: *Proceedings of the 26th annual international conference on machine learning*. ACM. 2009, pp. 1025–1032.

[232] Graham W Taylor, Geoffrey E Hinton, and Sam T Roweis. "Modeling human motion using binary latent variables". In: *Advances in neural information processing systems*. 2006, pp. 1345–1352.

[233] *Theatre*. URL: https://en.wikipedia.org/wiki/Theatre.

[234] Steven M Thurman, Martin A Giese, and Emily D Grossman. "Perceptual and computational analysis of critical features for biological motion". In: *Journal of Vision* 10.12 (2010), p. 15.

[235] Joëlle Tilmanne, Alexis Moinet, and Thierry Dutoit. "Stylistic gait synthesis based on hidden Markov models". In: *EURASIP Journal on advances in signal processing* 2012.1 (2012), pp. 1–14.

[236] Deepak Tolani, Ambarish Goswami, and Norman I Badler. "Real-time inverse kinematics techniques for anthropomorphic limbs". In: *Graphical models* 62.5 (2000), pp. 353–388.

[237] Steve Tonneau. "Motion planning and synthesis for virtual characters in constrained environments". PhD thesis. Rennes, France: INSA Rennes, 2015.

[238] Lorenzo Torresani, Peggy Hackney, and Christoph Bregler. "Learning motion style synthesis from perceptual observations". In: *NIPS*. 2006, pp. 1393–1400.

[239] Maxime Tournier, Xiaomao Wu, Nicolas Courty, Elise Arnaud, and Lionel Reveret. "Motion compression using principal geodesics analysis". In: *Computer Graphics Forum*. Vol. 28. 2. Wiley Online Library. 2009, pp. 355–364.

[240] Nikolaus F Troje. "Decomposing biological motion: A framework for analysis and synthesis of human gait patterns". In: *Journal of vision* 2.5 (2002), pp. 371–387.

[241] Nikolaus F Troje, Cord Westhoff, and Mikhail Lavrov. "Person identification from biological motion: Effects of structural and kinematic cues". In: *Perception & Psychophysics* 67.4 (2005), pp. 667–675.

[242] Arthur Truong, Hugo Boujut, and Titus Zaharia. "Laban descriptors for gesture recognition and emotional analysis". In: *The Visual Computer* 32.1 (2016), pp. 83–98.

[243] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley Series in Behavioral Science: Quantitative Methods, 1997.

[244] Eugene Tuv. "Feature Extraction: Foundations and Applications". In: Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. Chap. Ensamble Methods, pp. 187–204. ISBN: 978-3-540-35488-8. DOI: 10.1007/978-3-540-35488-8_6.

[245] Eugene Tuv, Alexander Borisov, George Runger, and Kari Torkkola. "Feature selection with ensembles, artificial variables, and redundancy elimination". In: *The Journal of Machine Learning Research* 10 (2009), pp. 1341–1366.

[246] University of Texas at Arlington. *Human motion database*. 2011. URL: http://smile.uta.edu/hmd/ (visited on 04/10/2014).

[247] Munetoshi Unuma, Ken Anjyo, and Ryozo Takeuchi. "Fourier principles for emotion-based human figure animation". In: *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. ACM. 1995, pp. 91–96.

[248] Raquel Urtasun, Pascal Glardon, Ronan Boulic, Daniel Thalmann, and Pascal Fua. "Style-based motion synthesis". In: *Computer Graphics Forum*. Vol. 23. 4. Wiley Online Library. 2004, pp. 799–812.

[249] Jakub Valcik, Jan Sedmidubsky, and Pavel Zezula. "Assessing similarity models for human-motion retrieval applications". In: *Computer Animation and Virtual Worlds* (2015).

[250] Herwin Van Welbergen, Ben JH Van Basten, Arjan Egges, Zs M Ruttkay, and Mark H Overmars. "Real Time Animation of Virtual Humans: A Trade-off Between Naturalness and Control". In: *Computer Graphics Forum*. Vol. 29. 8. Wiley Online Library. 2010, pp. 2530–2554.

[251] Gentiane Venture, Hideki Kadone, Tianxiang Zhang, Julie Grèzes, Alain Berthoz, and Halim Hicheur. "Recognizing emotions conveyed by human gait". In: *International Journal of Social Robotics* 6.4 (2014), pp. 621–632.

[252] P. Viviani and C. Terzuolo. "Trajectory determines movement dynamics". In: *Neuroscience* 7.2 (1982), pp. 431 –437.

[253] Ekaterina Volkova, Stephan De La Rosa, Heinrich H Bülthoff, and Betty Mohler. "The MPI emotional body expressions database for narrative scenarios". In: *PloS one* 9.12 (2014), e113647.

[254] Ekaterina P Volkova, Betty J Mohler, Trevor J Dodds, Joachim Tesch, and Heinrich H Bülthoff. "Emotion categorization of body expressions in narrative scenarios". In: *Frontiers in psychology* 5 (2014), pp. 623–644.

[255] H.G. Wallbott. "Bodily expression of emotion". In: *European Journal of Social Psychology* 28 (1998), pp. 879–896.

[256] Charles W Wampler. "Manipulator inverse kinematic solutions based on vector formulations and damped least-squares methods". In: *IEEE Transactions on Systems, Man, and Cybernetics* 16.1 (1986), pp. 93–101.

[257] Matt P Wand and M Chris Jones. *Kernel smoothing*. Crc Press, 1994.

[258] Jack M Wang, David J Fleet, and Aaron Hertzmann. "Gaussian process dynamical models for human motion". In: *IEEE transactions on pattern analysis and machine intelligence* 30.2 (2008), pp. 283–298.

[259] Jack M Wang, David J Fleet, and Aaron Hertzmann. "Multifactor Gaussian process models for style-content separation". In: *Proceedings of the 24th international conference on Machine learning*. ACM. 2007, pp. 975–982.

[260] Weiyi Wang, Valentin Enescu, and Hichem Sahli. "Adaptive Real-Time Emotion Recognition from Body Movements". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5.4 (2015), p. 18.

[261] Chris Welman. "Inverse kinematics and geometric constraints for articulated figure manipulation". PhD thesis. Simon Fraser University, 1993.

[262] Rainer Westermann, GUNTER STAHL, and F Hesse. "Relative effectiveness and validity of mood induction procedures: analysis". In: *European Journal of social psychology* 26 (1996), pp. 557–580.

[263] Jane Wilhelms and Allen Van Gelder. "Efficient spherical joint limits with reach cones". In: *Apr* 17 (2001), pp. 1–13.

[264] William A Wolovich and H Elliott. "A computational technique for inverse kinematics". In: *Decision and Control, 1984. The 23rd IEEE Conference on*. IEEE. 1984, pp. 1359–1363.

[265] Andy T Woods, Carlos Velasco, Carmel A Levitan, Xiaoang Wan, and Charles Spence. "Conducting perception research over the internet: a tutorial review". In: *PeerJ* 3 (2015), e1058.

[266] Shihong Xia, Congyi Wang, Jinxiang Chai, and Jessica Hodgins. "Realtime style transfer for unlabeled heterogeneous human motion". In: *ACM Transactions on Graphics (TOG)* 34.4 (2015), p. 119.

[267] Zhengzheng Xing, Jian Pei, and Eamonn Keogh. "A brief survey on sequence classification". In: *ACM SIGKDD Explorations Newsletter* 12.1 (2010), pp. 40–48.

[268] Junchao Xu, Joost Broekens, Koen Hindriks, and Mark A Neerincx. "Mood expression through parameterized functional behavior of robots". In: *2013 IEEE RO-MAN*. IEEE. 2013, pp. 533–540.

[269] Lei Xu, Adam Krzyżak, and Ching Y Suen. "Methods of combining multiple classifiers and their applications to handwriting recognition". In: *Systems, Man and Cybernetics, IEEE Transactions on* 22.3 (1992), pp. 418–435.

[270] Katsu Yamane and Yoshihiko Nakamura. "Natural motion animation through constraining and deconstraining at will". In: *IEEE Transactions on visualization and computer graphics* 9.3 (2003), pp. 352–360.

[271] Katja Zibrek, Ludovic Hoyet, Kerstin Ruhland, and Rachel McDonnell. "Evaluating the Effect of Emotion on Gender Recognition in Virtual Humans". In: *SAP '13*. 2013, pp. 45–49.