*Technical Report*

UCAM-CL-TR-787<br>
ISSN 1476-2986

Number 787

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

# Emotion inference from human body motion

Daniel Bernhardt

October 2010

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

*http://www.cl.cam.ac.uk/*

This technical report is based on a dissertation submitted January 2010 by the author for the degree of Doctor of Philosophy to the University of Cambridge, Selwyn College.

Technical reports published by the University of Cambridge Computer Laboratory are freely available via the Internet:

*http://www.cl.cam.ac.uk/techreports/*

ISSN 1476-2986

# Summary

The human body has evolved to perform sophisticated tasks from locomotion to the use of tools. At the same time our body movements can carry information indicative of our intentions, inter-personal attitudes and emotional states. Because our body is specialised to perform a variety of everyday tasks, in most situations emotional effects are only visible through subtle changes in the *qualities* of movements and actions. This dissertation focuses on the automatic analysis of emotional effects in everyday actions.

In the past most efforts to recognise emotions from the human body have focused on *expressive gestures* which are archetypal and exaggerated expressions of emotions. While these are easier to recognise by humans and computational pattern recognisers they very rarely occur in natural scenarios. The principal contribution of this dissertation is hence the inference of emotional states from *everyday actions* such as walking, knocking and throwing. The implementation of the system draws inspiration from a variety of disciplines including psychology, character animation and speech recognition. Complex actions are modelled using Hidden Markov Models and motion primitives.

The manifestation of emotions in everyday actions is very subtle and even humans are far from perfect at picking up and interpreting the relevant cues because emotional influences are usually minor compared to constraints arising from the action context or differences between individuals. This dissertation describes a *holistic* approach which models emotional, action and personal influences in order to maximise the discriminability of different emotion classes. A pipeline is developed which incrementally removes the biases introduced by different action contexts and individual differences. The resulting signal is described in terms of posture and dynamic features and classified into one of several emotion classes using statistically trained Support Vector Machines. The system also goes beyond isolated expressions and is able to classify natural action sequences. I use Level Building to segment action sequences and combine component classifications using an incremental voting scheme which is suitable for online applications.

The system is comprehensively evaluated along a number of dimensions using a corpus of motion-captured actions. For isolated actions I evaluate the generalisation performance to new subjects. For action sequences I study the effects of reusing models trained on the isolated cases *vs.* adapting models to connected samples. The dissertation also evaluates the role of modelling the influence of individual user differences. I develop and evaluate a regression-based adaptation scheme. The results bring us an important step closer to recognising emotions from body movements, embracing the complexity of body movements in natural scenarios.

# Acknowledgments

Embarking on a PhD is not an easy endeavour and I have been lucky enough to have had the support of the best colleagues, friends and family one could wish for. First and foremost I would like to thank my supervisor Peter Robinson. He has been an invaluable source of advice and support and his trust and generosity enabled me to gain unique experience. I would also like to thank my second supervisor Neil Dodgson for his support and Maja Pantic and John Daugman for their insightful comments.

A large part of my PhD work built upon the data carefully collected at the Department of Psychology, University of Glasgow. The data has been of tremendous value to me and I would like to thank everyone involved in the project at Glasgow for sharing the resource with the research community. In particular, I would like to thank Frank Pollick and Helena Patterson for sharing their insights and offering their help.

I could not have collected my own data without the help of a number of people. Most importantly, I would like to thank Joe Osborne for his work and shared enthusiasm. He wrote the audio environment and Vicon handlers and helped to collect audio samples. Shazia Afzal, Ursula Augsdörfer, Cecily Morrison, Bjarki Holm, Rok Strnisa and Phil Tuddenham offered their time as participants.

I have been lucky enough to be part of a stimulating research environment and I would like to thank Peter, Neil and Alan for the stimulating conversations over coffee and the BBQs. I would like to thank Metin Sezgin for many thought-provoking discussions, some of which were related to the work in this dissertation, others less so — I throughly enjoyed my time in SS12. Special thanks go to Metin, Shazia and Ian Davies for proof-reading parts of this dissertation. I am indepted to Cynthia Breazeal and everyone at the Personal Robots Group at MIT for providing me with the most rewarding experience which significantly shaped my own approach to research. Many thanks to Phil for being a formidable office mate and great source of advice and to Chris Nash for ensuring a good laugh in and outside the office.

My family has been a tremendous source of support, during stressful and more relaxed times. I would like to thank my parents Karlheinz and Michaela for enabling me to chase my dreams and my sister Marie for always being able to lift my spirits. Xixi, you have been my sunshine whenever I couldn't see the end of the tunnel. I am eternally grateful for the lovingly sobering reality checks without which I wouldn't be where I am today.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Towards emotionally intelligent machines

The ubiquitous deployment of computer technology throughout peoples' private and professional lives has been a trend which seems set to continue. New technologies often hold the potential to make our lives more pleasurable or manageable. However, there is a constant conflict between the utility technology provides and the complexity of operating it. The traditional command-and-control interaction paradigm has been largely technology-centred, leaving the human user to learn and adapt to its use. Three recent trends mean that this traditional paradigm will not be scalable to future technologies.

- The number of computers found in our environment is becoming large enough to make a traditional technology-centred interaction mode inefficient and frustrating. For example, the complexity of many modern-day car dashboards is likely to confuse even the most technology-savvy users. Often the requirement to interact with additional technologies such as satellite navigation or mobile phones increases the mental strain on drivers still further.

- Hardware advances mean that devices are constantly shrinking in size, making physical controls impractical. For example, the current generation of portable music players and mobile phones is hitting this limit and new interaction styles such as voice control are starting to become mainstream.

- The use of computer technology is becoming a fundamental requirement for a majority of people from all age groups, backgrounds and abilities. This means that technology-centric designs which are targeted at specialised and trained users are no longer optimal for all potential user groups.

A very popular goal for computer science research has therefore been to endow machines with more intelligence in order to reduce the amount of explicit information users need to provide through command and control interfaces. The goal is therefore to give machines more human-like abilities to communicate with users. One highly important aspect in the domain of human-human interaction is the communication of emotions. Being able to infer emotional states through non-verbal cues allows humans to empathise and reason about each others' underlying intents and goals. *Affective Computing* is the branch of computer science which aims to exploit the power of emotions to facilitate more effective human-machine interaction. Picard identified the goal of Affective Computing as giving machines the ability to recognise, express and regulate emotions [Pic97].

Multiple scenarios where these kinds of emotionally intelligent machines could aid human-machine interaction have been described over the last years. As the explicit control of all computational devices in our environment becomes overwhelming, explicit command and control could be replaced by wearable devices [LN04, KTP06] or sentient environments [KFN07] which automatically pick up a person's emotion-communicating cues, infer user intents and react to them appropriately. As electronic devices such as portable music players are becoming too small to make the physical interaction through buttons *etc.* impractical, new sensors such as motion-sensitive could accelerometers allow the communication of emotions to control the device either implicitly or explicitly (*e.g.* see eMoto's use of emotional gestures [FSH04, SSH05]). Finally, endowing technology with emotional intelligence is one way of making interactions more natural and therefore suitable for diverse user groups, such as autistic or elderly users. For example, research aiming to introduce robots into everyday scenarios usually addresses issues of emotion detection from human users and emotion modelling and expression within the robot [Bre04, FA05].

## 1.2 The role of the body

In this dissertation I will look in detail at the role of the body for communicating emotions. I am aiming to develop a computational model for detecting emotions from general body motions. This kind of system could eventually be used by wearable or sentient systems as well as social robots to sense the emotions of humans from cues in their movements.

A particular aim of this dissertation is to develop an approach which is as applicable as possible to real-world situations in which emotional body cues are likely to occur. In particular, I am not focusing on traditional human-computer interaction scenarios in which a user is engaged in front of a single desktop machine. Instead, I am primarily interested in the expression of emotions in everyday scenarios in which activities like walking, picking up objects *etc.* are modulated by emotional qualities.

This approach also means that I will avoid the analysis of previously studied *emotion archetypes*. These are body expressions which are exaggerated to express a certain emotion. While exaggerated emotional expressions may be naturally occurring during the communication with infants [TAD00, AZ07], in most everyday interactions our emotional expressions are softened by display rules and constraints arising from concurrent activities. While everyone would recognise an archetypal body expression of joy with an erect upper body, raised head and energetically raised arms, in a more natural scenario joy might only manifest itself in a subtly different walking style. In this dissertation I will focus on analysing the latter problem and to detect emotion as it appears in everyday activities.

## 1.3 Open research problems and challenges

Being such a salient aspect of human-human communication, automatic emotion recognition has been an active research area over the last decade. Most efforts have focused on separate modalities, targeting facial expressions and head gestures, voice quality, physiological signals such as skin conductance and more recently brain signals. Other researchers have studied the expression of emotions through different aspects of body motions such as posture, emotional gestures, archetypal expressions or movements qualities. In many cases those efforts have not been motivated by practical applications and early databases have been largely collected under controlled laboratory conditions. While many of the previous studies have made significant progress towards the construction of emotionally aware machines, there are a number of open research problems. In this thesis, I am aiming to advance the knowledge with respect to the following key problems which pose significant research challenges:

1. The communication of emotions through body expressions is generally *less understood* than for other modalities, especially the voice and face. The field of psychology which has in the past inspired many computational approaches has traditionally ignored bodily cues and more recent results need to be evaluated and incorporated into computational systems.

2. To the best of my knowledge there have been no previous attempts to infer emotions automatically from a *variety of everyday actions.* In this scenario most of the motion signal is dominated by the action performed and emotional variations are very subtle. The motions to analyse are also likely to be subject to greater statistical noise as the recording procedures are less controlled than for more archetypal expressions and actions can blend into each other seamlessly.

3. These kinds of natural scenarios also give rise to greater *inter-personal differences.* Although a few approaches for other modalities have targeted the issue of personal differences, it has been largely neglected by the affective computing community and is virtually unexplored for the field of body expressions.

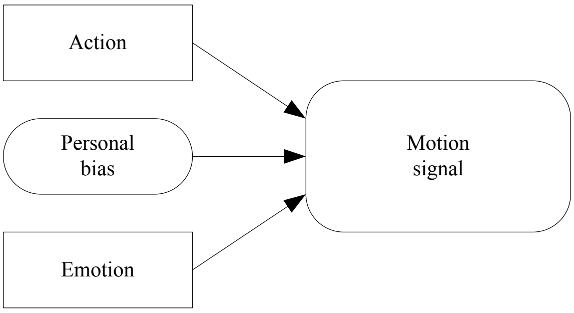

**Figure 1.1:** Factors affecting the final observable motion signal.

## 1.4 A holistic view of body movement

In order to meet these challenges, I will put forward a more holistic view than most previous work. Like previous studies, I will treat emotion recognition as a statistical pattern recognition problem. In order to detect different emotions I will consider both posture and movement qualities as predictive indicators. However, this dissertation will not treat emotion in isolation. In her PhD thesis, Helena Paterson recently discussed many of the issues relevant to this work from a psychological perspective and highlighted the significant three-way interactions between the individual, emotion and action [Pat02]. My work embraces these recent insights and analyses the impact of different categories of actions and the importance of individual movement styles for the task of emotion recognition (see Figure 1.1). The resulting holistic model is more robust to different action patterns and individual variations while capturing prevalent emotion cues.

I advocate the view that an emotion recognition system which is to function in a natural environment will require a certain "getting-to-know" period during which it builds a model for the person-specific movement parameters. Throughout this dissertation I will aim to make this personalisation as light-weight as possible. In other areas of intelligent human-machine interaction these kinds of personalisation are already standard procedure. For example, speaker adaptation in state-of-the-art speech recognition systems ensures that recognition performance is optimal for large-vocabulary and unstructured tasks.

## 1.5 Dissertation outline

At the heart of this dissertation lies the development of a pipeline which allows the inference of emotions from raw body motion signals. The pipeline incorporates various models of action patterns, the effect of individual movement style as well as emotional variations. The material will be presented as follows.

**Chapter 2** surveys the background material for this dissertation from a number of different disciplines. I will give evidence for bodily cues of emotional expression as described in psychological experiments and compare them to other modalities such as the face and voice. I will also cover the importance of body cues for the expression of emotions from a character animation perspective. The chapter then surveys previously developed systems which use these cues in different ways to detect emotions. Finally, I will present the corpus of body motions used for the majority of this dissertation.

**Chapter 3** serves as a practical introduction to a number of important issues relating to recognising emotions from body expressions. I will describe statistical features used to capture the emotional content from body posture and motion qualities. This chapter also explores the discriminability of different emotion classes and starts to explore the issue of personal differences in expression. In this chapter I am using a small corpus of archetypal expressions which I collected using a novel music-based environment.

**Chapter 4** moves away from archetypal expressions and describes methods for analysing complex everyday body movements. Initially, I focus on distinguishing different categories of actions such as walking, knocking and throwing. I then explore two approaches to understanding these kinds of complex actions in terms of basic motion primitives such as a series of arm raises or walking cycles.

**Chapter 5** combines the results from the motion analysis in Chapter 4 and the emotional features developed in Chapter 3 to build an emotion classification pipeline for isolated complex actions. In this chapter I also formally explore the issue of personal motion idiosyncrasies and describe a user adaptation framework akin to speaker adaptation in speech recognition systems.

**Chapter 6** explores how body motions change when they are performed in natural sequences rather than being recorded in isolation. I describe a method for segmenting sequences of complex actions and generalise the emotion recognition pipeline to deal with connected action sequences.

**Chapter 7** summarises the contributions made by this work and concludes with directions for future research based on my results.

## 1.6 Publications

Some of the results described in this dissertation have appeared in the following publications:

1. Daniel Bernhardt and Peter Robinson: Detecting emotions from connected action sequences. In *Proceedings of the International Visual Informatics Conference*, November 2009.

2. Daniel Bernhardt and Peter Robinson: Interactive control of music using emotional body expressions. In *CHI'08: Proceedings of the ACM Conference on Computer-Human Interaction*, pages 3117–3122. ACM, 2008.

3. Daniel Bernhardt and Peter Robinson: Detecting affect from non-stylised body motions. In ACII'07: *Proceedings of the Second International Conference on Affective Computing and Intelligent Interaction*, pages 59–70. Springer, 2007.

4. Daniel Bernhardt: Posture, gesture and motion quality: a multilateral approach to affect recognition from human body motion. In *ACII'07: Proceedings of the Doctoral Consortium at the Second International Conference on Affective Computing and Intelligent Interaction*, pages 49–56. 2007.

# Chapter 2

# Background

Although my research investigates primarily *computational* methods for the detection of emotions, it draws on many related fields for inspiration. Only very recently results on the human body have started to emerge out of work conducted by the affective computing community. The majority of relevant results are still to be found in psychological studies and models, systems developed for animating emotional body motion and recognition approaches developed for other modalities, primarily the face and voice. This chapter summarises the most important results from these fields. I conclude by introducing the data I will use throughout this dissertation to develop my computational models.

## 2.1 Modelling emotions

The term emotion has been used to denote a variety of concepts both in academic and everyday lay use alike. For example, it is commonly seen and used as a synonym for concepts such as *feeling*, *affect* or *mood* [SP01]. In this section I will give a succinct overview over the more recent consensus among psychologists about the conceptualisation of emotions. A suitable definition will set the scope for the computational models aimed at detecting emotions from body movements in the following chapters.

### 2.1.1 Emotions and other affective phenomena

Emotions are commonly seen as one of a number of different affective phenomena intrinsic in human experience such as mood, interpersonal stance, attitude and personality traits [Sch05a]. All of these affective phenomena have the power to bring about changes in the human physiology and psyche. They can be distinguished based on a number of dimensions including intensity, duration, as well as the degree of coordination between different modalities. Table 2.1 summarises these and additional dimensions. As opposed to some other affective phenomena, emotions are triggered by certain events and are of

**Table 2.1:** Affective phenomena and their properties (adapted from Scherer and Peper [SP01]). The extents along various dimensions range between 0 – +++ (no agreement – strong agreement).

| Affective phenomenon | Intensity | Duration | Synchronisation | Event focus | Extent of immediate behavioural effect |
|---|---|---|---|---|---|
| *Emotion*: relatively brief episode of synchronised response of all or most organismic subsystems in response to the evaluation of an external or internal event as being of major significance (*e.g.* angry, sad, joyful, fearful, ashamed, proud, elated, desperate) | ++ – +++ | + | +++ | +++ | +++ |
| *Mood*: diffuse affect state, most pronounced as change in subjective feeling, of low intensity but relatively long duration, often without apparent cause (*e.g.* cheerful, gloomy, irritable, listless, depressed, buoyant) | + – ++ | ++ | + | + | + |
| *Interpersonal stance*: affective stance taken toward another person in a specific interaction, coloring the interpersonal exchange in that situation (*e.g.* distant, cold, warm, supportive, contemptuous) | + – ++ | + – ++ | + | ++ | ++ |
| *Attitude*: relatively enduring, affectively colored beliefs, preferences, and predispositions towards objects or persons (*e.g.* liking, loving, hating, valuing, desiring) | 0 – ++ | ++ – +++ | 0 | 0 | + |
| *Personality traits*: emotionally laden, stable personality dispositions and behaviour tendencies, typical for a person (*e.g.* nervous, anxious, reckless, morose, hostile, envious, jealous) | 0 – + | +++ | 0 | 0 | + |

relatively short duration. Picard compares emotional responses to the attenuating sound level of a struck bell [Pic97]. The emotional responses from various bodily subsystems are highly synchronised and often cause a change in behaviour as the eliciting event is relevant to a person's well-being or goals. Mood, for example largely differs along most of those dimensions by generally having longer-term but milder effects on a person. While my work is largely targeting the detection of emotion, we will return to some of the characteristics of mood in Section 2.4.4.

Based on a clear understanding of the emotion concept, we can now turn to the question of how a computational system might detect different emotions from the human body or other modalities. For this we need to build a more detailed understanding of which aspects of a person's physical and mental state are affected by changes in emotion. The following definition by Scherer [Sch00] focuses on the effects that emotions have on several response *components*:

> Emotions are episodes of coordinated changes in several components (including at least neurophysiological activation, motor expression, and subjective feeling but possibly also action tendencies and cognitive processes) in response to external or internal events of major significance to the organism.

This definition is the essence of the Component Process Model of emotions and lists a number of components influenced by emotions. Each serves a different function [SP01]:

- peripheral physiological activation component: system regulation
- motor expression component: communication of reaction and behavioral intention
- subjective feeling component: monitoring of internal state and organism-environment interaction
- action tendencies (motivational) component: preparation and direction of actions
- cognitive component: evaluation of objects and events.

The first three components are commonly referred to as the *emotional response triad* [SP01]. Both the motor expressions and to some extent the peripheral physiological components can yield information for inferring emotions from directly visible components. Inferring emotions from these externally visible cues has been termed the *social-perceptual* approach and has been demonstrated to be a fundamental component used by humans to understand emotions in other human beings [TFS00]. This data-driven model of emotion recognition lends itself well to a computational implementation. In this dissertation I will focus on the motor expressions produced by the human body in particular to infer emotional states.

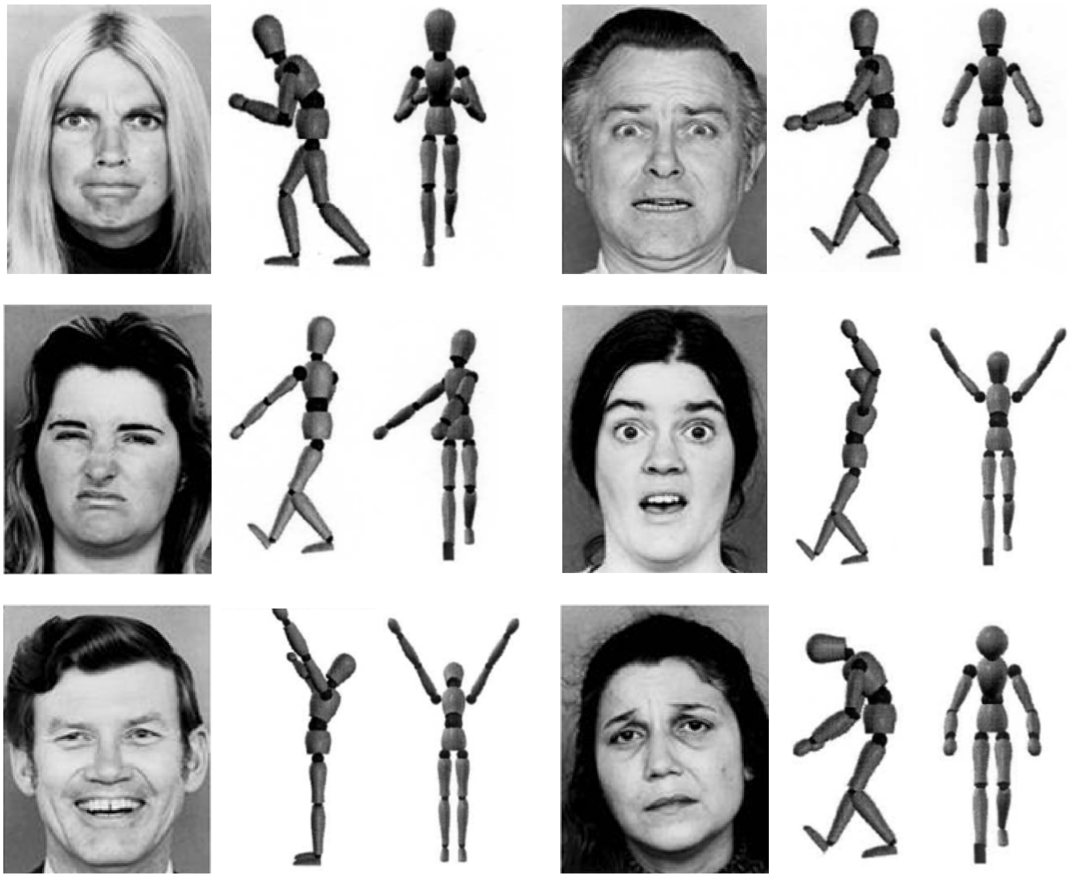

**Figure 2.1:** Ekman's basic emotions (angry, afraid, disgusted, surprised, happy, sad) as expressed in the face (from [EF76]) and through body posture (from [Cou04]).

### 2.1.2 Models of differentiation

From everyday use people are very familiar with differentiating between emotions using labels such as *cheerful*, *tense* or *angry*. Nevertheless, emotions are inherently complex phenomena, continuously developing and changing over time. It is therefore not surprising to find a number of different approaches to structuring and differentiating between different emotions in the psychology literature. Often different models are particularly suitable for describing a subset of the components of emotion (*e.g.* motor expression). While the model of discrete emotions is well-suited to address the motor and behavioural effects of emotions, dimensional and meaning-based methods have been argued to be particularly good for capturing subjective feelings. These three models were outlined by Scherer [Sch00] and I describe them briefly below.

#### Discrete models

The idea of discrete emotions stems from the neuropsychological idea that certain emotional reactions are innate in the form of neural circuits and neuromotor programs [Sch00].

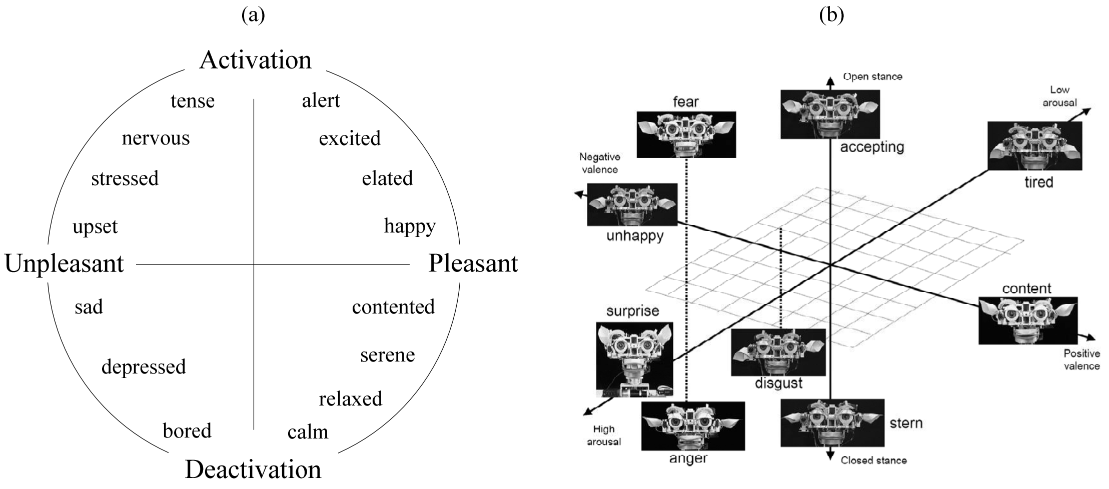

**Figure 2.2:** Dimensional emotion models proposed by (a) Russell [PRP05] and (b) Breazeal [Bre04].

A discrete view of emotions was first postulated by Darwin [Dar72] and has more recently been adopted by Ekman and Izard who argue that a number of basic emotions are expressed through corresponding prototypical facial expressions [Ekm92, Iza94]. The number of basic emotions proposed varies from author to author but is generally assumed not to exceed 15. Figure 2.1 shows six basic emotions expressed through facial expressions and body postures. One problem with discrete emotions is their inability to capture the large variety of distinct emotional experiences that we would intuitively postulate. It is therefore necessary to allow a blending of different basic emotions into more complex emotion mixtures.

### Dimensional models

The notion of emotion mixtures, especially within the realms of subjective feelings, leads to the idea of continuous emotion spaces. In these models emotions are usually placed within a low-dimensional space. The circumplex model proposed by Russell uses two orthogonal dimensions: valence (pleasantness) and arousal (activation) [PRP05]. The standard emotions as identified by Russell lie on a circle in this space (see Figure 2.2(a)). Although valence and arousal are the most commonly used dimensions for modelling emotions, other authors have used additional dimensions such as attention [Wun05], stance [Bre04] or action tendency [KSBB05] (see Figure 2.2(b)). Note how the latter additional dimensions help to capture aspects of the emotional process that go beyond the subjective feeling component.

**Table 2.2:** Body motions and postures associated with certain emotions as noted by Darwin [Dar72] (from Wallbott [Wal98]).

| | |
|---|---|
| Joy | Various purposeless movements, jumping, dancing for joy, clapping of hands, stamping, while laughing head nod to and fro, body held erect and head upright |
| Sadness | Motionless, passive, head hangs on contracted chest |
| Shame | Turning away the whole body, more especially the face, avert, bend down, awkward, nervous movements |
| Fear/terror/ horror | Head sinks between shoulders, motionless or crouches down, arms thrown wildly over head, arms violently protruded as if to push away |
| Anger/rage | Whole body trembles, intend to push or strike violently away, gestures become purposeless or frantic, pacing up and down, shaking fist, head erect, chest well expanded, feet planted firmly on the ground, one or both elbows squared or arms rigidly suspended by the sides |
| Disgust | Gestures as if to push away or to guard oneself, spitting, arms pressed close to the sides, shoulders raised as when horror is experienced |
| Contempt | Turning away of the whole body, snapping one's fingers |

### Meaning oriented models

These models are very much structured around linguistic concepts of emotions and their understanding within a society. Often emotional concepts are organised in hierarchies or taxonomies based on generality or semantic significance. For example, the OCC model developed by Ortony, Clore and Collins [OCC75] proposes a structure for different emotion concepts based on cognitive eliciting conditions involving events, objects and other agents. Another approach was taken by Scherer who built a comprehensive taxonomy of emotion terms and categories [Sch05a]. Meaning oriented models are usually very intuitive to understand for people otherwise unfamiliar with emotion theory. They are therefore often used for labelling emotion displays by untrained labellers. Because of the semantic structure and intriguing intuitive appeal, models like the OCC have often been proposed for implementations of real-world artificial intelligence-based systems.

## 2.2 Emotional body expressions

The first rigorous evidence for the expression of emotions through the body are found in Darwin's seminal work *The expression of the emotions in man and animals* [Dar72].

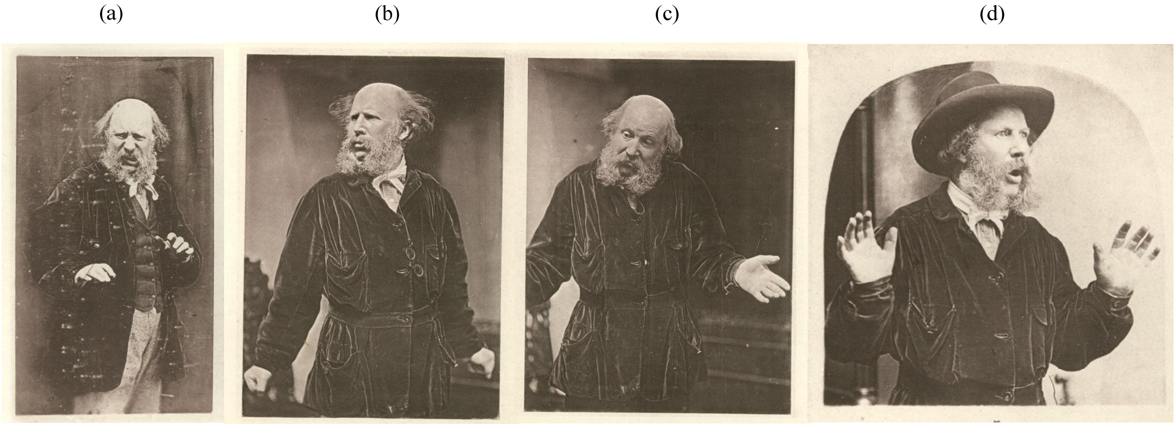

**Figure 2.3:** Archetypal expressions associated with (a) disgust, (b) anger, (c) helplessness and (d) surprise collected by Darwin [Dar72].

In it he lists behaviours specific to certain emotional categories, many of which are now regarded as basic emotions. Table 2.2 lists some examples. Darwin also collected a number of photographs depicting archetypal gestures associated with some emotions (see Figure 2.3). In the century that followed many researchers compiled similar lists of stereotypical expressions, *e.g.* in [GP06, BRI$^+$05, Wal98]. Other authors have argued for similarly distinct patterns of emotional body expressions along a number of more abstract dimensions, using terms such as *speed*, *force*, *energy*, *expansiveness* or *directness* [Mei89, SW90, MKZA99].

While Darwin seemed convinced that different emotions give rise to specific body movements and postures, some more recent psychological work has been more nuanced. Some psychologists even suggested that body expressions can be merely an indication of emotion intensity rather than an emotion itself [EF74]. I will come back to this view in Section 2.2.1. Other authors have argued that the body carries emotional information in much weaker form than originally claimed by Darwin and than may be accredited to other modalities like the face. An important attribute of the body is that it is primarily used to perform manipulatory actions and to facilitate locomotion. Emotions are hence often only noted as *secondary signatures* on top of those more fundamental actions [RCB98, ABC96]. Furthermore, several authors have mentioned that body posture and motion are susceptible to individual idiosyncrasies [Cou04, Mei89, Wal98, Ste92]. For example, Wallbott found that in one of his experiments the amount of variation observed in the data due to emotions was nearly matched by that due to individual differences in expressions [Wal98]. This brief overview hence suggests that there is significant emotion-related information in body expressions. However, there are other effects like underlying actions and individual idiosyncrasies which also have a large impact on body expressions.

In the next two sections I will expand on these themes while discussing emotional body expressions as treated within the humanities and animation communities.

### 2.2.1 Emotional body expressions and the humanities

The study of the role of different modalities in the communication (encoding and decoding) of emotions has a long tradition in psychology. It has been frequently noted, however, that conclusive results about the particular role of the body are missing due to a lack of studies [VdSRdG07, ADGY04]. Given the striking saliency of the human face as a communication channel of emotions, it is maybe not surprising that many early studies focused on the face. They often concluded that facial expressions account for the large majority of emotional cues. Work by Ekman lead to the influential *quantity but not quality* hypothesis [Wal98, Ekm64, EF74]. He postulated that body movements merely provide information of the intensity of an emotion. He argued that body motions could not give any indication about which emotion gave rise to it. This notion was in clear conflict with Darwin's original observations and many authors have since found convincing evidence for a systematic encoding of emotional "quantity" information in both body posture and gestures.

#### Emotions in body posture

Posture is usually defined as the *static* configuration of body parts, including head pose, as well as arm, leg and trunk configuration. Starting with Darwin (see Table 2.2 and Figure 2.3), there have been more or less detailed descriptions of specific postures adopted in connection with different emotions [BC01, Wal98, Mei89, Bul87]. More recent studies by Coulson and De Silva *et al.* explored the effect of different posture features more systematically [Cou04, DSBB04]. Both studies independently confirmed that features like head and chest bend, elbow bend and hand-shoulder distances are statistically significant indicators of emotions.

#### Emotions in body movement

Similarly systematic investigations have revealed the significance of *dynamic* motion cues for the communication of emotions. In fact, Atkinson *et al.* recently showed that static form and dynamic motion cues provide distinct contributions to the communication of emotion through the body [ADGY04]. In other studies authors have demonstrated the importance of dynamic motion cues for the encoding and decoding of emotions in different contexts including interpersonal dialogues [Bul87, CBF$^+$05], dance [CLV03, DTLM96], infant expressions [CSM93] and everyday actions [PPBS01]. Most of these studies stress the importance of certain movement qualities such as *jerkiness*, *energy* or *speed* as discriminating attributes. In 1990 these results had been predicted by Scherer and Wallbott who argued that emotions in the body are likely to be detectable through changes in the speed, rhythm and fluidity of movements [SW90]. This is mainly due to the fact that the human body has evolved to primarily support life-preserving functions. The overwhelming evidence therefore points to the importance of dynamic qualities when it comes to the encoding of emotions through the body.

**Structure of body expressions**

Social scientists have also been among the first to study human movement in a principled way. In order to facilitate their studies, researchers in the humanities have long sought to devise notation schemes which let them break down complex motion sequences into fundamental building blocks (primitives). The goal is usually to represent motions as objectively as possible and leave social or emotional interpretations to a later stage. Bull and Birdwhistell devised complex and complete systems for describing body posture and motion in anthropological and psychological observations [Bir70, Bul87]. The goal of Bull's *Body Movement Scoring System* [Bul87] was explicitly to develop a system as complete as the *Facial Action Coding System* [EF78] — now a defacto standard for coding facial expressions in psychology and more recently for the computational analysis of facial expressions. Birdwhistell called his motion primitives *kinemes* and argued for fundamental similarities between bodily and natural language. The similarity between body and spoken language will be a returning theme throughout this dissertation. Other coding systems have tried to explicitly incorporate the crucially dynamic aspects of body motion. For example, *Labanotation* [Lab75] was originally developed to record dance steps but was later extended with additional dimensions describing the spatial, dynamic and rhythmic aspects. Movements can be described along a number of *Effort* and *Shape* dimensions with attributes such as *indirect*, *light*, *sudden* or *spreading* [Hut77].

### 2.2.2 Emotional body animations

A very different field of study which has focused on emotional body motion is character animation. Since the very early days of animation there has been a strong focus on portraying characters as emotionally convincing in order to catch the imagination and empathy of audiences. Though less formal, there is an immense amount of insight that can be gained from the animation community. This kind of knowledge is captured well in a quote by John Lasseter, director and producer of many pioneering films by Pixar and Disney. He gives some insights from an early animation piece involving two extremely expressive characters in the form of desk lights (*c.f.* Figure 2.4) [Las87]:

> One character would not do a particular action the same way in two different emotional states. An example of this, in Luxo Jr., is the action of Jr. hopping. When he is chasing the ball, he is very excited, happy, all his thoughts on the ball. His hops are fast, his head up looking at the ball, with very little time

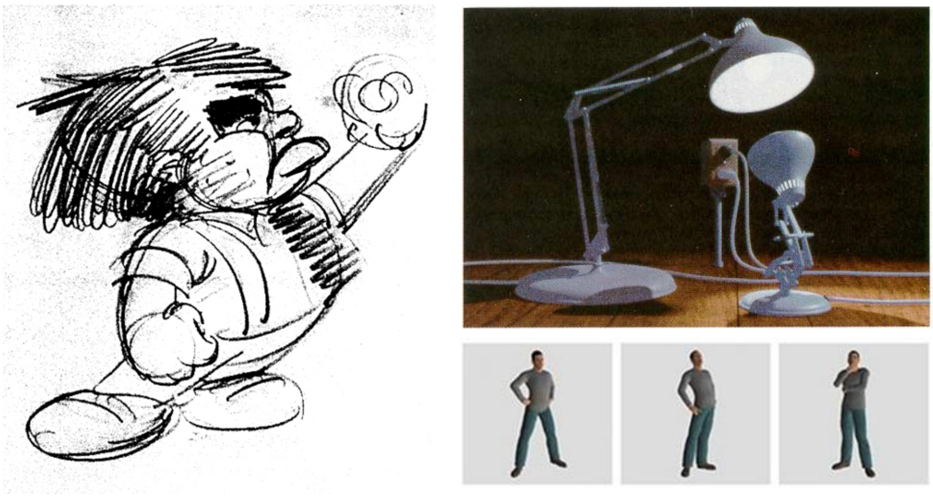

**Figure 2.4:** Examples of emotional body expressions in animation. Left: A sketch from *The illusion of life: Disney Animation* [JT95], top right: Luxo Jr. and Luxo Sr. in a Pixar animation from 1987 [Las87], bottom right: part of the cover art of the book *Character emotion in 2D and 3D animation* by Les Pardew [Par07]

> on the ground between hops because he can't wait to get to the ball. After he pops the ball, however, his hop changes drastically, reflecting his sadness that the object of all of his thoughts and energy just a moment ago is now dead. As he hops off, each hop is slower, with much more time on the ground between hops, his head down. [. . . ]
>
> No two characters would do the same action in the same way. For example, in Luxo Jr., both Dad and Jr. bat the ball with their heads. Yet Dad, who is larger and older, leans over the ball and uses only his shade to bat it. Jr., however, who is smaller, younger, and full of excited energy, whacks the ball with his shade, putting his whole body into it.

In order to make character behaviour believable, Lasseter highlights many of the issues already familiar to us from the insights of detailed psychological studies:

- influence of different emotions on the appearance of actions
- significance of *dynamic* motion qualities as well as body *posture*
- impact of individual idiosyncrasies on the appearance of actions.

Many of these ideas have been the subject of academic research in computer animation of film characters and more recently avatars. Figure 2.4 shows some examples of emotional body expressions from traditional and modern animation. Systems developed in recent research frequently aim to provide better means for animators to control the expressions of their characters. Many of the approaches are based on the notion of affective transforms applied to an underlying basic or neutral animation [VGS$^+$06]. One of the earliest approaches developed by Amaya *et al.* was based on two such transforms affecting the speed and spatial amplitude parameters of motions [ABC96]. They also noted that different emotional styles might be encoded through different phase relationships between various body joints. A more complex set of transformations has been described by Polichroniadis who proposes additional parameters such as constant posture offsets, body swing and shake [Pol01].

While Polichroniadis' parameters are motivated by diverse sources, Costa *et al.* used Laban's Effort/Shape framework discussed in Section 2.2.1. Their EMOTE system uses Effort and Shape parameters as a basis for interpolation between traditional animation key frames [CCZB00]. In a related system they demonstrate how the EMOTE parameterisation can be used to drive animations through natural language instructions like *walk proudly* [ZBC00]. In this case adverbs like *slowly*, *proudly* or *happily* map to specific Effort and Shape parameters affecting the quality of body motions. This verb/adverb analogy was also used by Rose *et al.* who treated adverbs like *happily*, *sadly*, *uphill* or *downhill* as parameters in their own right [RCB98]. By interpolating between actions of the same type (verbs), they derive new motions with different stylistic qualities.

Finally, Liu *et al.* describe a set of parameters which capture the biophysical elements of style, for example tendon elasticity and relative preferences of muscle uses at each joint [LHP05]. They describe a way to learn and represent emotional styles as certain combinations of these parameters. A similar abstract notion of style is employed by Brand and Hertzman [BH00]. They define *Style Machines* as stochastic models which can generate new motion sequences based on a style parameter and a probabilistic description of the underlying action.

#### Gestures

A relatively independent trait of body movements often discussed by the animation community are body gestures. They are seen as fundamental to creating believable characters or avatars [VGS$^+$06]. A gesture can be defined as a *visible action when it is used as an utterance or as part of an utterance* in the context of interpersonal communication [Cas08, Ken04]. Gestures are often regarded as intimately linked to verbal language [VGS$^+$06, mcn92] but are sometimes also more generally regarded as an action which manifests deliberate expressiveness [Ken04]. Douglas-Cowie *et al.* reported that in conversational contexts only 20% of these kinds of deliberate actions are informative of emotions [DCDM$^+$05]. They found that in real-world scenarios explicit body gestures were the least informative out of a number of cues from multiple modalities including speech, eyes, mouth, head and brows. Because affective gestures like arm raises, hand claps and the like are relatively easy to analyse with well-established gesture recognition algorithms, several researchers have in the past limited their modelling of emotional body signals to affective gestures [BRI+05, DSOMM06, CKC08]. In this dissertation I will focus on the less explicitly communicated emotions in everyday actions as summarised in the next section.

### 2.2.3 Everyday actions *vs.* emotion archetypes

We have seen in previous sections that body motions can be complex and that the body can perform different functions in different contexts. For this dissertation I will distinguish between two fundamental types of body motions: everyday actions and inherently affective gestures. I regard as affective or expressive gesture *any kind of body motion which is motivated by the deliberate expression of affective content.* Everyday or complex actions on the other hand are *motions which are motivated by goals other than communicating affect.*

Examples for everyday actions are activities such as walking, sitting down or certain kinds of gestures such as pointing movements or iconic gestures describing the spatial extent of objects. Expressive gestures, on the other hand constitute emotional body movements such as raising one's arms in joy or parts of expressive dance performances. One particularly pertinent class of affective gestures are *emotion archetypes* which are the kinds of exemplary motions which come to peoples' minds when they hear a certain emotion word [CS05]. Note that, while motions arising from everyday actions are not *intended* to convey emotions and are indeed usually carried out to achieve some other goal, *qualities* of those motions can convey certain emotional information [Cas08]. Indeed, many of the animation techniques described above target transformations of everyday actions in order to convey a certain emotional content. Lasseter's description of a happy *vs.* sad Luxo Jr. is a perfect example of how an action such as walking can have clear emotional qualities.

## 2.3 Detecting emotions from face, speech and body

Much research in the past has been devoted to the study of automatic emotion recognition from the face and human speech (for a comprehensive overview see [CDCT+02]). Comparatively little work has still been done on emotion detection from the human body — one of the major motivations for the work described in this dissertation. Given such a large volume of work on other modalities, it is tempting to assume that the techniques developed for the face and speech can be readily applied to motion signals. There are, however, fundamental differences which need to be addressed. In particular, despite the seeming similarities between facial and bodily expressions of emotions (after all the face is just a specialised subsystem of the human body), there is evidence for very different emotional patterns.

Compared to the human body, the face has been argued to be relatively free of functional constraints — its intricate muscles are specialised to produce subtle facial expressions [SW90]. Notable exceptions include activities such as speaking, chewing and squinting which can significantly impact facial movements. Although Pantic and Rothkrantz pointed out the significant impact of speech articulation on the detection and interpretation of facial expressions [PR02], much of the facial expression recognition research has focused on isolated expressions without such complicating actions. By the definition in Section 2.2.3, these expressions amount to affective gestures — the motions themselves contain immediate emotional meaning. This is most notably the case for facial expressions of basic emotions where early rule-based systems linked expressions to discrete emotions (*e.g.* smiling implies happiness). These systems ignore the presence of non-expressive speech motions often found in natural scenarios. Nevertheless, because speech is usually the only source of non-expressive signals in the face, an analysis on the basis of affective expressions alone is certainly justified and has produced some convincing results (see Section 2.3.1 for details).

Emotion detection from auditory speech signals, on the other hand, is very different. In speech large parts of the signal are shaped by the "non-expressive" verbalisation of words and phrases. Only certain sonic qualities are usually indicative of emotional meaning. Early research by Cowie and Douglas-Cowie identified a number of features in the *prosodic* domain which they showed to be linked to emotions [CDC96]. Over the following decade authors proposed many new features capturing speech qualities. We saw in previous sections that there is convincing evidence that emotions from the body are also likely to be encoded by *qualities* rather than explicit, affective motions and gestures.

Relatively few authors have discussed the role of motion *qualities* with respect to facial expressions. Studies by Pantic, Valstar, Cohn and others which have considered qualities [CS04, SKA$^+$06, VPAC06, VP07, VGP07] usually focus on the detection of spontaneous *vs.* posed expressions such as smiles and brow actions rather than the detection of emotions. Some exceptions include the work by Zhang and Ji who use dynamical models to capture durations of facial expressions and their effect on emotional interpretations [ZJ05]. Some other works exist which discuss dynamics of facial actions [PP06], how to use the dynamics of facial expressions for the detection of identity [LBF$^+$04], pain [LBL07] and the intensity of smiles [WLF$^+$09]. This dissertation focuses on the use of *motion qualities* rather than the detection of explicit gestures for emotion recognition — this makes it in some ways similar to the analysis of emotion in speech and is related to some of the recent works discussing emotion analysis from facial dynamics. Even though the research groups of Pantic and Cohn have in the past stressed the importance of facial dynamics for emotion analysis [PP06, Coh06], this is still a largely unexplored area.

It is worth noting at this point that other sources of emotional information have been investigated in the recent past. Those usually fall under the *physiological activation* component of the emotional response triad (*c.f.* Section 2.1.1). Popular sources of information are electromyogram, blood volume pressure, skin conductivity and respiration [PVH01, GPP08]. More recently researchers have started to analyse brain activity in order to infer emotional states [HDR08]. Because these methods are not concerned with the motor expression of emotions, I will largely focus on how emotion is expressed in the face and voice to draw inspiration for my body-based approach.

### 2.3.1 Automatic emotion recognition from facial expressions

Despite the underlying difference in expression of emotions, studies on facial expression recognition are relevant for this research. Fundamentally, both problems are commonly solved using pattern recognition techniques. The comprehensive literature on facial expression analysis provides a good background on the use of different machine learning techniques such as Support Vector Machines [BLV$^+$08, MK03], Dynamic Time Warping [RD08], Hidden Markov Models [ZTPH08, PBL02, KR05] or Dynamic Bayesian Networks [KR05] for emotion recognition. I will touch on all of these techniques throughout this dissertation.

Facial expression recognition has revealed a great deal about the importance of *dynamic* information when looking for subtle emotional cues. Early systems tackling a small number of basic emotions found that simple static models worked well in distinguishing between them [PR00]. As the systems aimed to analyse more subtle facial expressions, it emerged that dynamic information is very important for detecting emotions [Kal05, PP06, Coh06, RD08]. This work naturally lead to the analysis of more natural sequences of facial expressions rather than the artificial isolated displays often found in early databases. It is not surprising, however, that Pardás *et al.* report that detecting emotions from natural sequences is generally harder than when applied to isolated expressions [PBL02]. They also confirmed that functional movements produced by speech represent a major complication for facial expression recognition.

Other relevant areas include the development of a universal coding system for facial actions — the Facial Action Coding System [EF78]. Many similar systems have been proposed for the body in the past, but because of their complexity none has so far been widely adopted as a standard. Many facial analysis systems have relied on Ekman's basic action units as primitive building blocks and several authors have worked on the automatic extraction of action units from facial motion sequences [TKC02, VP06, BLF$^+$06, ZDlTCZ09, TLJ07]. I will investigate to what extent the analysis of body movements can benefit from similarly basic motion primitives.

Finally, studies on facial expressions have tackled issues concerning the impact of individual differences in expression on training and recognition performance. While systems trained and tested on the same subjects tend to perform very well, holding out certain subjects' data for testing produces worse results (*e.g.* see [CSG$^+$03]). This effect can be very significant. It means that recognition rates obtained for the former setup are not very good indicators for the *generalisation* performance of a system to previously unseen subjects or users — clearly a desirable property of a real-world system. Similar effects were found when using different *corpora* for testing and training [LBF$^+$04]. I am aiming to report reliable figures by employing a cross-validation technique which reliably estimates the recognition performance for unseen subjects.

### 2.3.2 Automatic emotion recognition from vocal expressions

The issue of subject differences has also been the topic of several voice-based emotion recognition systems. Speaker adaptation is now an established concept in the field of speech recognition and has been finding its way into the emotion recognition community. Vogt and André, for example, demonstrated that augmenting emotion recognition with a gender detection stage leads to better performance [VA06]. While the difference in gender may not be quite as pronounced for body motions as it is in speech [Mei89], the general ideas of adapting the recognition framework to user differences is useful. Other authors have stressed the more general importance of speaker normalisation for emotion recognition [SAE07, VSWR07, KCHL03]. All of these approaches are targeted at removing signal bias introduced by the individuality of speakers. In another study, Forbes-Riley and Litman investigated the role of additional context information such as gender and higher-level dialogue features [FRL04]. All of these techniques are immediately relevant for my work on human body expressions.

As it is mainly the *quality* of both speech and body motions which is important for the expression of emotions, it is also worth considering the features commonly used to capture emotional differences in speech. The most commonly used features are acoustic or prosodic in nature and hence capture the *quality* of speech:

- pitch (fundamental frequency)
- intensity (vocal energy)
- speech rate/duration
- pitch contour
- phonetic features.

Among those, pitch and energy have been argued to be the most informative for affect recognition [KCHL03]. Although the domains are very different, speech features such as *energy*, *speech rate*, *velocity* and *acceleration* of pitch [KCHL03, SS08] are surprisingly similar to the qualities which psychologists and animators have argued to be indicative of emotion in body movements. Like in the field of facial affect recognition, early works on emotion recognition from the voice often postulated specific acoustic correlations for emotion expressions. Those had among others been summarised by Cowie *et al.* [CDCT+02].

Basic emotions recorded under laboratory conditions are relatively easy to distinguish using these kinds of correlates. More natural speech as it arises in call centres, meetings and interviews, however, makes emotion recognition far more challenging. In those scenarios prosodic features are affect by emotion as well as a range of other linguistic factors like accents and intonation [Moz02]. this problem is reminiscent of emotion recognition from body motions in many every-day scenarios. The dynamics of one's body movements are often influenced significantly by the particular action one is carrying out – in the same way that speech qualities are significantly affected by functional linguistic factors as well as emotional factors.

As one possible solution to this problem the speech community has proposed the use of different features. Several authors have shown that higher-level lexical or linguistic features like occurrences of word repetitions, corrections or different uses of parts-of-speech and words can be very valuable for distinguishing emotions in more natural scenarios [LR04, BFH+03].

Finally, the pattern recognition techniques that have been employed to distinguish between different emotional patterns in speech have ranged from simple methods like k-Nearest-Neighbour, template matching and heuristic rules to state-of-the-art classifiers such as boosted decision trees or Support Vector Machines [ZPRH07].

### 2.3.3 Automatic emotion recognition from body expressions

Having mentioned the considerable progress made in emotion recognition from the face and voice, I now turn to the previous efforts in emotion recognition from the body. As in other fields of study, until recently the body had been largely ignored by the community as a source of affective information. Furthermore, the vast majority of efforts have to date focused on detecting emotions from *emotional gestures* and motions [KKVB+05, KSBB05, CLV03, BRI+05, GP07]. To gather representative data, subjects are in many cases instructed to act out certain emotions, which inevitably leads to explicitly affective gestures and archetypes.

One notable exception is some of the recent work by Castellano *et al.* who also consider some non-expressive scenarios such as a piano performance [CMC+08] or a simple non-propositional arm gesture [CVC07]. In both cases the data was acquired using a conventional DV camera. As the motions themselves carry little or no emotional meaning in these scenarios, they use *movement qualities* to distinguish emotions. For that purpose the InfoMus Lab at the University of Genova developed the EyesWeb Expressive Gesture Processing Library [CMV04]. It is designed to facilitate the real-time analysis of expressive gesture in full-body human movement based on computer vision algorithms.

Features used by Castellano *et al.* include the quantity of motion and contraction index of the upper body as well as velocity, acceleration and fluidity of limbs and head. When distinguishing between the four emotion classes *anger*, *joy*, *pleasure* and *sadness* their Bayesian Network-based classifier achieves a correct recognition rate of 61% in the gesture scenario. This seems considerably lower than typically quoted rates achievable from the face and voice. For the piano scenario their features did not show any clear patterns and a classification was not successful. This goes to show that detecting emotions in these kinds of scenarios is not easy. They hypothesise that the constraints inflicted by the characteristics of the task limit the expressiveness of the subject, especially when compared to scenarios which focus on emotional expressiveness like dance [Cas08, CLV03]. A further limitation is the low-fidelity representation of body motion and configurations owing to the single camera view of the subjects and only basic processing inside EyesWeb.

Another popular everyday motion context is human gait. Using force plates, Janssen *et al.* recorded the gait patterns of a number of subjects in four emotional styles: *neutral*, *joyful*, *angry*, *sad* [JSL$^+$08]. They found that the differences between individuals far outweighed the differences due to emotions. In fact, they note that they found inter-subject emotion recognition impossible. However, they noticed very clear patterns within subjects. They did not, however, attempt to normalise the data between subjects. Instead, they trained Neural Networks on the patterns from individual subjects and found an average recognition rate of 84%.

Progress on explicitly expressive movements has in the meantime been a bigger focus in the community. Early results were reported by Camurri *et al.* who looked at recognising emotions from expressive dance [CLV03]. They describe a processing stack based on the EyesWeb infrastructure to extract dynamic cues from the dancers' movements. They do not report formal classification results, but show that there are statistically significant effects between certain emotions and key qualities including time duration, contraction index, quantity of motion and fluency.

More recently Kapur *et al.* reported recognition results based on archetypal body motions [KKVB$^+$05]. They instructed subjects to freely enact four emotional states and recorded the motions using a 3D motion capture system (*c.f.* Section 2.4.3). They then trained and tested a number of classifiers on the data. They found that Neural Networks and Support Vector Machines were able to achieve a correct recognition rate of 84% while humans were able to classify the clips correctly 93% of the time. This illustrates that both humans and machines find recognising the emotional patterns in archetypal motions much easier than in more subtle everyday scenarios. In similar experiments by Gunes and Piccardi subjects were seated and recorded by a standard DV camera. Their expressive motions were distinguishable by a Bayesian Network-based classifier at a rate of 90% [GP05]. Further works by Gunes and Piccardi employ similar methodologies, together with a focus on fusing body signals with other channels [GP06, GP07, GPP08, GP08]. Using the same data Shan *et al.* presented a vision-based system which analyses the spatio-temporal features extracted from interest points in videos. Their SVM classifiers were able to distinguish between the 7 classes at an average rate of 0.75 using body features alone.

Another class of systems are based on the explicit modelling and detection of affective gestures such as hand clapping, arm raises or hands-on-face gestures. Three examples for these kinds of systems come from Balomenos *et al.* [BRI$^+$05], De Silva *et al.* [DSOMM06] and Castellano *et al.* [CKC08]. These kinds of systems usually make the assumption that certain gestures are indicative of certain emotions. The systems by Balomenos and de Silva model upper body motions using Hidden Markov Models and formulate the emotion recognition problem as a gesture recognition problem. De Silva *et al.* go one step further and also try to predict the intensity of displayed emotions based on factors like the detected mix of emotions and the environmental context, which in their case is the state of a computer game which the subject is playing. In Castellano *et al.*'s system emotions are not being modelled explicitly but are detected through a number of dynamic features such as quantity of motion, fluidity *etc.* Using a Bayesian Network they were able to distinguish between the gestures of eight emotion classes at an average rate of 67%.

Finally, Picard *et al.* and Kleinsmith *et al.* concentrated on posture cues in isolation. Picard *et al.* were interested in predicting a learner's interest level from his/her body posture [MP03, KBP07]. They recorded data using a pressure-sensitive chair in an authentic scenario. In subsequent experiments they found that shifts in the pressure patterns could predict the three states *high interest*, *low interest* and *neutral* with an accuracy of 76%. While Picard *et al.* were collecting data in an everyday context, Kleinsmith *et al.* used archetypal body expressions recorded in an artificial laboratory context and extracted the apex of the emotional expression as a posture of 3D body joints [DSBB04]. In [KSBB05] they find a mapping between a number of static posture features and the three emotion dimensions of arousal, valence and action tendency. They report that while the arousal dimension is very well represented in posture features, valence and especially action tendency introduce a large error. They attribute these problems to the *dynamic* cues which are missing in their static posture model.

## 2.4 The data used in this dissertation

We have seen that detecting emotions from explicitly expressive and archetypal body motions has yielded relatively good results in the past. For the majority of this thesis (Chapters 5 – 7) I will therefore focus on the largely unexplored and more challenging category of *everyday actions*. In the rest of this section I will present the source and nature of my data and describe why its use can facilitate some significant progress in the field beyond the insights acquired elsewhere.

I also decided to include a study of some expressive and archetypal motions in Chapter 3.

As my work on everyday actions covers largely new terrain, Chapter 3 serves as a reference to previously conducted work and gives a good overview over the general principles involved in the feature extraction process. Chapter 3 is largely self-contained and I will not cover the archetypal data here.

### 2.4.1 The question of ecological validity

In their recent article *Beyond emotion archetypes: Databases for emotion modelling using neural networks* Cowie *et al*. raise the important question of ecological validity of emotion databases used for the kinds of emotion recognition experiments described in this dissertation [CS05]. They remark that assembling databases for these kinds of experiments is very challenging due to various reasons ranging from practical and ethical to intellectual. Nevertheless, they criticise many of the traditional approaches because they do not focus on *emotion in action and interaction.* Cowie *et al*. warn that

> [. . . ] research should not slip into assuming that the empirical data needed to understand a topic equate to a collection of cases that reflect the *archetypal* images we associate with it.

They use the term archetypal to refer to instances such as Ekman's facial images of basic emotions or previously described affective gestures which we would expect to see very rarely in actual human interactions. They therefore demand that modern affective computing research should focus on *emotion as it appears woven through everyday activities.* They note that in experiments

> [. . . ] most of the variables which are affected by emotion are also affected, as much or more, by commonplace activities such as moving, speaking, or even thinking. As a result, their value as discriminators diminishes rapidly as one moves towards free situations where people are likely to be moving, speaking, or thinking. [. . . ]

In this dissertation I intend not to evade the complexity of real everyday situations. The database I am using captures many of the intricate interactions between actions, emotions and individuality. Dealing with this kind of data is clearly a challenge, but it is necessary if we want to move beyond archetypal displays of emotion and towards more ecologically valid situations.

Picard *et al*. published a useful list of heuristic dimensions [PVH01] along which the ecological validity of recorded data can be assessed. The dimensions are outlined below. Note that Picard *et al*. originally developed the dimensions in connection with the recording of physiological signals and implicitly assumed that data would be recorded in a realistic task environment.

- *Subject-elicited vs. event-elicited*: Does the subject purposefully elicit emotion or is it elicited by a stimulus or situation outside the subject's efforts?
- *Lab setting vs. real-world*: Is the subject in a lab or in a special room that is not their usual environment?
- *Expression vs. feeling*: Is the emphasis on external expression or on internal feeling?
- *Open-recording vs. hidden-recording*: Does the subject know that anything is being recorded?
- *Emotion-purpose vs. other-purpose*: Does the subject know that the experiment is about emotion?

While idealised realistic scenarios fall on one end of the spectrum (event-elicited, real-world, feeling-oriented, hidden-recording, other-purpose), Picard *et al.* acknowledge that these scenarios are virtually impossible to achieve in practice. A large number of motion databases populate the opposite end of the spectrum. Some notable exceptions come from the recent HUMAINE database [CCS$^+$07] which attempts to represent emotion as it appears in a variety of natural contexts. Although large parts of it score highly on many of the dimensions above (*e.g.* event-elicited, real-world, other-purpose), it is a fundamentally heterogeneous collection of video data which has not been annotated with emotion labels. As such, it will undoubtedly play an important role in future research but was unsuitable for this dissertation.

### 2.4.2 The Glasgow corpus

For the majority of the work described in this dissertation I used a corpus of body motions recorded by Frank Pollick *et al.* at the Psychology Department, University of Glasgow [MPP06]. The database was explicitly created to capture and represent the wide range of personal properties, including identity, gender and emotion, that are encoded in a person's movement. The database has a number of distinct merits:

- captures complex interactions between emotion and individual differences
- covers four distinct emotion classes: *neutral*, *happy*, *angry*, *sad*
- all samples are labelled by emotion, providing a ground truth for learning and evaluation
- covers four everyday action categories: *Knocking*, *Throwing*, *Lifting*, *Walking*
- actions recorded both in isolation and in natural sequences (see Figures 2.5 & 2.6): duration varies between 2–4 seconds for isolated actions; 18 seconds on average for action sequences

- contains a large number of actors and repetitions: (15 male + 15 female) × 10 repetitions per action and emotion
- very high quality of recording (see Section 2.4.3)
- freely available for research use
- emotion recognition results by humans exist for parts of the database.

These advantages and the fact that the database has not been used for computational analysis in the past make it ideal for the purpose of this dissertation. Above all, the database contains emotional movements as they arise from non-archetypal, everyday actions. This raises its ecological validity over many of the previously recorded corpora. Despite this, the data is of very high quality and was recorded under controlled laboratory conditions. Compared to other laboratory-based databases the Glasgow corpus scores relatively high on Picard's criteria as it uses emotion induction to simulate event elicitation and targets both feeling and expression. At the same time the data is immediately usable for machine analysis. This makes it significantly more useful than uncontrolled video recordings which dominate the HUMAINE database. The ecological validity of the data is further strengthened by the fact that actions were recorded in sequences as well as in isolation. Very often experimental designs force subjects to display only isolated expressions, starting from and returning to a common neutral pose. While this makes analysis tasks such as segmentation easier, it also produces less natural emotion portrayals. The analysis of action sequences forms an important part of this dissertation.

### Recording procedure

Subjects were asked to act out the different actions in the various emotional styles. In order to help elicit more natural displays, a scenario-based script was produced for each emotion/action combination. For example, the angry scenario for the *Throwing* action read as follows.

> Today you slept in, so you had to rush to get ready. Then on the way to work, a policeman flags you down and gives you a speeding ticket, although you were just keeping up with traffic. You finally get to work where a note is waiting for you denying your request for having Friday off; now you are furious. Standing by your desk, you reach for a bit of rubbish and slam it into the bin as your temper flares.

During the recording, the instructors were careful not to perform the actions themselves in order to allow for fully personal interpretations. The effects of this are clearly visible in the data. For example, while some subjects throw objects with large, sweeping arm

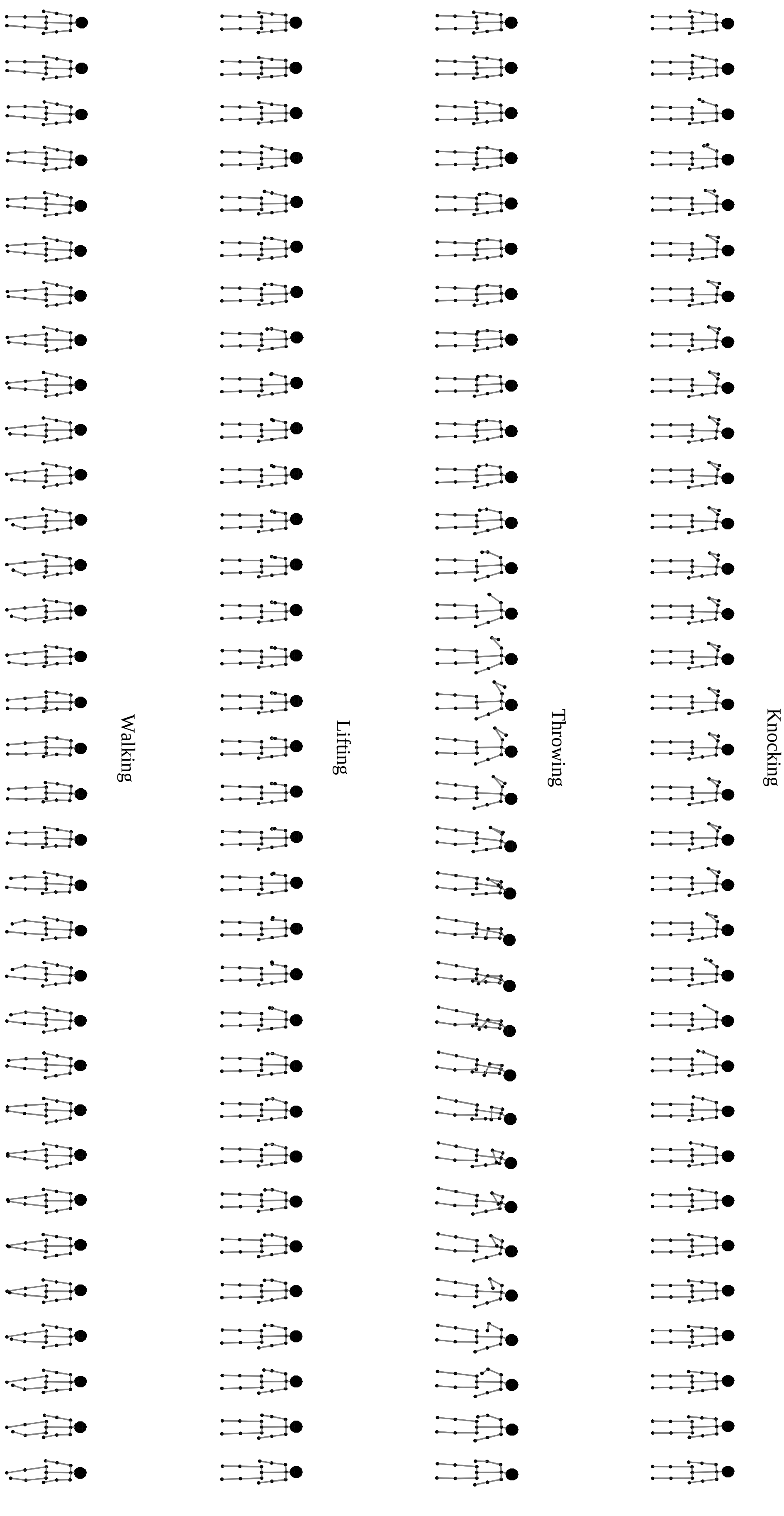

**Figure 2.5:** Examples of isolated actions found in the Glasgow corpus. From top to bottom: neutral *Knocking*, angry *Throwing*, happy *Lifting*, sad *Walking*.

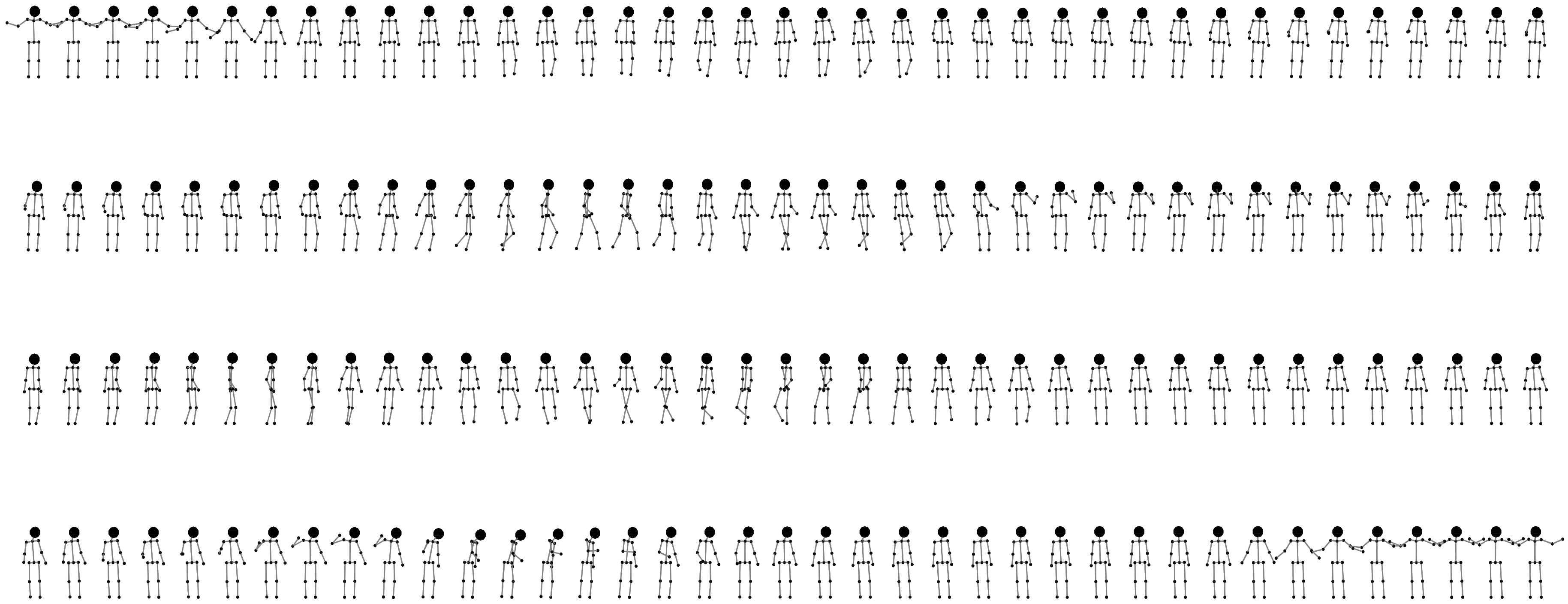

**Figure 2.6:** Example of a an *angry* action sequence from the Glasgow corpus. All sequences in the corpus follow the protocol: *TDown*, *Walking*, *Lifting*, *Walking*, *Knocking*, *Walking*, *Throwing*, *TUp*.

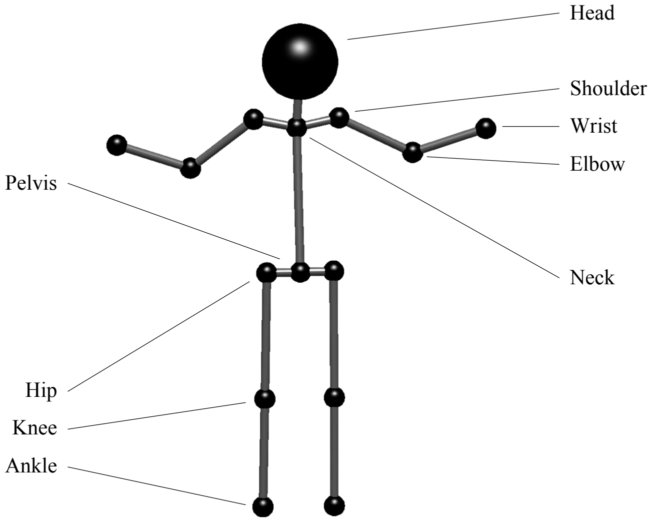

**Figure 2.7:** Body representation in the Glasgow corpus.

movements, others only use small wrist flicks. At the beginning of each recording, subjects were asked to stand in a T-pose, leading to an initial *TDown* action as the subjects lowered their arms into a neutral position before performing the first action. Similarly, subjects were asked to return to a T-pose at the end of the recording, leading to a *TUp* action. Figure 2.6 shows a series of frames extracted from one of the action sequences, which all follow the same protocol: *TDown*, *Walking*, *Lifting*, *Walking*, *Knocking*, *Walking*, *Throwing*, *TUp*. The creators of the corpus found it easy to supply the isolated actions complete with segmentation points. The sequences, however, are unsegmented in the corpus.

After the recording, emotional displays were not verified by independent judges. The emotional ground truth is hence derived solely from each individual's interpretation of the emotional state together with the eliciting scenario. While this is a common way to define the ground truth in the field, alternative or additional definitions have been employed occasionally including self-judgements of recordings and judgements by trained judges or peers [DCW$^+$08].

## 2.4.3 Motion capture

The data recorded for the corpus is of very high quality. As opposed to some other databases, the Glasgow corpus contains complete 3D position data for the 15 major body joints as shown in Figure 2.7. While subjects performed the actions, their bodies were tracked by multiple black-and-white 2D cameras of a commercially available Vicon motion capture system. In post-processing steps those camera images were combined to compute very accurate 3D joint positions at a data rate of 60Hz. The quality of this kind of data recording is clearly superior to the 2D recording of body shapes with a standard DV camera at the typical rate of 25Hz. Furthermore, by using infrared markers the motion-captured data is relatively noise-free (although significant manual post-processing is usually necessary to account for occlusions). This compares well to simpler marker-less techniques which can find it hard to track body parts such as the hands and head reliably [Cas08].

Of course, using motion capture technology comes at a cost. Firstly, subjects need to be fitted with infrared reflectors before recording. This can be both time-consuming and intrusive. As a result the motions might appear less natural. Also, data of such a high quality can currently only be obtained in a laboratory setting. With marker-less technology maturing, however, comparable accuracies could be achievable with far less intrusion in real-world settings in the future.

### 2.4.4 Mood or emotion?

The final point of this chapter returns to the issue of different affective phenomena. While the Glasgow corpus was recorded to capture the influence of *emotion* on body movement, the definition of emotion in the strict sense as presented in Table 2.1 only seems to be reflected to some extent. We saw that emotions are generally defined as *relatively brief episodes of synchronised response to the evaluation of an external or internal event.* In particular when looking at the scenario description given earlier, we see that the influence of the *Throwing* action notionally derives from events which far precede it, namely a sleep-in, the traffic and a speeding ticket. It is normally assumed that emotional responses do not last longer than a few seconds up to a minute. Furthermore, the emotional response in this case is clearly not a synchronised response from multiple modalities — instead the corpus reflects the *subtle motion qualities* which change because of the emotional state of the subject.

One way in which to reconcile this apparent mismatch is to postulate that the corpus captures a person's *mood*, a distinct affective concept which acts over far longer periods of time and at much less intense levels. A different view is communicated by Cowie and Schröder who propose a differentiation between *episodic* emotions (emotions in the traditional, narrow sense) and *pervasive* emotions (in a broader, more inclusive sense) [CS05]. Whichever way we look at it, it is clear that the dynamic qualities of emotions and the varying temporal scales at which affective processes impact on real-world situations are currently not very well understood and therefore largely ignored by current emotion recognition research. A resolution of this theoretical debate about the nature of emotions is beyond the scope of this dissertation. Instead, I shall continue to use the term *emotion* to refer to the concept of an affective process which may pervade large parts of a person's action in a broad sense.

# Chapter 3

# An experimental corpus of expressive body motions

In Section 2.2.3 I discussed the fundamental difference between archetypal and everyday body motions in relation to emotional expression. Although the ultimate goal of this work has been the development of an approach for non-archetypal emotion expression, this chapter will introduce a corpus of explicitly expressive body motions. As an important component, the chapter will document the recording of the expressive motion corpus. There are several motivations for treating these kinds of expressive body motions at this point. It will lay the foundations and serve as an introduction to the data processing and feature extraction techniques discussed in later chapters. The data will allow us to get a feel for the way different emotions manifest themselves in statistical motion features, and to what extent distinct individuals express emotions differently. I will also start to discuss the problem of defining decision boundaries to classify between different emotions.

Although some prior work on expressive motions exists, my experiments make new research contributions. While many studies dealing with expressive body gestures tend to focus on short episodes or portrayals of expression, this chapter specifically discusses sustained motions produced over many minutes. This chapter documents the use of a mixture of techniques to elicit and sustain a variety of emotional expressions from the subjects. Furthermore, the data has been recorded from untrained subjects, rather than professional actors or dancers. Before discussing the data recording and analysis in detail, I will give a brief overview over previously conducted studies using expressive motions and emotion archetypes.

## 3.1 Emotion in expressive body gestures

Almost all automated emotion recognition experiments so far have made use of explicitly expressive movements (in the sense defined in Chapter 2). By far the most common

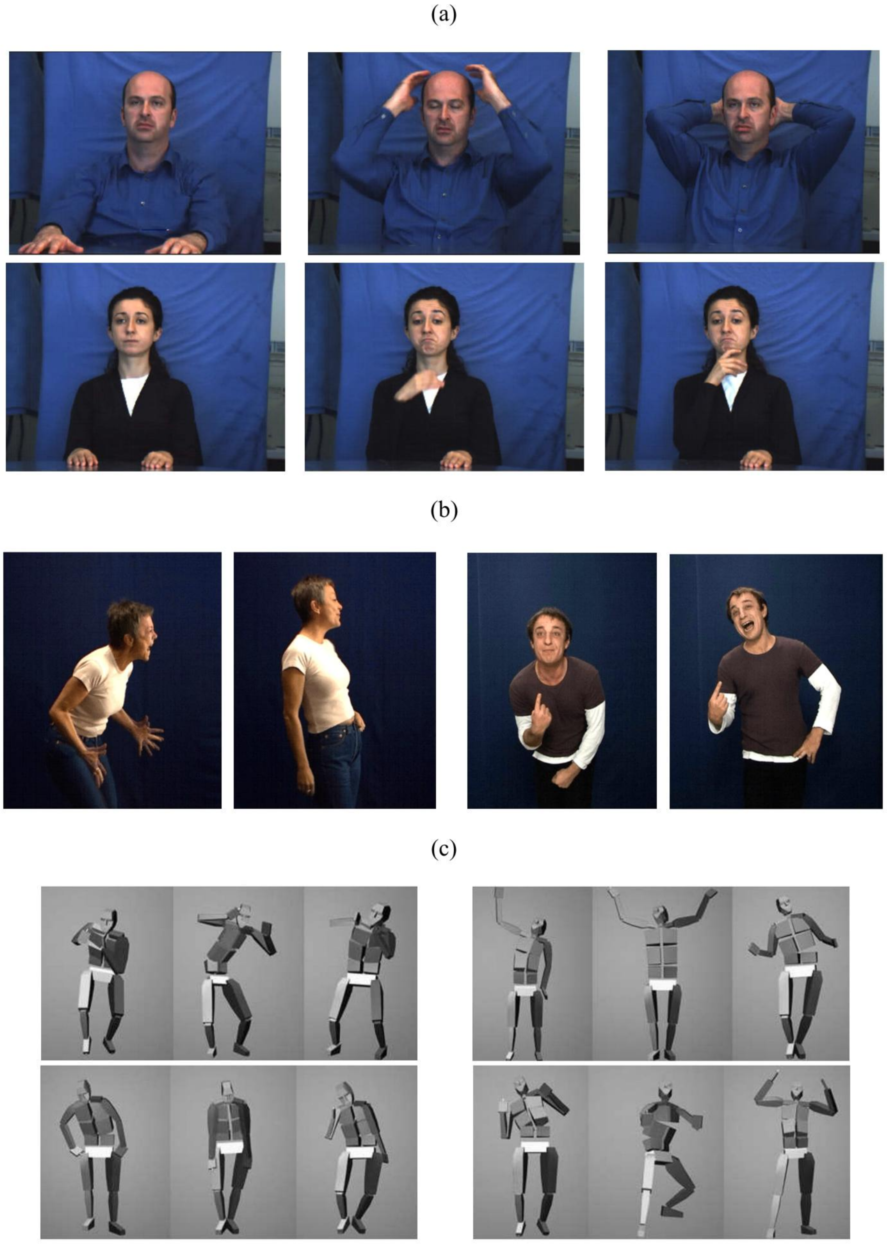

**Figure 3.1:** Examples of archetypal/expressive emotion corpora: (a) the FABO corpus [GP06] (b) the GEMEP corpus [BS07] (c) De Silva *et al.*'s affective posture corpus [DSBB04].

form of expressive motions are acted emotion portrayals. For this class of emotion data, actors (professional or otherwise) are instructed to act out certain emotions, usually with very few restrictions placed on the body movements allowed. Two databases of this kind which have been used in the past are the FABO and GEMEP databases (see Figure 3.1). Both databases include video recordings of the subjects' upper bodies and were recorded using small scenarios in order to help elicit convincing emotional displays. Scenarios might range from paragraphs of text describing an emotion-eliciting situation to one line descriptions such as "It was just announced that you won the biggest prize in the lottery". An overview of both databases together with some other corpora discussed in

**Table 3.1:** Corpora of expressive emotional body motions.

| | GEMEP<br>[BS07] | FABO<br>[GP06] | Kapur<br>*et al.*<br>[KKVB+05] | De Silva<br>*et al.*<br>[DSBB04] | Cambridge<br>corpus |
|---|---|---|---|---|---|
| subjects | 10 | 23 | 5 | 13 | 6 |
| professional actors | ✓ | | | | |
| professional dancers | | | ✓ | | |
| untrained | | ✓ | ✓ | ✓ | ✓ |
| basic emotions | 5 | 6 | 4 | 4 | 3 |
| other emotions | 13 | 3 | 0 | 0 | 2 |
| sustained emotions | | | | | ✓ |
| video cameras | 2 | 1 | | | |
| 3D motion capture | | | ✓ | ✓ | ✓ |
| reviewed during recording | ✓ | | | | |
| professional director | ✓ | | | | |
| expert ratings | ✓ | ✓ | | | |
| novice ratings | | | ✓ | ✓ | |
| self-ratings | | ✓ | | ✓ | |
| publicly available | ✓ | ✓ | | | |

the literature is provided in Table 3.1. While both databases were in fact created to allow multimodal analysis of face, body gestures and/or voice during emotional expressions, they were also used for body-only analyses. For example, Gunes and Piccardi found that 90% of the body motions portraying six different emotions in FABO could be classified correctly using a Bayesian Network classifier [GP05]. Other researchers have compiled similar portrayal-based corpora using higher fidelity motion capture technologies (see Table 3.1). Kapur *et al.* captured acted portrayals with a Vicon motion capture system and showed that four basic emotions could be classified correctly with an accuracy of 92% [KKVB+05]. Kleinsmith *et al.* investigate how the apex of emotional portrayals can be analysed in terms of emotional postures in order to discriminate different emotion classes [KSBB05, KBB07]. Many of these studies confirm that these kinds of emotion portrayals tend to produce very stereotypical reactions (archetypes), which seem to be consistent across different subjects. For example, Gunes and Piccardi find that *uncertainty* in their dataset can always be detected as a combination of shoulder shrug with palms pointing up [GP07]. Similarly, Balomenos *et al.* argue that there is a high correlation in their data between the emotional displays of joy and the gesture of a high frequency hand clap [BRI+05]. Similar observations apply to other studies.

The large interest in these kinds of scenario-based movements has been a fairly recent development. In the past, psychologists interested in the connection between emotion and body motion have often turned to dance. Refer to Section 2.2.1 for an overview of related studies. At this point it suffices to mention a recent study by Camurri *et al.* which compares the recognition performances of human observers and automated techniques for distinguishing emotions expressed by professional dancers [CLV03]. Although no algorithms are trained to distinguish different emotions, many motion-based cues are described which the authors argue to be correlated with emotion classes. Dance performances offer several features which make them particularly interesting for my goal of classifying everyday and naturally sequenced body motions. As opposed to archetypes, dance sequences tend to be continuous, without the artificial restriction of having to start and finish with a neutral pose. Secondly, as was done by Camurri *et al.*, archetypal portrayals can be explicitly inhibited by prescribing a sequence of motions which has to be performed. This results in emotional variations of a fixed motion sequence. The motion data of the Glasgow corpus is similar in this respect.

## 3.2 Collecting expressive body motions - a musical environment

In the previous section I discussed a number of studies which presented automatic techniques for analysing and, in part, distinguishing emotions in expressive body movements. This section introduces my own effort which shall serve two broad goals:

1. To confirm the viability of detecting emotion in expressive body movement
2. To inform the development of a computational approach to detecting emotions in the everyday body movements from the Glasgow corpus.

To this end, I am less interested in the stereotypical gestures elicited by scenario-based setups. As in previous studies, the setup should give individuals the freedom to express emotions according to their preference and without any significant physical constraints. Furthermore, as the Glasgow corpus features longer sequences of body motions, I am interested in recording sustained expressive motions rather than isolated portrayals. Finally, as in the Glasgow corpus, the data should be recorded from multiple subjects and for multiple emotion classes.

Recording sustained expressive body motion without a powerful elicitation mechanism can be very difficult. Most of the quoted studies tend to adopt a fairly simple procedure of acting short emotional displays, usually based on a brief scenario description and the desired emotion class to be exhibited. This method has several shortcomings. A frequently stated issue is the ecological validity of the displays recorded in this way. It is a common debate as to whether subjects are displaying in some sense true emotional reactions or simply acting the way they believe an according emotional reaction should look. Indeed, this can lead to the already mentioned risk of collecting stereotypical reactions rather than reactions as they would be observed outside the laboratory. It has been reported that experimenters sometimes have to prompt specific gestures when the subjects feel unfamiliar with a scenario [GP06] or may opt to prescribe the gestures altogether [CKC08]. Using a particular scenario to elicit expressive movement is therefore not an optimal solution in my case.

To help inform my recording setup, I will give a brief survey of the methods used in psychology to elicit certain emotional responses. *Mood induction procedures* (MIPs) commonly discussed in the psychology literature include:

- **Velten MIP**: subjects are confronted with self-referential statements ("I am really disappointed by this test result.")
- **Imagination MIP**: subjects are instructed to remember an emotional event in their lives
- **Story MIP**: subjects are instructed to identify with the protagonist in a story/scenario
- **Film MIP**: subjects are instructed to identify with the protagonist in a film clip
- **Music MIP**: subjects are exposed to emotionally evocative music
- **Feedback/Social Interaction/Gift MIP**: subjects are engaged in various kinds of social scenarios to elicit certain emotions.

Recently conducted experiments within the affective computing community seem to favour a small subset of these. The studies mentioned in Section 3.1 most commonly use a form of Story MIP, which stimulates the subjects' imagination to identify with a situation and its protagonist. While the Film and Story MIPs have been argued to be the most effective for certain mood induction scenarios [WSSH96], it is also well-known among researchers in the field that music can have a very strong emotion-inducing effect [Gil08]. Importantly for my goals, and as opposed to the commonly used Story MIP, music can provide a sustained source of stimulation. This is in contrast to more commonly employed setups using a Story MIP, where subjects/actors would have to act an emotion without assistance once they are familiar with the scenario.

One disadvantage of only using music stimuli by themselves is that their interpretations are extremely subjective. While I would like to encourage individual expressions of emotion, different emotional interpretations of the same music piece would significantly complicate the construction of a useful dataset. In order to establish a ground truth, experimenters usually supply subjects with an explicit instruction about the intended emotional expression. This is a common technique and an independent review of MIPs found that including this kind of instruction in the form of emotion labels with the Music MIP improves its effectiveness [WSSH96]. I confirmed that without the instructions one pilot subject felt unsure about the expected interpretation of the music and was hence not very expressive. The level of expressiveness improved markedly when I included the emotion labels alongside the music stimuli.

Including the emotion labels meant that choosing music pieces which "correctly" and uniquely expressed the intended emotion was not as critical as without the labels. The main requirements for the music were:

- engaging/immersive in order to stimulate sustained movement
- carrying as strong emotional connotations as possible.

I decided to use film music primarily as it largely fulfills the above criteria. I carried out the music selection process together with an undergraduate student in order to ensure a degree of objectivity. We started from a set of approximately 100 pieces, from which he selected a number of excerpts with particularly strong expression. We went through an iterative review process to filter out a number of pieces and according emotional classes which were particularly strong. This is the list of emotion classes and corresponding music pieces used in the final recording setup:

**Neutral**: *Neutral*: A recording of an orchestra tuning up before a concert

**Happy**: *Happy 1 & 2*: Two excerpts from George Enescu: "Romanian Rhapsody"

**Worry**: *Worried 1*: Excerpt from Howard Shore: "A journey in the dark" (film music for "Lord of the Rings")
*Worried 2*: Excerpt from James Horner: "Falkirk" (film music for "Braveheart")

**Sadness**: *Sad 1*: Excerpt from Niki Reiser: "Kai's Death" (film music for "Beyond Silence")
*Sad 2*: Excerpt from Edvard Grieg: "Ase's Death" (from "Peer Gynt Suite No 1")

**Fear**: *Afraid 1*: Excerpt from James Horner: "Revenge" (film music for "Braveheart")
*Afraid 2*: Excerpt from Howard Shore: "The passage of the marshes" (film music for "Lord of the Rings")

These pieces were confirmed to convey the intended emotions by five independent judges. Details of the according two experiments can be found in Appendix A.

In order to make experimental conditions equal for each subject, we developed an environment to play these excerpts according to a pre-determined script. The environment allowed pieces to be played for a certain amount of time and seamlessly fade between them in order to create an uninterrupted and emotionally engaging experience. I also added the ability to display the emotion class labels. In the final experiments, they were projected onto a screen in an otherwise darkened room. The system also allowed the logging of event times such as transitions between different emotional stimuli and the entering and exiting of the subject in the recording area. These allowed for a post-hoc association of body movements with the corresponding emotional labels.

## 3.3 Data recording

The recorded data was largely meant to be used in an exploratory fashion. For this purpose the data would be more interesting if it exhibited a lot of variation in the way emotions could be expressed. In the later sections we shall look at a quantitative and statistical analysis of the data. For this, I required a reasonably large number of data samples in order to make statistical measures robust and meaningful. In the traditional portrayal-based setup, this usually means asking for a number of repetitions from each subject or recruiting a very large number of subjects. One problem with asking for explicit repetitions is that it biases the subjects towards copying the same kind of actions multiple times, hence inhibiting more natural data variation. In order to encourage data variation for each emotion class, I gave subjects between 30 and 50 seconds at a time to continuously express an emotion. The periods for which different emotions were to be expressed were then concatenated in a pseudo-random sequence. Transitioning and continuous expression was supported by music and labels as described in Section 3.2. Once the long stretches of emotional expression had been recorded, I could consider shorter subregions and extract multiple samples from them.

In order to draw general conclusions from the Glasgow corpus, it is also important to consider a reasonably large variety of subjects, preferably with a good spread across genders and age. Another subject-specific aspect which may have a large impact on the observed data is the issue of using trained versus untrained subjects. Several corpora of expressive motions have in the past been collected of trained actors or dancers [BS07, CLV03]. This usually has the advantage that subjects are used to their body motions being recorded and analysed. It could be argued that, as a result, actors may be able to produce more realistic and expressive renderings of emotional expressions. Several reasons, however, favour using untrained subjects in my case:

1. There is an increasing interest in the affective computing community to move away from purely acted data. Using trained dancers in my scenario would be a step in the wrong direction.

2. As far as I am aware, no corpus has so far been recorded of untrained subjects expressing emotions through their body in connection with music. There have been studies, however, of professional dancers expressing emotions.

3. My analysis of the recorded data will inform the analysis of the everyday actions in the Glasgow corpus (see Section 2.4). This data was recorded from untrained subjects which makes the use of untrained subjects in this instance desirable.

I recruited six subjects, three male and three female. They were all members of the Department of Computer Science at the University of Cambridge and participated on a fully voluntary basis. I screened participants informally beforehand to establish whether they would feel comfortable with exhibiting expressive body motions. Each session lasted roughly one hour with 20 minutes taken up by the actual recording of data. At the start of the session, subjects were given an information sheet which outlined the procedure (see Appendix B for details). The most important instructions for this study were that

- subjects should try and exhibit motions throughout the whole session (*i.e.* not just when the stimuli changed)
- subjects were free to express the emotions in any way they wanted
- subjects were encouraged to be as expressive as possible.

To record the body movements, I used Vicon motion capture technology (see Section 2.4) which required subjects to wear an upper-body tracking suit and a specially prepared head cap. I only tracked the upper body and head in order to avoid the more significant discomfort for the subjects resulting from wearing a full body tracking suit. I set up the capture space to allow foot movement in an area of 2 × 2 metres which created a total capture volume of $4 \times 4 \times 2.5$ metres for unrestricted upper body and arm motions. Whenever subjects were asked to perform, no-one else was present in the room. There was very little illumination in the room in order to avoid distractions and allow for a maximal emotional effect from the music.

After calibrating the skeleton model as part of a Vicon recording session, I gave subjects five minutes to get familiar with the setup. During that time the environment was controlled by a "warm-up" script which cycled through the following set of emotions: *neutral*, *happy*, *worried*, *sad*, *afraid*. Subjects were encouraged to become comfortable with moving in the space and to find expressions which they believed convincing for conveying the according emotions. After the opportunity to ask any more questions, the main script started and controlled the environment for the next 14 minutes. During that time all body movements of the participant were recorded.

I recorded the 3D skeleton parameters at 100Hz using the Vicon software. This included 3D positions of the shoulders, elbows and wrists as well as the pelvis, neck joint and head. At the same time, the environment logged all changes in emotional stimuli (music and displayed label). Once I had captured all subjects, this allowed me to associate each frame in the movement recording with a certain emotion class (ignoring the regions where stimuli from different emotions overlapped). In the next two sections I analyse this data statistically with the goal of linking the observed joint movements to associated emotion labels automatically.

## 3.4 Defining emotion-communicating features

This section will outline the basic processing techniques for turning time series of joint positions into higher level motion descriptors. Within the machine learning and pattern recognition communities, these kinds of descriptors are usually referred to as features. In a conventional machine learning framework, the derived features will be supplied to trained algorithms for classification into different emotion classes. The goal is hence to derive a set of features which is maximally discriminative between the different emotion classes, while being invariant to irrelevant transformations within a class. Expressed in feature space, a selection of motion samples from two different emotion classes should ideally exhibit a large between-class variability while maintaining a small within-class variability. Choosing a set of discriminative features normally requires considerable human intelligence including domain knowledge. In the field of affective computing, domain knowledge can come in the form of qualitative or quantitative psychological studies. My work makes extensive use of the results from psychological studies, as well as useful insights from computer animation as discussed in detail in Chapter 2. First, however, we shall turn to a number of transformations to achieve basic feature invariances.

### 3.4.1 Invariances

The major invariance that is required for analysing motion capture data is invariance with respect to global skeleton position and orientation as the absolute world coordinate system can be arbitrary. As a result I am assuming that an expressed emotion does not influence the direction a subject is facing or where in the space he/she is located. While this might not necessarily be the case in all real life situations, it is a necessary assumption if no context information exists about the environment as in the two corpora used in this dissertation. In order to achieve this invariance, I define a body-centred coordinate system which moves and rotates with the subject (Figure 3.2). The body-centred coordinate origin lies at the pelvis of the skeleton. In order to define the local coordinate axes, I first define the coronal plane in terms of the shoulder-shoulder and spine-pelvis vectors. The z-axis is then defined as pointing perpendicular to the coronal plane. The y-axis is defined as the vector pointing from spine to pelvis, and together

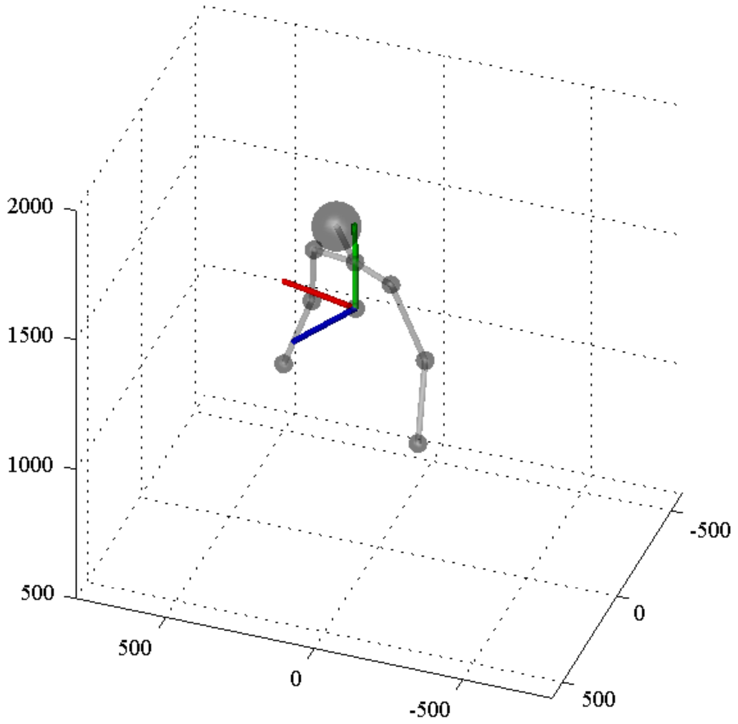

**Figure 3.2:** World and derived body-local coordinate systems. The body-local system's origin is located on the skeleton's pelvis and local x, y and z-axes are shown in red, green and blue respectively.

with the z-axis defines the sagittal plane. Finally, the x-axis is constructed perpendicular to the sagittal plane (and will necessarily lie in the coronal plane). For all subsequent processing, the joint data is mapped to this body-local coordinate system. I furthermore normalise all joint translations by the total arm length of the subject. This ensures that posture encodings are invariant with respect to absolute body size.

### 3.4.2 Posture descriptors

Joint position data is good for capturing a postural description of the body. In particular, I am tracking the body-centred coordinates of all 9 captured body joints

$$\boldsymbol{j}_i = (\hat{x}, \hat{y}, \hat{z}) \qquad \text{for } 1 \leqslant i \leqslant 9. \tag{3.1}$$

For emotion recognition each joint is therefore modelled as having three independent positional degrees of freedom. This representation clearly does not reflect all skeletal constraints. However, I will later use various forms of feature selection to remove statistical redundancies where necessary. Things such as raised arms versus lowered arms or a tilted versus upright head are discernable using these posture features alone. Time series of posture data are also an important component for recognising actions as I will discuss in Section 4.2. As far as emotion recognition is concerned, however, work from various related fields suggests that emotional content is also contained in the dynamic aspects of body motion (see Chapter 2). Cues such as the speed with which limbs are moving, the energy exerted through or the smoothness of a movement have been associated with emotional meaning by various authors. Kleinsmith *et al.* who have studied the automatic classification of static posture features into emotion class have repeatedly remarked that motion features such as direction and velocity are probably necessary to better recognise certain emotions [KSBB05]. This seems in agreement with psychological findings which state that humans have trouble distinguishing certain emotions without dynamical cues [Cou04, DSBB04]. Kapur *et al.* found that humans presented with moving point-light representations of human body joints can distinguish between the four emotions *happy*, *sad*, *angry* and *afraid* with an accuracy of 93% [KKVB+05]. In an independent study using the same emotional classes and recording methods De Silva *et al.* found that only presenting the most expressive static *posture* in the form of an avatar reduces the human recognition ability to 69% [DSBB04]. They observed that significant information was lost by reducing the dynamically expressed emotions to static postures and that as a result postures of different emotion classes could look similar. The lack of dynamic cues therefore probably played a major role for the significantly lower recognition performance in De Silva *et al.*'s experiment over that of Kapur *et al.* This result is supported by the numerous psychological studies I presented in Section 2.2. To capture the dynamic qualities of body movements I therefore derive additional features — speed, acceleration and jerk — which I describe next.

### 3.4.3 Dynamic descriptors

Mathematically speaking speed, acceleration and jerk have straightforward definitions. They are the first, second, and third derivatives of the joint position data with respect to time. As we are dealing with discretised time series data, they can be approximated by taking finite differences between neighbouring samples. In terms of physical significance they can be seen as describing different qualities of a motion:

1. Joint positions capture the posture or static qualities of the body: $\boldsymbol{j}_i(t)$.

2. Speed is related to the kinetic energy of a motion and hence describes how energetic a movement looks:

$$\text{speed}_i(t) = |\boldsymbol{v}_i(t)|^2 = |\boldsymbol{j}_i(t) - \boldsymbol{j}_i(t-1)|^2. \tag{3.2}$$

3. Acceleration is related to the force applied to move the body and can distinguish sudden from sustained or smooth motions. As force is a direct result of the amount

of muscle contraction, acceleration can also be seen as giving an indication for the amount of muscle tension:

$$\text{acceleration}_i(t) = |\boldsymbol{a}_i(t)|^2 = |\boldsymbol{v}_i(t) - \boldsymbol{v}_i(t-1)|^2. \tag{3.3}$$

4. Jerk captures the rate of change of acceleration. A large jerk indicates significant changes in acceleration and hence changes in applied force while a small jerk is indicative of motions such as circular motion at constant speed [FH85]:

$$\text{jerk}_i(t) = |\boldsymbol{a}_i(t) - \boldsymbol{a}_i(t-1)|^2. \tag{3.4}$$

In order to capture the expressive elements of arm movements, I thus compute normalised joint positions as well as speed, acceleration and jerk for the hand and elbow joints. Because these different channels are physically related to each other by the operation of differentiation, it might at first appear redundant to capture all of them. However, during the feature extraction process there are a number of non-linear operations, most notably the temporal aggregation described in Section 3.4.4. It is therefore absolutely necessary to capture those as different features. We will see in Section 5.4 that speed features, acceleration features *etc.* do indeed carry statistically independent information.

Another channel which I make use of is the head. In particular, head posture can be very indicative of emotional state. For example, a lowered head might indicate sadness. A common skeletal representation of the head (which is also found in the Glasgow corpus) is a single head joint connected to the neck joint. From this representation I can derive two angles which capture the head posture:

1. the rotation around the body-local x-axis (pitch).
2. the rotation around the body-local z-axis (roll).

Note that it is not possible to determine the third degree of freedom of head rotation (yaw) from this representation. Also, due to the simplicity of the assumed head model, it is only possible to register larger-scale head motions involving the cervical vertebrae. In order to derive higher precision measurements from the head, I would need to model it in more detail. Since other researchers have already studied the role of head motions in emotional displays in detail [Kal05], I will only include the pitch and roll angles, as well as the overall head speed, acceleration and jerk values in my analysis. The main focus, however, will be directed towards the significance of the rest of the body.

Table 3.2 summarises the channels used to represent the posture and dynamic qualities of joint movements. For each joint, I represent posture by the three normalised coordinates $(\hat{x}, \hat{y}, \hat{z})$. The dynamics are captured by the three scalar quantities of speed, acceleration and jerk. Together with the channels for the head this adds up to 29 channels, excluding

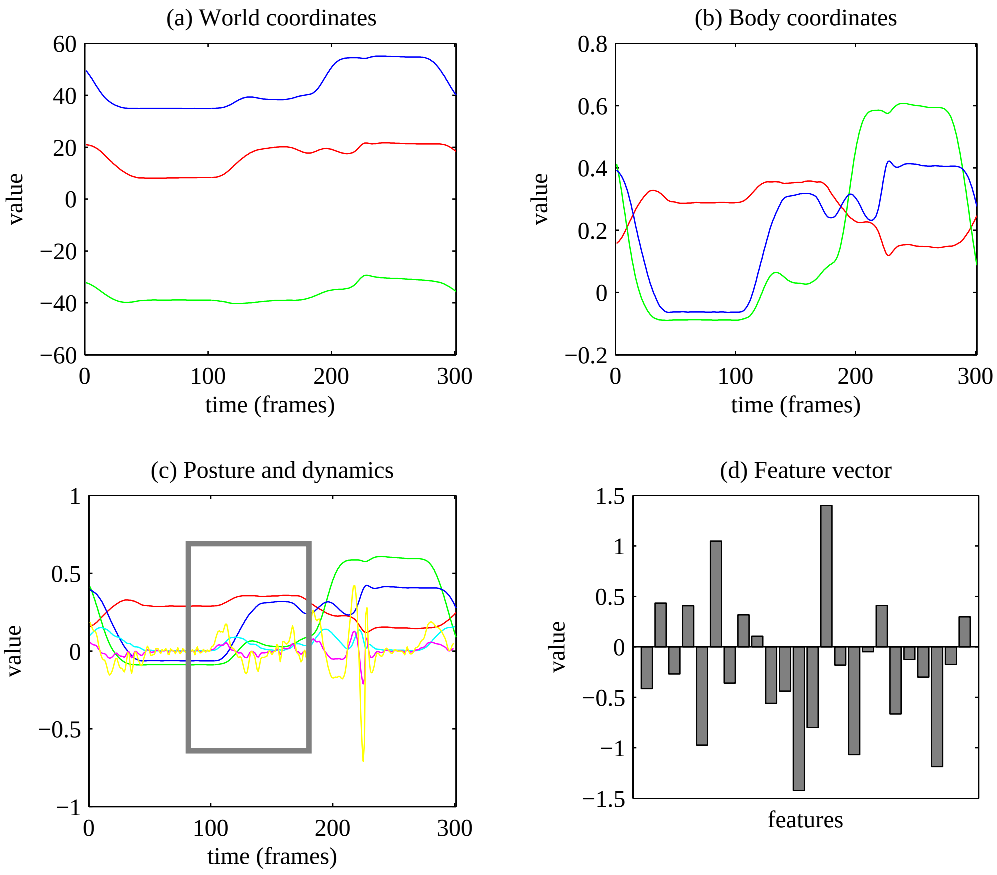

**Figure 3.3:** Feature extraction for the right wrist joint. (a) Joint positions in world coordinates $x$, $y$, $z$. (b) Body-centred coordinates $\hat{x}$, $\hat{y}$ and $\hat{z}$. (c) Posture and additional dynamic measures speed, acceleration and jerk. A grey time window of one second is highlighted. (d) 24-dimensional feature vector computed as median, standard deviation, minimum and maximum over the time window for each channel.

**Table 3.2:** Channels capturing posture and motion dynamics of the upper body.

| Body joint | Posture | Dynamics | # Channels |
|---|---|---|---|
| right elbow | $\hat{x}$,$\hat{y}$,$\hat{z}$ | speed, acceleration, jerk | 6 |
| right wrist | $\hat{x}$,$\hat{y}$,$\hat{z}$ | speed, acceleration, jerk | 6 |
| left elbow | $\hat{x}$,$\hat{y}$,$\hat{z}$ | speed, acceleration, jerk | 6 |
| left elbow | $\hat{x}$,$\hat{y}$,$\hat{z}$ | speed, acceleration, jerk | 6 |
| head | pitch, roll | speed, acceleration, jerk | 5 |
| | | | **29** |

**Table 3.3:** Subset of statistical features used for the exploration of expressive body motions from the Cambridge corpus.

| Body joint | Channels | Statistical features computed | # Features |
|---|---|---|---|
| right wrist | $\hat{x}$,$\hat{y}$,$\hat{z}$, speed | median, deviation | 8 |
| left wrist | $\hat{x}$,$\hat{y}$,$\hat{z}$, speed | median, deviation | 8 |
| head | pitch, roll, speed | median, deviation | 6 |
| | | | **22** |

the lower body. Figure 3.3(a)-(c) summarises the definition of posture and dynamic descriptors. For clarity of presentation, the figure focuses on the motion of the right hand only. Similar transformations are applied to the left hand and elbow joint positions. Also, this scheme generalises in the obvious way to the additional knee and foot joints when they are considered in future chapters.

### 3.4.4 Statistical feature definitions

The final step is turning the multiple *continuous channels* of motion descriptors into time-discrete *feature vectors*. The method adopted in this work is to consider a segment of the continuous signal at a time, say between time frames $t_{\min}$ and $t_{\max}$. The simplest way of determining $t_{\min}$ and $t_{\max}$ is to use a sliding window of fixed length and to produce new feature vectors at a constant rate as the window moves along the signal. If it is possible to build a better understanding of the underlying motion, I can also divide the signal into more meaningful segments. For example, I may want to detect certain semantic elements such as “hand raises” and compute features for those rather than using a sliding window. This issue will become more interesting once we are dealing with more complex everyday actions and will be the topic of Chapter 4. In the absence of any prior knowledge about the structure of the expressive motion data, using a sliding window is sufficient. For the analysis in this chapter, I will therefore use a sliding window of a fixed length of one second (100 frames) to determine $t_{\min}$ and $t_{\max}$. While I verified that using a bigger window size can lead to a better classification performance, using too big a window will decrease the number of samples that are available for training and evaluation. Also, the real time response time of a classifier will be dependent on the window size, making a longer window less desirable.

For a given $t_{\min}$ and $t_{\max}$, the goal is to find a concise numeric description of the signal. Throughout this dissertation I will use four statistical measures to describe the signal over the interval $[t_{\min}, t_{\max}]$:

1. sample median
2. sample standard deviation

3. sample minimum
4. sample maximum

Considering the features derived from all 29 channels as described above would give rise to a 116-dimensional feature space (29 channels × 4 statistical measures). Clearly, many of the resulting feature dimensions will be highly correlated, for example through factors such as the correlation of median and maximum values of a signal. There are also biological constraints which mean that elbow motion is clearly not independent from hand motion. It is therefore meaningful to consider certain subsets of this feature space for initial exploration. For the rest of this chapter, I will define my feature vector as the median and standard deviation derived from the posture and speed of both hands and the head. This amounts to a 22-dimensional feature vector which is more manageable for this initial exploration (see Table 3.3). In the final section of this chapter I will confirm that these features do indeed capture important elements of emotional qualities. Selecting good features from a large set in a more systematic fashion is an important problem and I will return to it in Section 5.4.

## 3.5 Exploring the data

In this section I am going to analyse the feature space constructed for the Cambridge corpus. The treatment will be largely exploratory in nature and is aimed at building an understanding and intuition for the general structure of the data at hand. Some of the questions that are relevant at this point are:

1. Is the feature set described in Section 3.4 suitable for capturing differences in emotional expressions?
2. From a pattern classification standpoint, to what extent do the different emotional classes appear to be separable?
3. Is there any interesting structure to the data which I can exploit later to build effective classifiers?

The above questions are meant as a guidance for the data exploration and are not intended as problem statements to be answered conclusively. However, insights gained in this section will motivate the approaches adopted in later chapters. I will make use of two common data exploration techniques which have distinct advantages: Multiple Discriminant Analysis (MDA) and Multivariate Density based Discriminant Functions (MD-DFs). The first half of this section will focus exclusively on *visualising* the complex feature space with the help of MDA. Visualisation is a crucial step in building an intuition around the structure of the data. The second half focuses on *quantifying* the separability between the different emotion classes in the data with the help of MD-DFs.

In Section 3.4 I described in detail the construction of a 22-dimensional feature space capturing motion qualities of the upper body. It is very hard to build an intuition around and visualise the structure of such high-dimensional data. A commonly used technique to familiarise oneself with a set of data is a 2D scatter plot. In order to produce this I need to make use of a dimensionality reduction technique such as MDA. While MDA attempts to preserve as much detail about the feature space as possible in the lower-dimensional visualisation, such a radical projection is likely to suggest a worse separability than is possible in the original higher-dimensional representation. In the second half of this section I will therefore make use of all 22 dimensions to define MD-DFs. MD-DFs are specifically aimed at computing boundaries between the emotion classes which minimise classification errors, assuming a parametric distribution for each class. Note that in this chapter I will only look at the feature space as a whole to decide if the features contain enough information to separate the emotion classes. I will provide a detailed analysis of individual features in Chapter 5.

### 3.5.1 Multiple Discriminant Analysis

In Section 3.4 I accumulated a number of features which previous studies suggested may carry some information about emotion in body movements. I was not particularly careful to pick a small set of features, however, which makes visualising the constructed feature space virtually impossible. The most common way of dealing with this problem is to linearly combine original features $\phi_1 \dots \phi_i$ to create new dimensions for visualisation,

$$\hat{\phi}_n = w_{n_1}\phi_1 + w_{n_2}\phi_2 + \ldots + w_{n_i}\phi_i. \tag{3.5}$$

The problem is how to define weights $w$ such that a maximal amount of information about the data is preserved by this transformation. Two classical approaches exist to defining the weights. Principal Component Analysis (PCA) computes weights such that the new dimensions preserve the maximum amount of *variation* in the original data. Multiple Discriminant Analysis (MDA) on the other hand finds weights such that defined classes of data are maximally *separated* after the transformation. Although both techniques amount to different visualisations of the same underlying data, MDA promises to give more useful insights as I am ultimately interested in the *separability* of the emotion classes. As visualisation requires a two-dimensional representation of the data, I jointly compute weights for the two transformed dimensions $\hat{\phi}_1$ and $\hat{\phi}_2$.

The details of MDA are well understood and documented [DHS00]. For my purposes it suffices to note that the solution for the weights $w$ has an algebraically closed solution and is unique up to linear transformation on the resulting space. This means that the visualisations I am deriving are for all practical purposes unique. MDA determines the weights $w$ such that the $\frac{\text{between-class-scatter}}{\text{within-class-scatter}}$ ratio is maximised in the resulting space and thus ensures that the underlying classes are maximally separated in the transformed representation. Formally, MDA maximises the objective function

$$J(\boldsymbol{W}) = \frac{\det(\boldsymbol{W}'\boldsymbol{S}_B\boldsymbol{W})}{\det(\boldsymbol{W}'\boldsymbol{S}_W\boldsymbol{W})} \tag{3.6}$$

to find the optimal weight matrix $\boldsymbol{W}$ where $\boldsymbol{S}_B$ and $\boldsymbol{S}_W$ are the between and within-class scatter matrices respectively for a set of emotion classes $\mathcal{E}$ and are defined as

$$\boldsymbol{S}_W = \sum_{e\in\mathcal{E}} \sum_{\boldsymbol{\phi}_i \text{ of class } e} (\boldsymbol{\phi}_i - \boldsymbol{\mu}_e)(\boldsymbol{\phi}_i - \boldsymbol{\mu}_e)' \tag{3.7}$$

$$\boldsymbol{S}_B = \sum_{e\in\mathcal{E}} n_e(\boldsymbol{\mu}_e - \boldsymbol{\mu})(\boldsymbol{\mu}_e - \boldsymbol{\mu})' \tag{3.8}$$

Figures 3.4 & 3.5 show the features transformed using MDA and colour-coded by emotion class. In particular, Figure 3.4 shows a plot of all six subjects' data combined, while Figure 3.5 visualises the data for each subject individually. Inspecting the figures leads to several observations:

1. There is a clear pattern to the data in terms of distribution of emotion class. In Figure 3.4, each class occupies a distinct region of the feature space. This means that despite the lack of explicit instructions, subjects tend to express emotions in similar ways. Also, the features defined in Section 3.4 are clearly able to capture some of the emotional differences.

2. The emotional classes seem separable to some extent, but in two dimensions there is a significant overlap. *Happy* seems to be the class which is most easily separated from the others. Some emotions, such as *worried* and *sad* tend to occupy similar regions in the plots and are hence probably harder to separate. Some of these observations are subject-specific.

3. Despite the overall tendency of samples of the same emotion to cluster together across different subjects, we see a number of subject-specific differences.

   - The exact position and variation of individual emotion regions can vary between subjects (*e.g.* compare the *happy* samples for subjects 1 and 3).
   - The overall variation of samples in the space can vary between subjects. This might be interpreted as different levels of expressivity in different subjects (*e.g.* subject 4 appears more expressive than subject 5).

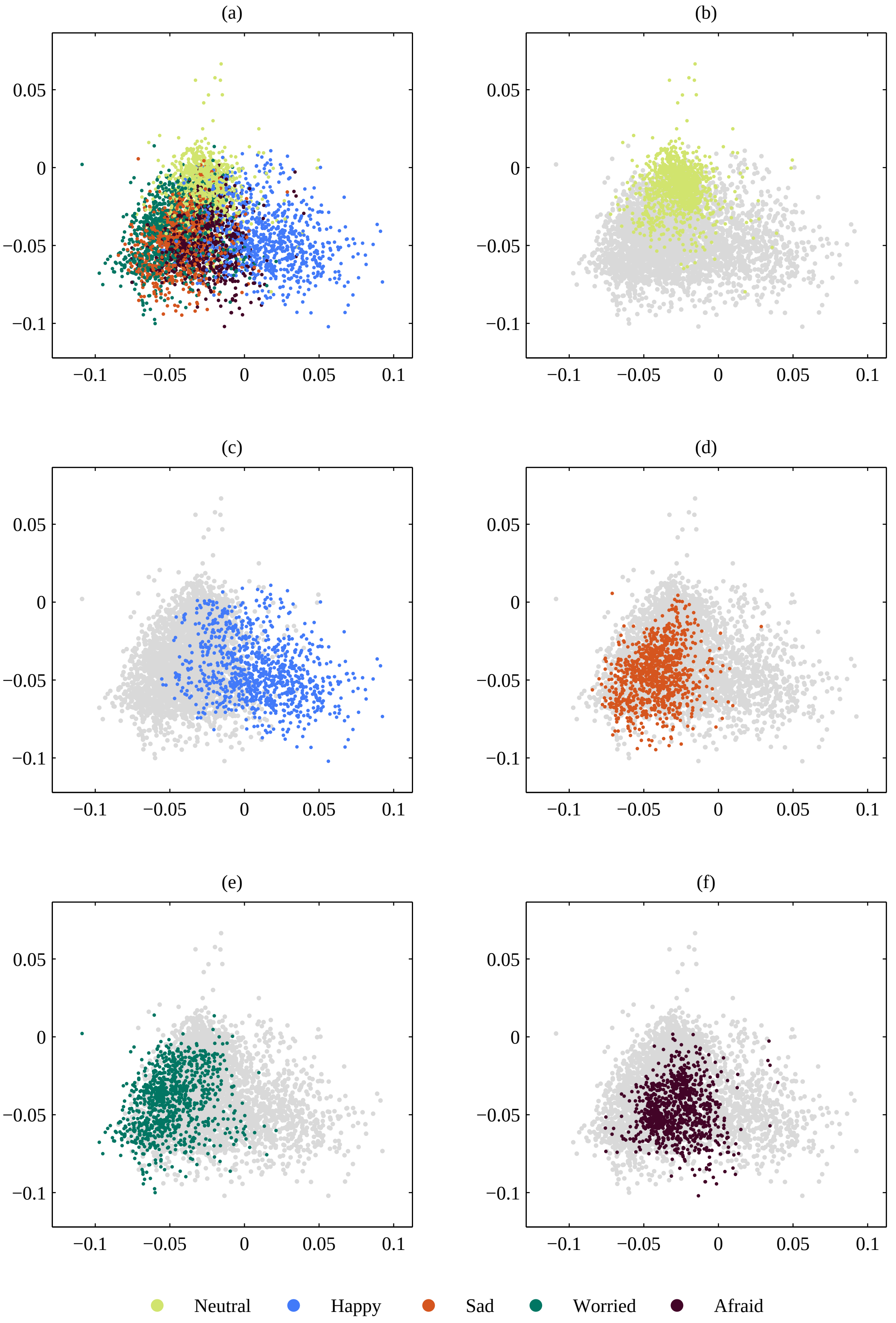

**Figure 3.4:** MDA-based projection of the features extracted from the Cambridge corpus. Features from all six subjects are combined. (a) all emotion classes; (b)-(f) individual emotion classes are highlighted and all other classes are combined.

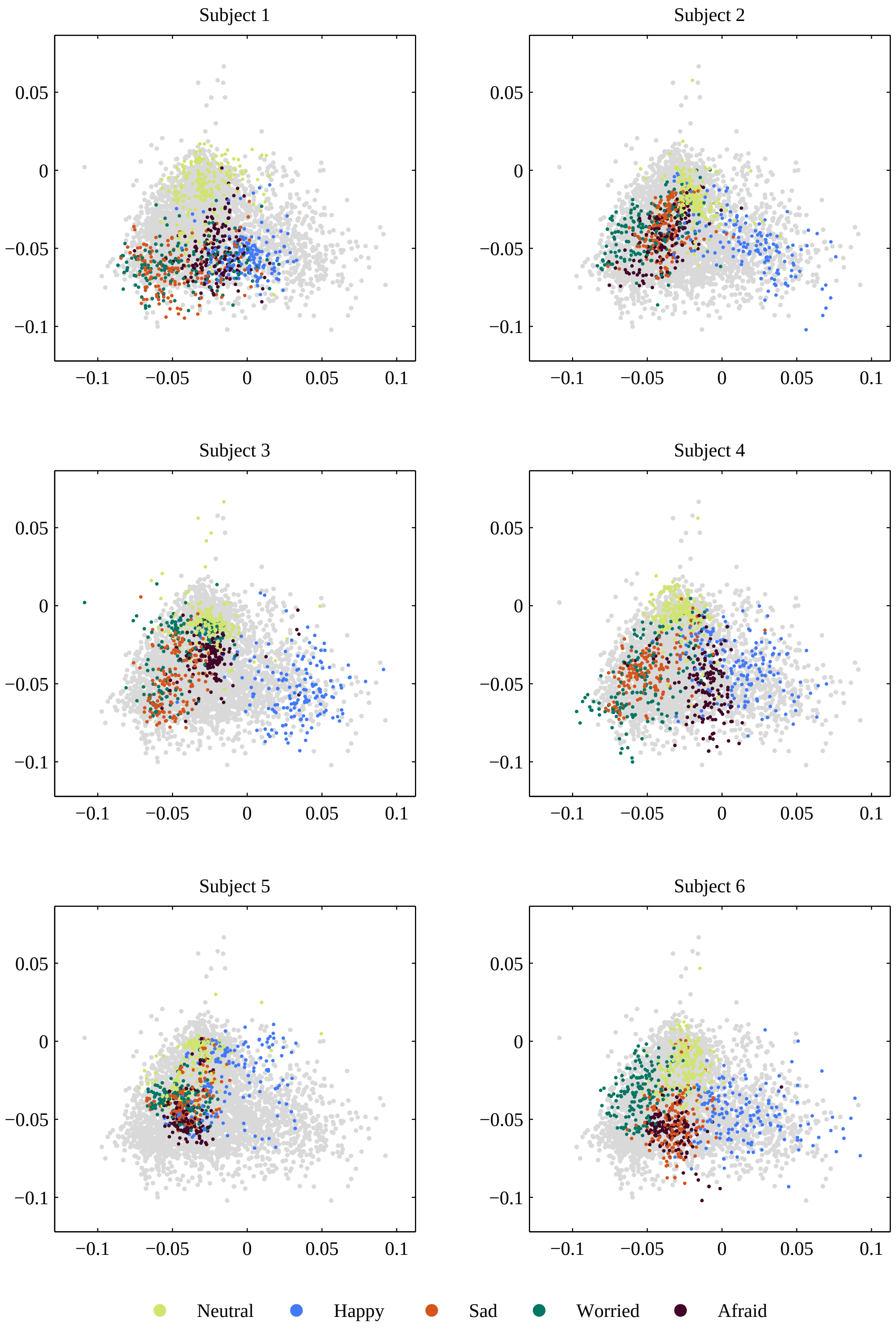

**Figure 3.5:** Features extracted for the six subjects recorded for the Cambridge corpus. The features are projected using a global MDA transform computed on the set of features from all subjects. The gray background represents samples from other subjects.

- Which emotions are separable can vary between subjects (*e.g. neutral* is very distinctive for subject 1, *happy* seems most easily separable for subject 3).

We should remember that we are looking at a 2D projection of the feature data. Some of my observations are likely to be affected by that. In this case dimensionality reduction was necessary for plotting the data in a 2D space. As a result emotion classes might appear less separable than they are in higher dimensions. Most computational pattern recognition techniques can deal with data of much higher dimensionality. In order to get a more comprehensive and quantitative view of the separability of the data I will turn to one such technique — multivariate density discriminant functions.

### 3.5.2 Multivariate density discriminant functions

MD-DFs are simple discriminant functions defined through multivariate normal densities. In Figure 3.4 we saw that emotion classes occupy well-defined and connected regions in feature space which makes this normal assumption suitable. Each emotion class $e$ is modelled by its own distribution $N_e$ — in this case parameterised by a 22-dimensional mean vector $\boldsymbol{\mu}_e$ and covariance matrix $\boldsymbol{\Sigma}_e$. The distributions are also called the discriminant functions for their respective classes. They express the degree of belief that a certain sample belongs to a certain class. Decision boundaries are then defined between classes where the surfaces of the discriminant functions intersect. The distribution parameters can be estimated by standard methods in order to minimise the error rate [DHS00]. Depending on the assumptions I make about the parameters, I obtain boundaries between classes of various complexities. For example, if I assume that the covariance matrix of each emotion class is the same ($\boldsymbol{\Sigma}_e = \boldsymbol{\Sigma}$) I obtain piecewise linear boundaries. Relaxing this assumption gives rise to piecewise-quadratic boundaries. The complexity of the discriminant functions can also be controlled by enforcing a diagonal versus allowing a full covariance matrix. Covariance matrices are usually restricted if there is not enough training data to reliably estimate all degrees of freedom without overfitting to the data. Having computed the parameters, I estimate the separability of the space in terms of the correct classification rate. In this section I will count a sample $s$ of class $\hat{e}$ as classified correctly if and only if $N_{\hat{e}}(s) \geqslant N_e(s)$ for all $e \in \mathcal{E}$.

Table 3.4 gives the results for computing the classification rates on all subjects and the combined space for the different covariance assumptions. In each case feature vector samples were extracted using the sliding window, leading to around 800 samples per subject (160 per emotion class). The results for the first six rows were hence obtained by fitting five Gaussian discriminant functions based on 160 samples each. The last row of results was obtained by combining the data samples of all subjects. In that case the discriminant functions were hence estimated based on $160 \times 6 = 960$ samples each. In

**Table 3.4:** Misclassification rates for the Cambridge corpus using discriminant functions derived from normal density models with different covariance assumptions. Functions were constructed for each subject individually (top rows) and for a combined space of all subjects' features (last row).

| Subject | Linear | Linear | Quadratic | Quadratic |
|---|---|---|---|---|
| | (full cov.) | (diag. cov.) | (full cov.) | (diag. cov.) |
| subject 1 | 0.85 | 0.68 | 0.94 | 0.75 |
| subject 2 | 0.75 | 0.68 | 0.89 | 0.60 |
| subject 3 | 0.80 | 0.67 | 0.88 | 0.56 |
| subject 4 | 0.81 | 0.74 | 0.92 | 0.76 |
| subject 5 | 0.64 | 0.55 | 0.84 | 0.47 |
| subject 6 | 0.90 | 0.81 | 0.95 | 0.73 |
| combined | 0.64 | 0.52 | 0.65 | 0.45 |

each case the classification rates were computed based on all samples (no explicit training and test sets). We can see the following:

1. The feature set works relatively well in this modelling scenario. Even the most restrictive model (linear with diagonal covariance) achieves classification rates between 70% and 80%.

2. As expected the emotion classes seem more separable than the MDA plots suggested. This is especially true when building models for each subject individually and using quadratic boundaries with full covariances. In that case classification rates range between 84% and 95%.

3. The complexity of the classifier can have a significant effect on the separability of the space. In particular, the most complex classifier is able to separate the classes best. This is only the case when considering the subjects individually.

4. Subject-specific differences are apparent in two ways:

   - data from some subjects is harder to discriminate than from others (*e.g.* classification rates for subject 6 are significantly higher than those for subject 5)

   - the very step of modelling individual subjects' data separately improves the separability significantly. This did not seem very apparent from the MDA analysis and is consistent across all classifier complexities. This suggests that there are subject-specific patterns of expression which I should take into consideration when building a classifier.

It is important to realise that these numbers are not necessarily good predictors for actual performance of real classifiers on new data. I did not train and test the discriminant functions on different data sets. Overfitting the discriminant functions is a potential problem, especially with little data and complex classifiers. In particular, the good performance of the quadratic classifier with full covariances is probably due to overfitting and would not generalise to new data. The plots in Figures 3.4 & 3.5 suggest, however, that ideally different covariances should be estimated for different emotion classes as the distributions show different amounts of variation and different orientations in feature space (*e.g.* compare variations and orientation of classes *neutral* and *happy* in Figure 3.4).

### 3.5.3 Implications for future chapters

We can take away the following insights from this chapter:

1. The statistical features capturing body posture and movement dynamics (Table 3.3) are appropriate for exposing emotional differences in expressive body motions. In future chapters we will see that they work equally well for everyday body motions. I will also take a more careful look at the contribution from different kinds of features in Section 5.2.

2. The use of classifiers which can define boundaries of different degrees of complexity may be important. This question could not be answered conclusively in this chapter due to a relatively limited amount of data and the resulting problem of overfitting. To investigate this issue further I will be using Support Vector Machines for my experiments in Chapters 5 and 6. By using the kernel method, they are able to construct decision boundaries of varying complexities.

3. Subject-specific differences are clearly observable in the recorded data. In the remaining part of this thesis I will develop methods to capture personal differences in emotional body expression. Importantly, I will decouple the modelling of personal differences from the emotion classification problem. This will allow me to treat personal differences as a factor just as important as the influence of emotion on a motion signal.

Before developing the idea of personal motion idiosyncrasies further, I will turn to modelling and analysing complex everyday actions in the next chapter.

# Chapter 4

# Modelling complex human body motion

This chapter is going to introduce and discuss complex everyday actions. In Chapter 2 I mentioned that these are *motions which are motivated by goals other than communicating affect*. It is the kind of motion which forms the core of this dissertation. Despite their real-world relevance, everyday actions and activities have so far been largely ignored by the affective computing community. In this chapter I will therefore present fundamental processing methods, which will allow me to effectively analyse complex actions and body movements. The focus of this dissertation will be on the everyday actions contained in the Glasgow corpus: *Knocking*, *Throwing*, *Lifting* and *Walking*. I will discuss a common generative approach to modelling these kinds of actions. Subsequently, I will present two methods for analysing the structure of actions at a finer level of detail, the level of motion primitives. While I will use examples from the Glasgow corpus throughout this chapter, my aim is to develop methods which are as general as possible. This includes the ability to generalise to new action categories and an ability to add new categories without any significant user intervention.

## 4.1 Problem description

Complex everyday motions commonly arise from some practical, real-world need such as the desire to reach and move an object. Indeed, the human body has evolved to perform very sophisticated everyday tasks which require a large amount of coordination of different parts of the body. As a result, many everyday actions exhibit quite distinct spatio-temporal structures of movements. In addition, actions which involve different body parts may be carried out in parallel, such as drinking a glass of water while walking. Actions which are carried out concurrently may affect each other [WLH07]. However, the modelling of these kinds of influences is beyond the scope of this dissertation. I will focus on the factors of action, individual movement idiosyncrasies and emotion with the goal of detecting emotion.

The most common and most studied action type is probably human gait [Whi02, Whi96, KSR$^+$04, LG02]. It has evolved to provide the most energy-efficient form of locomotion. At the same time, there are many factors which influence the exact timing of gait patterns. Those factors range from constants such as gender [LG02] and long-term factors such as height [Abd02] to mid-term factors such as age and body weight [TE01] and short-term effects such as emotion [CG07, JSL$^+$08]. In consequence, human gait has been studied by many researchers to *predict* variables like gender, identity and more recently emotion.

Most previous work has focused on very specific action categories such as walking. In those cases, the intricate knowledge about the action at hand and the orchestration of component movements clearly helps to distinguish the effects of factors like identity or emotion from action-specific factors. In the case of walking, for example, frequently used features include very specific variables such as stride length and cadence [CG07, Abd02]. The same argument applies to my more general case of detecting emotions in different types of everyday actions. For example, when trying to recognise an emotion from an arm raise, it is useful to know whether the arm is carefully lifting a glass of water or whether it is raised to knock on a door. For a knocking motion, hand speed is likely to give a good idea about the emotional state of a person [PPS01], while other parameters such as the height of lifting might be more constrained by the goal of knocking on the door. In the former case of lifting a glass, both speed and height of lifting are likely to be heavily constrained by the action. In conclusion, in order to recognise emotions from everyday motions, it is necessary to know which action is being performed. I will present a suitable action recognition framework in Section 4.2.

In addition to action-level patterns, I will investigate subaction-level patterns. This analysis assumes that complex actions are composed of simpler movements, sometimes referred to as motion primitives [HG04, FMJ02]. Rather than treating a whole action as supplying one piece of evidence for recognising an emotion, I will investigate if it is possible to obtain additional insights by considering the more detailed structure of actions. Using the previous knocking example, the first approach derives features from a whole *Knocking* action. By adding some knowledge about the common structure of *Knocking* actions, the primitive-based approach derives features from its components individually: arm raise, repeated knocking, arm lowering. I will describe two semi-supervised ways of defining and extracting these primitives from isolated actions in Section 4.3. My approach also accounts for the fact that many actions such as *Walking* and *Knocking* can have cyclic elements (see Section 4.3.3). Finally, I am going to measure the benefits of analysing emotions at the level of motion primitives in Chapters 5 and 6.

## 4.2 Generative modelling of actions using HMMs

In the absence of any context information, actions have to be defined and identified by the spatio-temporal trajectories of body joints. Formally, I represent an action category $c$ as a set of joints $\mathcal{J}_c$ and a description of its elements' movements over time, $\lambda_c$. I will regard two action categories $c_1$ and $c_2$ as temporally compatible if $\mathcal{J}_{c_1} \cap \mathcal{J}_{c_2} = \emptyset$. If two actions are temporally compatible, they can occur simultaneously as their joint movements are independent. These kinds of simultaneous movements are observable in the sequential actions of the Glasgow corpus.

The rest of this section will largely deal with the problem of describing the joint movements $\lambda_c$ in terms of Hidden Markov Models.

### 4.2.1 HMM formalism

The temporal evolution of joint parameters in an action can be captured by a Hidden Markov Model (HMM). HMMs are a well-known stochastic method for modelling temporally evolving systems which produce observable outputs. At any one time, the system can be in one of $s$ hidden states $\omega_1 \ldots \omega_s$. $\boldsymbol{\omega}^T$ denotes the entire state sequence a system goes through from time frame 1 to $T$. A system's state $\omega(t)$ at time $t$ is governed by a Markov process, that is $P(\omega(t+1) = \omega_i | \boldsymbol{\omega}^t) = P(\omega(t+1) = \omega_i | \omega(t))$. The set of state transition probabilities $P(\omega(t+1) = \omega_j | \omega(t) = \omega_i) = a_{i,j}$ define the transition matrix $\boldsymbol{A}$. At each time frame, the system emits an $n$-dimensional vector of visible observations $\boldsymbol{v}(t)$. The probability of observing a particular output is conditioned only on the current state $P(\boldsymbol{v}(t) = \boldsymbol{v} | \omega(t))$. The elements of $\boldsymbol{v}$ can either be drawn from a discrete alphabet or a continuous set of values. In general, joint movements will exhibit complex trajectories in position and speed and I therefore model $\boldsymbol{v}$ as a vector of continuous variables. Hence, $P(\boldsymbol{v}(t) = \boldsymbol{v} | \omega(t))$ needs to be a continuous probability density function. I will use the normal density as its successful use has been widely documented [Rab02]. The parameters needed to specify the observation probabilities are therefore a $d$-dimensional mean vector and covariance matrix $P(\boldsymbol{v}(t) = \boldsymbol{v} | \omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.

Every action category $c$ is modelled by a separate HMM $\lambda_c$. A full HMM $\lambda_c$ is defined by:

- the number of states $s$
- the transition matrix $\boldsymbol{A}$
- the prior belief vector $\boldsymbol{\pi}$ about the initial state of the system, where $\pi_i = P(\omega(1) = \omega_i)$
- the dimensionality $d$ of the observation vector $\boldsymbol{v}(t)$

- the observation distribution parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$.

Because different actions may be defined in terms of different numbers of joints, I do not assume that observation vectors for different categories are of the same dimensionality, *i.e.* $d_{c_1} \neq d_{c_2}$ in general. This requires some care during classification as I will discuss in the next section.

### 4.2.2 Action models

In this section I am going to describe the parameters for models used to represent the actions in the Glasgow corpus. They are developed with the ultimate goal to be general enough to apply to both isolated actions, where only one action happens at a time, and action sequences where certain actions may happen concurrently. For example, in the Glasgow corpus multiple actions can be interleaved such that a person is walking while already starting to raise the arm for subsequent knocking. In order to model such parallelism I divide the set of joints used to represent the body into two independent regions, the upper body and the lower body. The upper body $\mathcal{J}_U$ comprises the head, left and right arm joints and the neck. The lower body $\mathcal{J}_L$ comprises the pelvis as well as the left and right leg joints. All actions in the Glasgow corpus are defined as either upper body actions or lower body actions. To ensure that upper body actions $c \in \mathcal{C}_U$ can happen in parallel with lower body actions $c \in \mathcal{C}_L$, I require that upper body actions can only be defined in terms of upper body joints and lower body actions can only use lower body joints:

$$c \in \mathcal{C}_U \quad \Leftrightarrow \quad \mathcal{J}_c \subseteq \mathcal{J}_U \tag{4.1}$$

$$c \in \mathcal{C}_L \quad \Leftrightarrow \quad \mathcal{J}_c \subseteq \mathcal{J}_L \tag{4.2}$$

I am dividing action categories into upper and lower body actions as follows:

$$\begin{aligned} \mathcal{C}_U &= \{\textit{Knocking}, \textit{Throwing}, \textit{Lifting}, \textit{TUp}, \textit{TDown}, \textit{other}\} \\ \mathcal{C}_L &= \{\textit{Walking}, \textit{other}\} \end{aligned}$$

Note that I am including the T-pose actions in this set as it is necessary to distinguish them in the action sequences (*c.f.* Section 2.4.2). Furthermore, I am defining the *Walking* model in terms of the lower body only. This allows me to model a *Walking* action happening in parallel with parts of upper body actions as outlined above. In order to be able to assume that the upper body is always performing some action, even if the subject is only walking, I am adding an *other* action category. Similarly, if the subject is standing still while performing a *Throwing* action, the lower body will register an *other* action.

### Observation variables

The essence of an action is its sequence of posture and movement changes. I am therefore using the joint position and speed information as the observation variables (see Section 3.4). In addition to those, I also define

- upper body twist: angle between the normals to the coronal planes (z-axes) defined on the upper body (using the shoulder joints) and the lower body (using the hip joints): $\text{twist}(t) = \arccos(\boldsymbol{z}_{\text{upper}}(t) \cdot \boldsymbol{z}_{\text{lower}}(t))$.
- global body speed: this is not captured by the other dimensions as they are all encoded relative to the position of the pelvis joint. The global body speed is defined as a scalar difference in position of the pelvis joint $\boldsymbol{j}_{\text{pelvis}}$ over time. Because the recording quality is so high, I managed use $\Delta t = 1$: $\text{speed}_{\text{glob}}(t) = |\boldsymbol{j}_{\text{pelvis}}(t) - \boldsymbol{j}_{\text{pelvis}}(t-1)|^2$.

Because actions are defined in terms of a joint set $\mathcal{J}_c$ only a certain subset of these variables can be used as observation variables for each action. For example, $\mathcal{J}_{\text{Walk}}$ does not contain any upper body joints. Therefore, $\lambda_{\text{Walk}}$ does not output any variables derived from the upper body, such as elbow position or hand speed. Furthermore, actions do not have to make use of all variables derived from their assigned body part. For example, a (right-handed) *Knocking* model only outputs variables derived from the right half of the upper body and $\mathcal{J}_{\text{Knock}} = \{\text{right hand}, \text{right elbow}, \text{right shoulder}\} \subset \mathcal{J}_U$. Since there are no explicit left and right upper body regions, this is not in order to support concurrent left and right hand actions. Instead, this reduction in output feature space is useful to avoid overfitting the model to aspects of training samples which should be irrelevant for an action model. For example, consider the model for a right-handed knocking action. It is important that the model is defined independently of whatever motions the left arm exhibits during training. The T-pose motions on the other hand should be defined in terms of both the left and right arms. Table 4.1 gives a summary of the variables included in the observable output vectors for each action model.

It is important to note that these observation variables are hand-optimised to allow a good differentiation between the actions in the Glasgow corpus. For example, the left hand variables are only considered for the *TUp* and *TDown* actions. Left-handed *Knocking*, *Throwing* and *Lifting* actions are hence not supported by the current system. The full set of possible observation variables (posture, speed, body twist *etc.* of all body joints) would certainly be expressive enough to model a much bigger number of actions. However, generalising the method described here to automatically choose the best variable subset for each model in order to avoid overfitting would be an interesting challenge which is beyond the scope of this dissertation.

**Table 4.1:** HMM action model parameters. Each model uses a subset of the observation variables from the right arm (RA), left arm (LA), right leg (RL), left leg (LL), global body twist and global body speed. The table also indicates cyclic models as well as the number of states used for each model.

| Action | RA | LA | twist | RL | LL | speed | cyclic | #states |
|---|---|---|---|---|---|---|---|---|
| **Upper body actions:** | | | | | | | | |
| *TUp* | ✓ | ✓ | ✓ | | | | | 5 |
| *TDown* | ✓ | ✓ | ✓ | | | | | 5 |
| *Knocking* | ✓ | | ✓ | | | | | 9 |
| *Lifting* | ✓ | | ✓ | | | | | 9 |
| *Throwing* | ✓ | | ✓ | | | | | 9 |
| *other* | ✓ | ✓ | ✓ | | | | ✓ | 5 |
| **Lower body actions:** | | | | | | | | |
| *Walking* | | | | ✓ | ✓ | ✓ | ✓ | 5 |
| *other* | | | | ✓ | ✓ | ✓ | ✓ | 5 |

### HMM topology

We now turn to the definition of HMM state topologies. In many cases HMMs are an attractive choice because the hidden states can represent familiar concepts. For example, in speech modelling hidden states can capture the notion of phonemes. The structure of body motions has been studied far less and attempts to define concepts akin to phonemes for movements have not gone beyond early experimental stages. I therefore do not think of the hidden states as capturing a particular real-world concept — with the exception of determining the *number* of hidden states based on the subjective complexity of the action (see below). As proposed by many authors, I am using a left-to-right structure for the HMMs of acyclic actions. As Rabiner pointed out they are a natural choice for modelling signals changing over time [Rab02]. In particular, all upper body actions in the Glasgow corpus are modelled as acyclic (Table 4.1). In a left-to-right HMM the only states accessible from state $\omega_i$ are the states $\omega_i$ and $\omega_{i+1}$. Hence the transition matrix $\boldsymbol{A}$ is an upper bidiagonal matrix with $a_{i,j} = 0$ for $j \notin \{i, i+1\}$. Furthermore, the system is assumed to start in state $\omega_1$, *i.e.* $\pi_1 = 1$. Table 4.2 shows a sample transition matrix for a *Knocking* motion with 9 hidden states.

Intuitively, more complex and varied actions need to be modelled with more hidden states. In order to determine the number of hidden states for each action model, I therefore split my actions roughly into two groups: simple and complex. I am modelling simple actions (*Walking*, *TUp* and *TDown*) with 5 hidden states while the more complex actions (*Knocking*, *Throwing* and *Lifting*) are modelled with 9 hidden states. I empirically validated that changing this number of states does not improve the ability of the models to discriminate between different action categories. Table 4.1 shows the number of states used for each

**Table 4.2:** Sample transition matrix for action *Knocking*. Non-zero entries are highlighted in bold.

| | | to state | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| from state | 1 | **0.935** | **0.065** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 2 | 0.000 | **0.927** | **0.073** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 3 | 0.000 | 0.000 | **0.918** | **0.083** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 4 | 0.000 | 0.000 | 0.000 | **0.923** | **0.077** | 0.000 | 0.000 | 0.000 | 0.000 |
| | 5 | 0.000 | 0.000 | 0.000 | 0.000 | **0.969** | **0.031** | 0.000 | 0.000 | 0.000 |
| | 6 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | **0.921** | **0.079** | 0.000 | 0.000 |
| | 7 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | **0.926** | **0.074** | 0.000 |
| | 8 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | **0.930** | **0.070** |
| | 9 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | **1.000** |

action model. It also shows that *Walking* is modelled by a cyclic HMM with five states. The only difference between cyclic and acyclic models is that for a cyclic HMM with $s$ states, in general $a_{s,1} \neq 0$ and $\pi_i \neq 0$. In other words, once a full cycle of the motion has occurred, it may start again from the beginning and there is no unique starting state.

In order to complete the model definitions it is necessary to estimate the transition parameters $\boldsymbol{A}$ and the observation parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ from training data. After describing the training and classification procedures in the next section, I will demonstrate that the described parameters are suitable in Section 4.2.4.

### 4.2.3 Training and classification

Given an HMM topology I optimise the transition and observation parameters based on a set of training data observations $\mathcal{V}$. There is no known analytical method for finding the globally best set of parameters in the sense of maximising $P(\boldsymbol{V}|\lambda)$. However, the Baum-Welch procedure is an iterative method which can find a good locally maximised set of parameters [DHS00]. It uses an Expectation-Maximisation approach which starts with an initial guess of parameters. The procedure iteratively computes the expected number of state transitions from and between all states based on the current set of parameters and training data observations. In the maximisation step it finds an improved set of parameters based on those transition statistics. Baum proved that following this procedure results in convergence towards a locally maximal set of parameters. Clearly, the choice of initialisation parameters determines which local maxima in the often complex optimisation landscape are reachable. In order to increase the chance of finding a *globally* maximal solution, one can use multiple random initialisations. For left-to-right HMMs I am using an initialisation technique known as Viterbi initialisation [YKO$^+$00, NSHU$^+$94]. It initially assumes a temporally uniform assignment of states to observations and recomputes the parameters a number of times based on statistics calculated from decoding the observations using the Viterbi algorithm.

Decoding via the Viterbi algorithm is one way of quantifying the likelihood of a certain model producing a given observation sequence. In particular, it finds

$$\boldsymbol{\omega}^* = \arg\max_{\omega} P(\boldsymbol{\omega}, \boldsymbol{V}|\lambda). \tag{4.3}$$

That is, it finds the most likely sequence of states to have produced the observation sequence. Once the optimising state sequence is known $\boldsymbol{\omega}^*$, it is easy to also find the probability associated with that state path. In fact, the Viterbi algorithm uses dynamic programming in such a way that both the maximising state sequence and its likelihood are always computed. Given a way of computing $P(\boldsymbol{\omega}^*, \boldsymbol{V}|\lambda_c)$ I can *classify* sequences into action categories by computing

$$c^* = \arg\max_{c\in\mathcal{C}} P(\boldsymbol{\omega}^*, \boldsymbol{V}|\lambda_c). \tag{4.4}$$

Another commonly used way of classifying a given observation sequence is to consider the *total* likelihood of a certain model producing the sequence $P(\boldsymbol{V}|\lambda)$. This quantity can be found using the dynamic programming-based Forward algorithm. Classification then proceeds by finding $\arg\max_{c\in\mathcal{C}} P(\boldsymbol{V}|\lambda_c)$. For my purposes I will be using the Viterbi likelihood rather than the Forward likelihood. Above all, it gives rise to a more efficient solution in Chapter 6 where I am using Level Building to connect my individually trained HMMs to parse and classify *sequences* of actions. Using the Viterbi likelihood allows me to solve the segmentation problem using an efficient dynamic programming algorithm (*c.f.* Section 6.2.2), while using the Forward likelihood would lead to an exponential growth in complexity.

### Likelihood normalisation

In order to find the best model during classification, Viterbi likelihood values need to be compared. Care needs to be taken if those values were produced by HMMs with output vectors of different dimensionality. It is important to note that for the computation of likelihoods $P(\boldsymbol{\omega}^*, \boldsymbol{V}|\lambda_c)$, the Viterbi algorithm uses the observation probability densities in order to get a point measure for the likelihood of a certain (continuous and vector-based) observation. In general, these density measures will get smaller with an increasing number of observation dimensions. For classification this means that without an appropriate normalisation, classifications would always be biased towards models with lower-dimensional observations.

Assume that models $\lambda_1 \dots \lambda_n$ have observation vectors of dimensionality $d_1 \dots d_n$ respectively with $d_i \leqslant d_{i+1}$ such that $d_{max} = d_n$. The Viterbi algorithm would calculate the unnormalised likelihoods of the $i$th observation in a sequence as

$$L_{\lambda_m} = N(\boldsymbol{\phi}^i_{\lambda_m}, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m). \tag{4.5}$$

for models $\lambda_m$, where $\boldsymbol{\phi}_{\lambda_m}$ denotes the $d_m$-dimensional representation of the $i$th symbol in the feature space of model $\lambda_m$. Classifications based on these likelihoods would be biased towards $\lambda_1$ and against $\lambda_n$. In order to normalise the likelihoods, I am introducing *pseudo variables* into each model during the calculation of the Viterbi likelihoods. The number of variables introduced depends on the number of output dimensions originally associated with the model. In particular, the number of variables introduced into model $\lambda_m$ is $d_{max} - d_m$. When computing the normalised likelihoods, each of the virtual variables is defined to have a marginal observation likelihood of

$$\tilde{L}_{\lambda_m} = L_{\lambda_m}^{1/d_m}, \tag{4.6}$$

*i.e.* the same likelihood as one of the original observation variables assuming independence. To compute the overall normalised observation likelihood $\hat{L}$, the original likelihood is combined with the likelihood of the virtual variables assuming independence:

$$\begin{aligned}\hat{L}_{\lambda_m} &= L_{\lambda_m} \times \tilde{L}_{\lambda_m}^{\frac{d_{max}-d_m}{d_m}} \\ &= L_{\lambda_m}^{\frac{d_{max}}{d_m}} \end{aligned} \tag{4.7}$$

Note that this normalisation will not affect the computed Viterbi path itself but only the Viterbi likelihood. I can now apply the training and classification techniques to the isolated actions in the Glasgow corpus.

### 4.2.4 Action classification results

In order to verify the effectiveness of the HMM-based classification approach, I trained and tested the 8 HMMs as summarised in Table 4.1 on the isolated Glasgow motion data. I am aiming to answer three questions:

1. How well do trained models generalise to new action samples from the training subjects?
2. How well do trained models generalise to actions from new subjects?
3. Are the features effective for modelling and distinguishing the actions?

I designed two experiments to answer the questions. In order to answer Question 1, Experiment 1 uses 10% of the motion samples randomly selected from the 30 subjects to

**Table 4.3:** HMM-based action recognition performance (mean and standard deviation) for actions *TUp* (TU), *TDown* (TD), *Knocking* (Kn), *Throwing* (Th), *Lifting* (Li), *Walking* (Wa) in Experiment 1 (testing on subject seen during training). Upper and lower body actions are shown as separate sections which include the *other* (ot) action categories.

| | classified as | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Truth | TU | TD | Kn | Th | Li | ot | Wa | ot |
| TU | **1.00**<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 |
| TD | 0.00<br>±0.00 | **0.99**<br>±0.00 | 0.00<br>±0.00 | 0.01<br>±0.00 | 0.00<br>±0.00 | 0.01<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 |
| Kn | 0.00<br>±0.02 | 0.00<br>±0.00 | **0.99**<br>±0.01 | 0.01<br>±0.01 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 |
| Th | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | **1.00**<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 |
| Li | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.01<br>±0.02 | **0.99**<br>±0.00 | 0.00<br>±0.02 | 0.00<br>±0.00 | 0.00<br>±0.00 |
| ot | 0.01<br>±0.02 | 0.00<br>±0.02 | 0.00<br>±0.01 | 0.04<br>±0.00 | 0.00<br>±0.01 | **0.95**<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 |
| Wa | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | **1.00**<br>±0.00 | 0.00<br>±0.00 |
| ot | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | **1.00**<br>±0.00 |

train the models. Experiment 2 is designed to answer Question 2 and uses 10% of the motion samples, but selected from only three distinct subjects to train the models. Hence, the number of training samples were the same for both experiments, but the coverage of subjects was significantly different. Each experiment was run 10 times with different training samples. For each iteration I classified the remaining 90% of the samples in both setups. I used custom Matlab code and Murphy's HMM toolbox [Mur] to run the experiments.

Tables 4.3 and 4.4 give the recognition results in the form of confusion matrices for Experiments 1 and 2 respectively. In both cases, the diagonals dominate. Most previously unseen samples have been classified correctly. In answer to Question 3, this suggests that the features I described in Section 4.2.2 are able to capture the fundamental differences between the considered actions. If we inspect Tables 4.3 and 4.4 in more detail, we see an interesting difference. The classification errors observed in Table 4.3 are minimal. The models were able to discriminate between the actions almost perfectly. The picture changes if we look at Table 4.4. When the models have to classify actions after being trained on samples from a small population, performance drops significantly for some action categories. In Experiment 2 the average recognition rates are lower and the variance across iterations is larger than in Experiment 1. The effect is pronounced for *Lifting* actions which are interpreted as *Throwing* actions 20% of the time. This is because the

**Table 4.4:** HMM-based action recognition performance (mean and standard deviation) for actions *TUp* (TU), *TDown* (TD), *Knocking* (Kn), *Throwing* (Th), *Lifting* (Li), *Walking* (Wa) in Experiment 2 (generalisation to new subjects). Upper and lower body actions are shown as separate sections which include the *other* (ot) action categories.

| | classified as | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Truth | TU | TD | Kn | Th | Li | ot | Wa | ot |
| TU | **0.99**<br>±0.02 | 0.01<br>±0.02 | 0.00<br>±0.01 | 0.00<br>±0.00 | 0.00<br>±0.01 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 |
| TD | 0.01<br>±0.01 | **0.97**<br>±0.02 | 0.01<br>±0.02 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.01<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 |
| Kn | 0.00<br>±0.00 | 0.00<br>±0.00 | **0.94**<br>±0.09 | 0.06<br>±0.09 | 0.00<br>±0.00 | 0.00<br>±0.01 | 0.00<br>±0.00 | 0.00<br>±0.00 |
| Th | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.01<br>±0.02 | **0.93**<br>±0.10 | 0.00<br>±0.00 | 0.06<br>±0.09 | 0.00<br>±0.00 | 0.00<br>±0.00 |
| Li | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.02<br>±0.01 | 0.07<br>±0.04 | **0.80**<br>±0.10 | 0.13<br>±0.12 | 0.00<br>±0.00 | 0.00<br>±0.00 |
| ot | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.01<br>±0.02 | 0.00<br>±0.00 | **0.99**<br>±0.02 | 0.00<br>±0.00 | 0.00<br>±0.00 |
| Wa | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | **1.00**<br>±0.00 | 0.00<br>±0.00 |
| ot | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.00<br>±0.00 | 0.29<br>±0.17 | **0.71**<br>±0.17 |

*Lifting* model overfitted to the few subjects that it was trained on. This occurs because *Lifting* actions are very simple. *Throwing* actions on the other hand are more complex and the trained *Throwing* models therefore sometimes fit previously unseen *Lifting* data better than the overfitted *Lifting* models. The same applies to the *other* lower body actions which, in the Glasgow corpus, are significantly less complex than the *Walking* actions.

Experiments 1 and 2 highlight that while HMMs are a perfectly suitable technique for distinguishing different isolated actions from each other, there are some caveats which must be kept in mind to make them most effective. One domain-specific constraint is that I am likely to deal with actions and hence models of different complexity. My results show that in those cases it is essential to ensure that the variation in the training data is as representative as possible in order to avoid overfitting. In part, the confusion problems are also due to the generative nature of my HMM training approach. The Baum-Welch procedure focuses on modelling the *generation* of a particular category of signal through Expectation Maximisation rather than building a model that *distinguishes* it from different categories. Because the differences in appearance between the different action categories are generally large, it is not necessary to resort to discriminative training methods in this context. However, when we turn to distinguishing different emotion classes in Chapters 5 and 6, I will make use of a discriminative approach which is concerned with modelling the differences between classes rather than the generation of data from a particular class.

### 4.2.5 Related work

The classification of time series data into distinct categories is a very common pattern recognition problem. In his seminal paper, Rabiner describes a way of representing speech signals using HMMs with continuous observation vectors [Rab02]. This is now standard practice among speech recognition researchers. When computer vision research became popular, researchers quickly adopted HMMs in other fields such as body gesture recognition [WB99, WH99], automatic sign language interpretation [SP95] and activity analysis [MSSS04, OHG02]. Earlier approaches made use of static pattern matching and dynamic time warping [WH99], but the field has been dominated by the advantages of HMMs and more general statistical frameworks such as Dynamic Bayesian networks (*e.g.* see [DCXL06, KR05]).

The modelling of the human body as independent regions for the purpose of action analysis has been proposed by several authors in the past. Lv and Nevatia present a framework for modelling arm, head and torso-based actions independently [LN06]. They also label actions by three additional categories: stationary (such as sitting), transitional (such as getting up) and periodic (such as waving a hand). Finally, they use HMMs and boosting to classify 3D joint trajectories into action categories. A structural body model is also proposed by Vacek *et al.* [VKD05]. They define a tree-based description of multiple levels ranging from "General Body" to "Hand". Activities are described in terms of the nodes of the tree depending on which body parts are involved. Interestingly, they also define additional context groups like "Objects" and "Places". From a recognition standpoint, having knowledge about this kind of context would help enormously in distinguishing different actions. In the case of the Glasgow corpus, however, there is no explicit context information. The role of context in activity recognition has also been discussed by other authors. In fact, Bobick uses the idea of context to draw clear distinctions between movements, activities and actions [Bob97]. He defines activities in terms of sequences of individual movements which can be recognised readily without context information. For a higher level understanding of actions, on the other hand, one requires information about the environment and larger temporal relationships. My work does not aim to recognise actions in Bobick's sense. For me it will be enough to have a basic understanding of the motion from which emotions are to be deduced.

On top of independent body parts, my action models also account for actions defined in terms of different feature sets. My approach has been inspired by Vogler and Metaxas' work on parallel HMMs for sign language recognition [VM01, VM99]. They model the left and right hands as two independent HMMs whose outputs are combined. Similarly to actions defined by either one or two hands (*TDown* versus *Knocking*), they have to deal with two-handed *vs.* one-handed signs. In order to make HMM probabilities comparable, they propose to assign the probability from the active hand to that of the inactive hand for one-handed signs. My approach is a generalisation which allows a fully flexible definition of feature vectors for each model. The output probabilities are then normalised in proportion to the relative number of dimensions as defined in Equation 4.7.

Action recognition as tackled in this chapter has received a lot of attention from the computer vision community. Building generative models for a set of predefined action categories has been a well-studied problem [Gav99, Ced95]. More recently the community has focused on problems with additional complexities. Probabilistic models related to HMMs are still commonly used however. Peursum *et al.* [PVW07] use factored state hierarchical HMMs to model both action phases and joint rotations as hidden states to be inferred from multiple camera views. Clearly, the additional vision-related problems such as occlusions and camera projection lead to a more complex model structure and inference can hence only be carried out using expensive particle filtering. They also do not model the internal body state by variables such as movement speed and twist. A similar approach is followed by Weinland *et al.* [WBR07]. They also have a hierarchical HMM structure where 3D body state is modelled as a second hidden layer. Their body model is, however, based on a small set of exemplars per hidden action state rather than a factored set of joint angles. Lv *et al.* also use the same idea and represent each action by a set of 3D key poses [LN07]. Modelling actions by a small number of discrete key poses rather than temporally continuous distributions as my HMMs has the advantages of being less computationally expensive during the recognition phase and being inherently robust to certain kinds of stylistic differences. Of course, key poses need to be picked carefully and the amount of detail retained in them is dependent on what and how many actions the system needs to distinguish between.

Finally, Turaga *et al.* describe a solution to the slightly different problem of action mining from video data [TVC07]. They model actions as cascades of smaller action elements, each element modelled in turn by a linear dynamical system. They decide not to use Markov Models for connecting their action elements because of the inflexibility of state duration distributions – a problem I will come back to in Section 6.2.2. Their focus is explicitly on discovering action patterns and the recognition ability is hence not as accurate. Their framework is nevertheless very interesting for scenarios where it is impractical to train models for each action in advance. Their notion of action elements is very similar to the idea of motion primitives which I will discuss in the remainder of this chapter.

## 4.3 Sub-motion modelling

It has been my goal from the outset to make emotion recognition action-aware. The first step has been achieved in the first part of this chapter. By recognising distinct actions

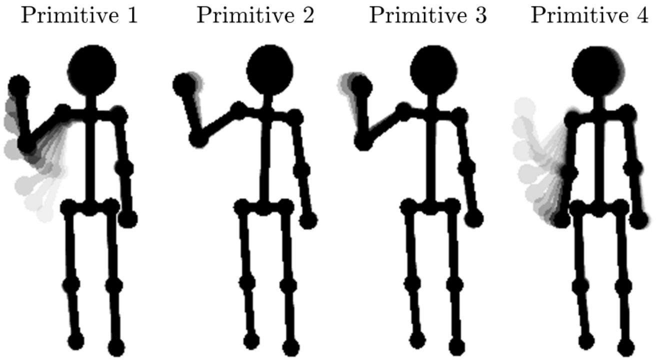

**Figure 4.1:** Four phases of a *Knocking* motion exhibiting distinct peaks of motion energy. The clustering-based algorithm detects each of the phases as a separate motion segment. Each segment is labelled with one of four automatically derived motion primitives. The primitives coincide with the semantically meaningful basic actions “Raise arm”, “Knock”, “Retract”, “Lower arm”.

I can train different emotion classifiers on each. I will be able to use knowledge such as: *Walking* always exhibits leg movement whereas a stamping foot during a *Throwing* action might actually indicate something about the underlying emotion. These are global observations. But is it possible to do better? I am going to analyse movements not only at the action level, but look deeper into their structure. This might allow me to capture structural differences between emotional effects which would be obscured if I only considered actions at a global level.

This section therefore presents two ways to segment actions into fundamental movements, which I call *primitives*. The goal is to define and extract the motion primitives for complex actions. The exact definition of a motion primitive will be part of the method description in each case. Whatever the exact definition, however, I think of a primitive as a section of movement which appears repeatedly throughout a number of complex actions, such as an arm raise or a walking step. In Chapters 5 and 6 I will extract emotional features from these primitives as well as from the action as a whole.

### 4.3.1 Clustering regions of high motion energy

The first approach to defining motion primitives is purely data-driven and works by processing the motion signal in a bottom-up fashion. It assumes that motions can be modelled as a composition of strokes - segments of approximately straight line motions. Let us consider a couple of examples from the Glasgow corpus.

A *Knocking* action can be decomposed into a sequence of strokes as follows (see Figure 4.1): arm raise, hand forward, hand backward, hand forward, ..., arm lower. Similarly, *Lifting* actions can be decomposed into: arm reach, lift object, replace object, arm lower.

The common property which marks the end of a stroke in these cases is a significant

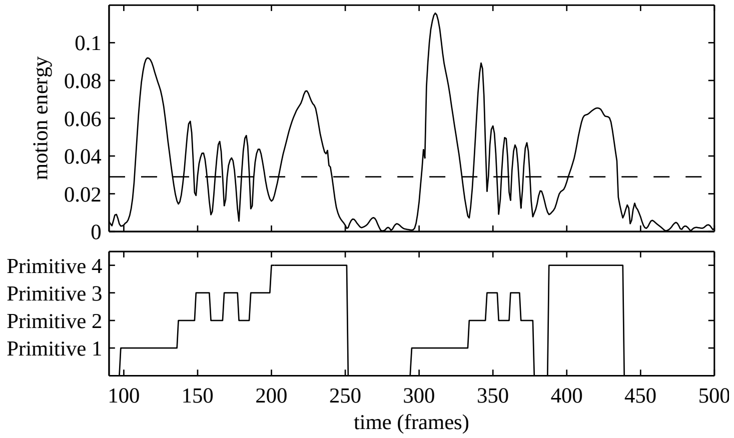

**Figure 4.2:** Objective function $E(t)$ (top) with automatically calculated optimal segmentation threshold $\tau_{opt} = 0.029$ for part of a repeated knocking motion separated by periods of no motion. The bottom shows the parse of this motion sequence into four motion primitives and periods of no motion. Note that one potential primitive is missed around frame 380 because its peak energy is below $\tau_{opt}$.

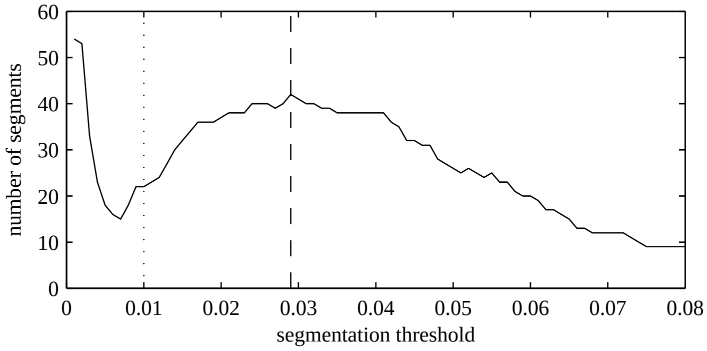

**Figure 4.3:** $numseg_E(\tau)$ for a repeated knocking motion sampled for thresholds between 0.001 and 0.08. The diagram also shows $\tau_{opt}$ (dashed) and $\tau_0$ (dotted).

decrease in hand speed. For example, when the hand reaches the door during a knock it stops or slows down significantly before it is retracted. Hence, I define motion primitives as periods of high motion energy, separated by local energy minima.

In order to capture this notion mathematically, I define an objective function $E(t)$ which is a measure for the overall motion energy at time frame $t$. In many ways this concept of energy is analogous to that employed for the segmentation of speech into phonemes or words [WLZ03] and can also be found in work by Fod *et al.* [FMJ02]. Let $\dot{\theta}_{t,g}$ denote the speed of the $g$th joint at time frame $t$. Then I can define the body's total motion energy as a weighted sum of squared joint speeds.

$$E(t) = \sum_{k=1}^{n} w_k \dot{\theta}_{t,k}^2 \tag{4.8}$$

The weights $w_k$ can be used to give different weightings to different body joints. In the work I used $w_k = 1$ for all $k$. $E$ will be large for periods of energetic motion and will remain small during periods of low motion energy. Figure 4.2 shows $E$ for repeated *Knocking* actions. The figure shows two repetitions separated by a period of no motion. As expected, the observed energy peaks coincide with primitive actions such as arm raises or individual forward and backward movements during the knock. Furthermore, periods of no activity are marked by distinctively low energy. Local minima in $E$ can be observed whenever the trajectory of the right arm changes direction. As the next step I find the local minima in $E$ as follows.

1. Compute $E$ for the whole motion sequence.
2. Threshold the signal at a threshold $\tau$. Mark all frames $t$ for which $E(t) > \tau$.
3. Find all connected regions of marked frames and regard them as individual motion segments.
4. Extend the segments to the preceding and succeeding local minima of $E$.

Obviously, the choice of $\tau$ has a major impact on the number and nature of segments. I could simply use an empirically derived threshold. If this method is to generalise to other actions, however, it is necessary to devise an automatic way of finding an optimal $\tau$. The following algorithm finds a threshold which will exhibit all major motion segments (energy peaks) while filtering out small-scale motions due to low-level signal noise. For every pair $(E, \tau_n)$ I obtain a number of segments by thresholding $E$ at $\tau_n$. Let $numseg_E(\tau_n)$ be the function which computes the number of segments for any such pair. Figure 4.3 shows *numseg* for the motion in Figure 4.2 and sampled at various thresholds. We see that noise is mainly registered during periods of low energy (e.g. between frames 250–300 and 450–500 in Figure 4.2). Let $\tau_0$ be an empirical noise threshold. Then the optimal threshold $\tau_{opt}$ is defined as the threshold which maximises the number of major motion segments.

$$\tau_{opt} = \arg\max_{\tau}\{numseg_E(\tau)\} \qquad \text{subject to } \tau_{opt} > \tau_0. \tag{4.9}$$

Ideally, I would like to group the extracted segments into semantically meaningful clusters representing similar primitive motions. For example, all individual knocking segments should be associated with each other. One approach to define such primitives would be to use a comprehensive list as devised by Bull or Birdwhistell to transcribe their psychological or anthropological observations [Bul87, Bir70]. Due to their generality, however, these sets are large. Many of the listed primitives are irrelevant for any particular context. Indeed, context often governs the *affective and social meaning* of movements [Bir70]. I therefore adopt a more context-dependent approach to the definition of motion primitives. It is based on the clustering of a set of segments derived from example motions which are representative for a certain context. For individual actions the number of motion primitives is therefore rather small. It would be possible, however, define primitives which span more complex contexts such as “every-day activities” or “interpersonal conversations”.

### Definition of cluster-based primitives

Having segmented all the knocking movements from the Glasgow database as described above, I need to find a representation for the segments which allows me to compare and cluster them. I therefore consider the time series of joint positions of the extracted segments and time-normalise them. This is done by resampling each segment at 25 equally spaced intervals. I also subtract the segments’ means in order to capture the *relative* motion rather than the absolute body configurations. Next, I wish to group the segments into semantically distinct categories. From the observations at the beginning of this section I hypothesise that the *Knocking* actions can be divided into four basic phases: lift arm, repeatedly knock and retract, lower arm. I can therefore use a $k$-means clustering algorithm with $k = 4$. In a completely unsupervised scenario without any prior knowledge of the number of motion primitives, one could choose a clustering technique which can automatically determines an optimal number of clusters such as hierarchical [Joh67] or Markov clustering [vD00]. A further exploration of this idea is an area for future work. The following steps summarise the algorithm to compute a set of motion primitives from a set of example motions:

1. Segment the set of motions based on the objective function $E(t)$.

2. Time-normalise all segments by linear resampling. Subtract sample means.

3. Cluster the normalised segments using the $k$-means algorithm.

4. The clusters (or cluster centroids) represent the motion primitives.

The four derived motion primitives for the set of *Knocking* motions are visualised in Figure 4.4. All primitives have been time-normalised to 25 frames and each plot shows the four rotational degrees of freedom of the right arm (3 shoulder + 1 elbow). Having defined the primitives, a new motion can now be *parsed* by following steps 1 and 2 as outlined above and replacing steps 3 and 4 by an assignment to the closest cluster centroid (most similar primitive) using the nearest-neighbour approach and a Euclidean distance metric.

### Evaluation

Figure 4.2 illustrates how a repeated *Knocking* motion (energy curve shown on top) has been parsed into a sequence of primitives (bottom). The plot confirms that the parse is

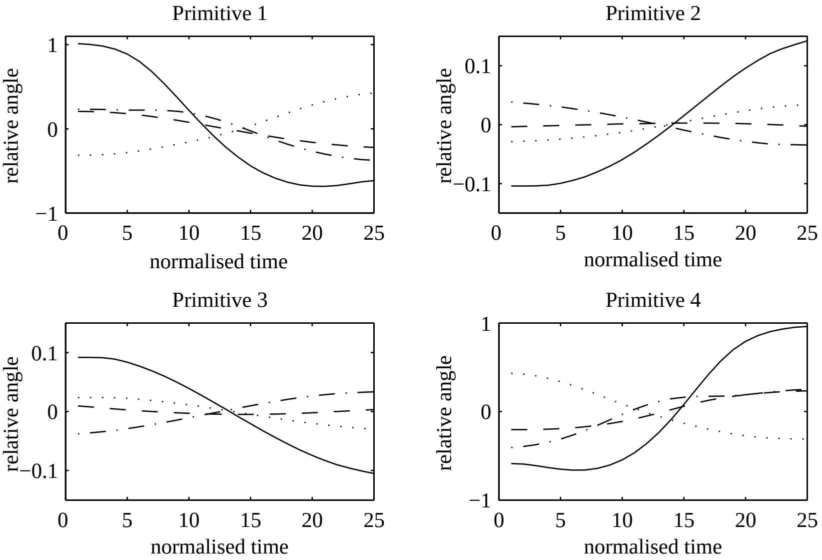

**Figure 4.4:** Four motion primitives derived by $k$-means clustering. Different degrees of freedom of the right arm are represented as separate curves. In each case the segments' means have been subtracted to arrive at a relative encoding.

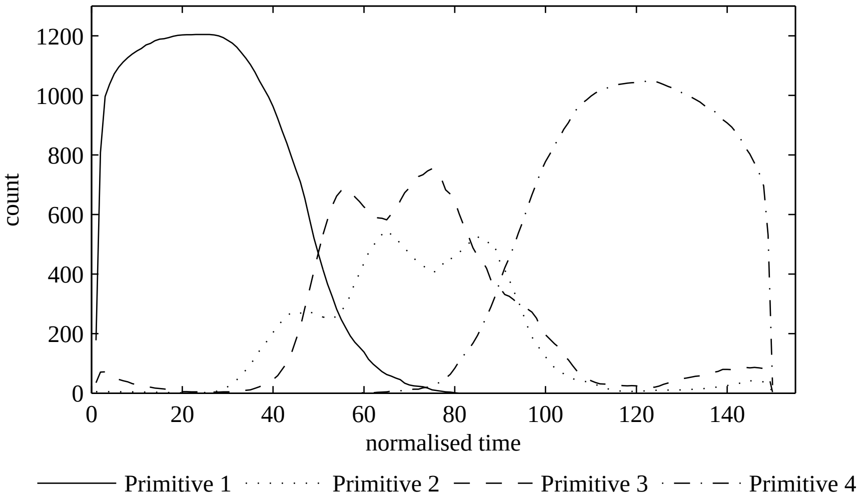

**Figure 4.5:** Distribution of motion primitives after parsing 1200 *Knocking* motions. All parses were time-normalised to 150 frames.

consistent with my initial expectations. In particular, Primitives 1 and 4 correspond to the larger scale motions of raising and lowering the right arm while Primitives 2 and 3 tend to capture the smaller scale knocking motions.

In order to evaluate the approach at a larger scale I parsed all 1200 *Knocking* samples in the Glasgow corpus. In order to visualise the parses I time-normalised them to 150 frames by linear sampling and interpolation. Figure 4.5 shows the distributions of primitives at each frame. We can clearly see that for the vast majority of samples, Primitives 1 and 4 were detected correctly at the beginning and end respectively. As expected, the smaller-scale knocking primitives are occupying the central region of the graph. We can also see the expected alternations between Primitives 2 and 3.

For a more complete quantitative evaluation see Appendix C.

#### Shortcomings

In [BR07] I applied this technique very successfully to isolated *Knocking* motions and went on to demonstrate how to recognise emotions on their basis. In a more general context, however, the method has several shortcomings:

1. Most importantly, the method provides no mechanism to allow user-defined motion primitives beyond specifying the number of cluster centroids. In many circumstances prior knowledge may be available which could usefully guide the definition. In this dissertation, I am primarily interested in the emotion-communicating qualities of the extracted primitives. Human experts might have a good intuition about which elements of an action communicate emotions particularly well. For example, it is probably more robust to consider the knocking phase (consisting of repeated forward and backward motions of the hand) as one primitive. The current method does not allow for this kind of top-down knowledge to be incorporated.

2. The segmentation method heavily relies on the existence of distinct energy minima in $E(t)$, detectable using the global threshold $\tau_{opt}$. In some cases the minimum-based definitions may result in more primitives than I want, while in other cases it may miss sections of movements which seem important but do not have distinctive enough energy maxima. We already encountered an example of this problem for *Knocking* actions in Figure 4.2. As another example, most *Throwing* motions exhibit a recoil phase after the throw, which is not usually separated from the main throw by a distinct energy minimum. This problem tends to be amplified when actions happen in sequence (Figure 4.6).

3. The clustering approach to defining motion primitives does not make use of temporal information. Primitive extraction may be more robust if the method exploits high-level knowledge such as “arm raise” primitives always appear at the *beginning* of *Knocking* and *Throwing* actions.

In the following section I will present a second algorithm which addresses these problems.

### 4.3.2 Manual definition based on multiple sequence alignment

Following the problem descriptions in Section 4.3.1 the motivations for the second approach are two-fold. Firstly, I would like to make it as easy as possible to exploit human intelligence in defining motion primitives. Two usability-related requirements guided the design of this algorithm:

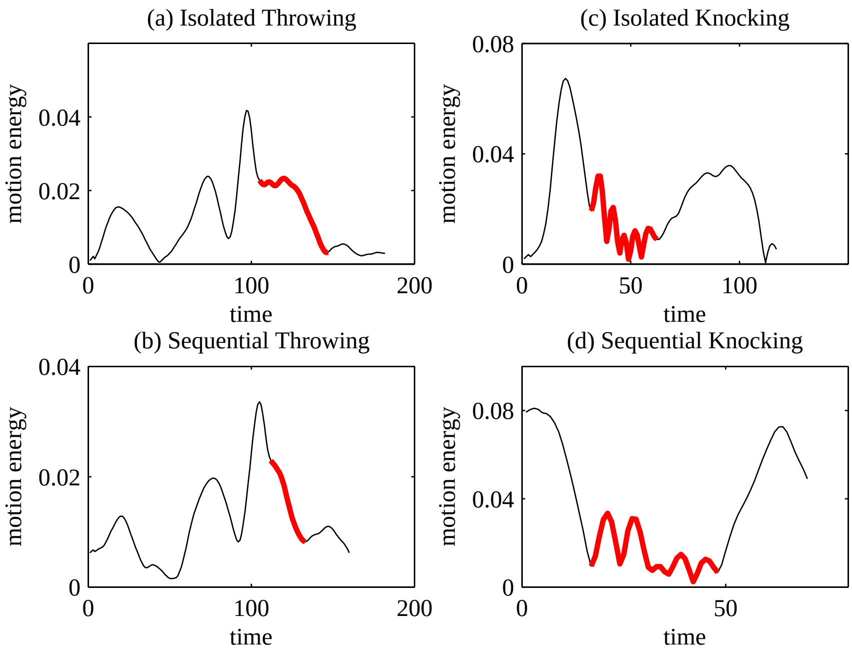

**Figure 4.6:** Problematic motion segments in *Throwing* and *Knocking* examples. The regions highlighted in red are interesting candidates for motion primitives but are difficult to extract using a global energy threshold. Graphs (b) and (d) are taken from action *sequences* for which a threshold-based analysis is generally harder as they exhibit smoother energy profiles with less distinct minima.

1. It should be easy for a human to define motion primitives without the need to acquire any technical knowledge about the underlying algorithms.
2. The time needed by a human to define the primitives should be as small as possible.

Secondly, I am aiming to define primitives in the context of their structural significance in an action. In the previous approach, primitives were computed in a solely bottom-up fashion based on their appearance - first as periods of high motion energy and then by clustering them. The new approach has to fulfill the following algorithmic requirements:

1. Primitives are no longer defined in terms of energy minima and maxima.
2. It should be possible to specify primitives as arbitrary motion segments.
3. The algorithm must be able to extract these segments from new motions automatically.

The approach which I developed first aligns a number of example motions in time. It then computes a combined representation of the motions as a single, combined time series. This aligned representation of all example motions serves as the *master description* of the action. The time series is visualised in a human-understandable form. I subsequently define motion primitives by picking regions from the master description. The chosen primitives can then be translated back to segments of the original signals. For extracting primitives from a new sequence, I can then apply top-down reasoning by first aligning it to the master description and then picking the regions which accord with the previously defined primitives. Note that the alignment of motions is not immediately meaningful for cyclic actions such as walking. I am therefore focusing on the three non-cyclic actions *Knocking*, *Lifting* and *Throwing*. I will discuss the analysis of cyclic actions under this scheme in Section 4.3.3.

### Dynamic Time Warping

My approach relies on the ability to align two or more time series in time. I am basing my solution on Dynamic Time Warping (DTW) [SC78]. DTW and similar dynamic programming-based warping algorithms have been used extensively in the past for aligning time series like speech [MR81b, MR81a] and gestures [DP93] as well as other sequential data such as DNA sequences [NW70]. In its original form DTW is used to find an alignment between two data series $\boldsymbol{V}_1^{T_1} = \{\boldsymbol{v}_1(1), \ldots, \boldsymbol{v}_1(T_1)\}$ and $\boldsymbol{V}_2^{T_2} = \{\boldsymbol{v}_2(1), \ldots, \boldsymbol{v}_2(T_2)\}$. As before, $\boldsymbol{v}(t)$ is a feature vector as observed at time frame $t$. For alignment, the feature vectors consist of dimensions capturing joint positions and velocities. The goal of the alignment of two sequences is to account for the non-linear timing variations in signals $\boldsymbol{V}_1^{T_1}$ and $\boldsymbol{V}_2^{T_2}$. This is expressed through a time warp function $W(x)$ which associates samples of sequence $\boldsymbol{V}_1^{T_1}$ with samples of sequence $\boldsymbol{V}_2^{T_2}$, for example

$$\begin{aligned}
W(1) &= (\boldsymbol{v}_1(1), \boldsymbol{v}_2(1)) \\
W(2) &= (\boldsymbol{v}_1(1), \boldsymbol{v}_2(2)) \\
&\ldots \\
W(x) &= (\boldsymbol{v}_1(i), \boldsymbol{v}_2(j)) \qquad (4.10) \\
&\ldots \\
W(\tilde{T}) &= (\boldsymbol{v}_1(T_1), \boldsymbol{v}_2(T_2))
\end{aligned}$$

Hence, $W(x)$ defines the time-warped sequences $\tilde{\boldsymbol{V}}_1^{\tilde{T}}$ and $\tilde{\boldsymbol{V}}_2^{\tilde{T}}$ as

$$\tilde{\boldsymbol{v}}_1(x) = W(x)[1] \qquad (4.11)$$

$$\tilde{\boldsymbol{v}}_2(x) = W(x)[2] \qquad (4.12)$$

Note that $\tilde{\boldsymbol{V}}_1^{\tilde{T}}$ and $\tilde{\boldsymbol{V}}_2^{\tilde{T}}$ now have the same length $\tilde{T}$. The quality of the alignment of two time frames can be quantified through a distance function $\delta(\boldsymbol{v}_1, \boldsymbol{v}_2)$. I am using the normalised Euclidean distance, or Mahalanobis distance

$$\delta(\boldsymbol{v}_1, \boldsymbol{v}_2) = \left( \sum_{i=1}^{N} \frac{(x_i - y_i)^2}{\sigma_i^2} \right)^{\frac{1}{2}}. \tag{4.13}$$

The total alignment distance $\Delta$ attributed to a particular warp function $W$ is calculated as a sum of individual frame-wise distances.

$$\Delta(W) = \sum_{t=1}^{\tilde{T}} \delta(W(t)). \tag{4.14}$$

DTW uses an efficient dynamic programming solution to find the alignment $W^*$ which minimises the total alignment distance $\Delta(W^*)$ and hence represents the best alignment. The basic DTW algorithm to align two sequences has the following inputs and outputs:

- **Inputs**
  - a distance function $\delta(\boldsymbol{v}_1, \boldsymbol{v}_2)$: calculates the distance between two frames of observation $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$
  - two sequences $\boldsymbol{V}_1^{T_1}$, $\boldsymbol{V}_2^{T_2}$.
- **Output**
  - $W^*$: the best alignment of $\boldsymbol{V}_1^{T_1}$ and $\boldsymbol{V}_2^{T_2}$.

Figure 4.7 shows the alignment results $W^*(x)$ of pairs of *Knocking*, *Throwing* and *Lifting* sequences in red. Figure 4.7 (left) plots the coarsely sampled correspondences between the two time series. Note that the plot only shows the aggregate motion energy while the underlying DTW algorithm aligns the sequences based on the multi-dimensional representation $\boldsymbol{V}$. Figure 4.7 (right) visualises $W^*(x)$ as a path through the dynamic programming matrix. The examples highlight some of the strengths of DTW. In the *Knocking* sequences, we see that although the knocks appear at different times in the sequence and with different intensities, the individual forward and backward hand movements have been aligned very accurately. The *Throwing* example highlights a particular strength of this approach. In one of the two sequences, the throwing segment is very strong and powerful, resulting in a major energy peak between frames 90 and 110. In the second sequence, the throwing segment is almost unnoticeable. Inspecting the original recordings reveals that the subject threw the object using a wrist-flick, resulting in a flat energy region. However, despite this difference, the DTW algorithm has aligned the two throwing segments perfectly — an association which would have been impossible using the previous method. These examples highlight how DTW can exploit global action structure to associate semantically equivalent regions in action sequences which differ in appearance.

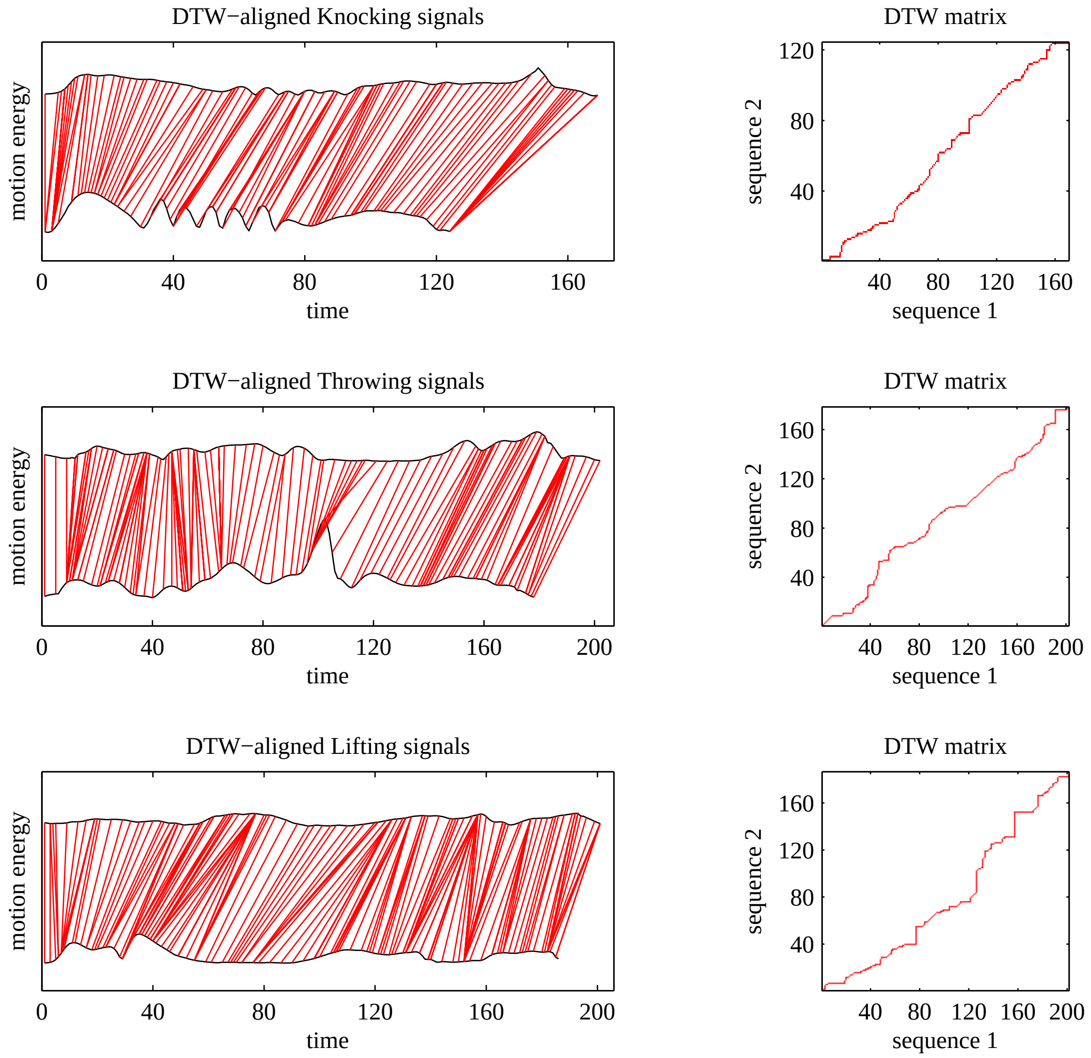

**Figure 4.7:** Results of aligning two sequences produced by different subjects and with different emotions. Left: frame-wise correspondences $W^*(t)$ sampled at regular time intervals. Right: full visualisation of $W^*$ as a trail through the alignment matrix.

### Aligning multiple sequences

Trying to align more than two sequences through dynamic programming soon becomes intractable. A common approximation is repeated pair-wise alignment [HS88]. For example, if I am trying to align 4 sequences $\boldsymbol{V}_1 \ldots \boldsymbol{V}_4$, I can first align $\boldsymbol{V}_1$, $\boldsymbol{V}_2$ and $\boldsymbol{V}_3$, $\boldsymbol{V}_4$ independently using DTW. This produces two compound sequences $\boldsymbol{V}_{12}$ and $\boldsymbol{V}_{34}$. I then align $\boldsymbol{V}_{12}$ and $\boldsymbol{V}_{34}$ using an adapted version of the original pairwise DTW algorithm. In my implementation I represent compound sequences such as $\boldsymbol{V}_{12}$ as sets. I therefore generalise the DTW algorithm to work on *sets* of sequences rather than individual sequences as follows:

- **Inputs**
  - a distance function $\delta_S(\boldsymbol{v}_1, \boldsymbol{v}_2)$: calculates the distance between two sets of

observation frames $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$.

  – two sets of sequences $\mathcal{V}_1$ and $\mathcal{V}_2$.

- **Output**

  – $W^*$: the best alignment of $\mathcal{V}_1$ and $\mathcal{V}_2$.

It is necessary for all sequences within a set $\mathcal{V}$ to have the same length. This allows me to index $\mathcal{V}$ uniquely by frame as is required by the DTW algorithm. Note that sequences derived through a number of pairwise DTW alignments will always be of the same length and hence repeated alignment will always produce sets $\mathcal{V}_1$ and $\mathcal{V}_2$ which satisfy the requirement. Of course, elements from set $\mathcal{V}_1$ do not need to have the same length as elements from set $\mathcal{V}_2$ in order to be able to align them.

There are several options to define a $\delta_S$ which generalises the original distance function $\delta$. The most accurate solution to finding the distance between two sets of points might be to calculate the average of all pairwise distances between the set of elements:

$$\delta_S(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{1}{|\boldsymbol{v}_1||\boldsymbol{v}_2|} \sum_{\boldsymbol{v}' \in \boldsymbol{v}_1} \sum_{\boldsymbol{v}'' \in \boldsymbol{v}_2} \delta(\boldsymbol{v}', \boldsymbol{v}'') \tag{4.15}$$

Computing this distance has complexity $O(n^2)$ and in practice becomes too slow for aligning large numbers of sequences. I am therefore using the distance between the set averages which can be computed in $O(n)$ time:

$$\delta_S(\boldsymbol{v}_1, \boldsymbol{v}_2) = \delta(\frac{1}{|\boldsymbol{v}_1|} \sum_{\boldsymbol{v}' \in \boldsymbol{v}_1} v', \frac{1}{|\boldsymbol{v}_2|} \sum_{\boldsymbol{v}' \in \boldsymbol{v}_2} v') \tag{4.16}$$

Although this distance definition effectively reduces every set of sequences to a single average, we will see that this approach works very well in practice. This definition allowed me to align all sequences in the Glasgow corpus and the result is very convincing (see Figure 4.9). Given $n$ sequences to align, I am using the generalised DTW algorithm as follows:

1. Assign each sequence to its own set $\mathcal{V}_i$.

2. Pick two sets $\mathcal{V}_i$, $\mathcal{V}_j$ and align their sequences using the generalised DTW algorithm. Assign the aligned sequences to their own new set $\mathcal{V}_{ij}$ and remove $\mathcal{V}_i$, $\mathcal{V}_j$.

3. If more than 2 sets are left, goto 2.

4. The single remaining set contains all aligned sequences.

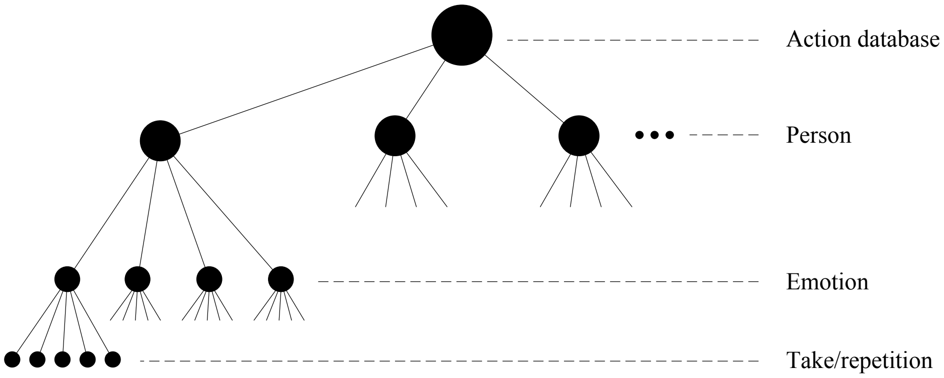

**Figure 4.8:** Part of the guide tree used for aligning a large number of sequences. Nodes in the tree represent aligned sets of sequences. Leaves are special sets containing only one sequence. The tree is heuristically constructed to pairwise align sequence sets of maximal similarity at every stage.

### Alignment order

The algorithm outline above does not mention the order in which the sequences should be aligned. It is usually regarded as most desirable to start with the most similar sequences. For example when aligning DNA sequences, evolutionary scientists commonly construct a *guide tree* first which describes the distance in relationship between the sequences [HS88]. If the tree is binary and rooted, it can be used to give a unique order for the alignments. Starting from the individual sequences at the leaves, larger sets are created by pairwise alignment of the child sets to produce their parent set. For aligning the Glasgow corpus of actions, I am also constructing a guide tree based on the known origin of the data. The tree defines a hierarchy of relationships depending on whether the motions originate from

1. the same take (recording)
2. the same emotion
3. the same person.

A part of the resulting tree is shown in Figure 4.8. Note that it is not binary but its unique root represents the set of all aligned motion sequences. All sequences of the database are represented in the leaves of the tree. Internal nodes represent sets of aligned sequences. As the tree is not binary, there is some ambiguity to the alignment order at each level. If a node has more than two children their combined alignment is produced by aligning sequences at random, while trying to keep the intermediate sets as balanced as possible. The root node represents the set $\mathcal{V}^*$ of all aligned sequences. The pairwise alignments we saw in Figure 4.7 were in fact taken from the set $\mathcal{V}^*$. They show that even after aligning over 1000 sequences, the result is meaningful for individual sequence pairs.

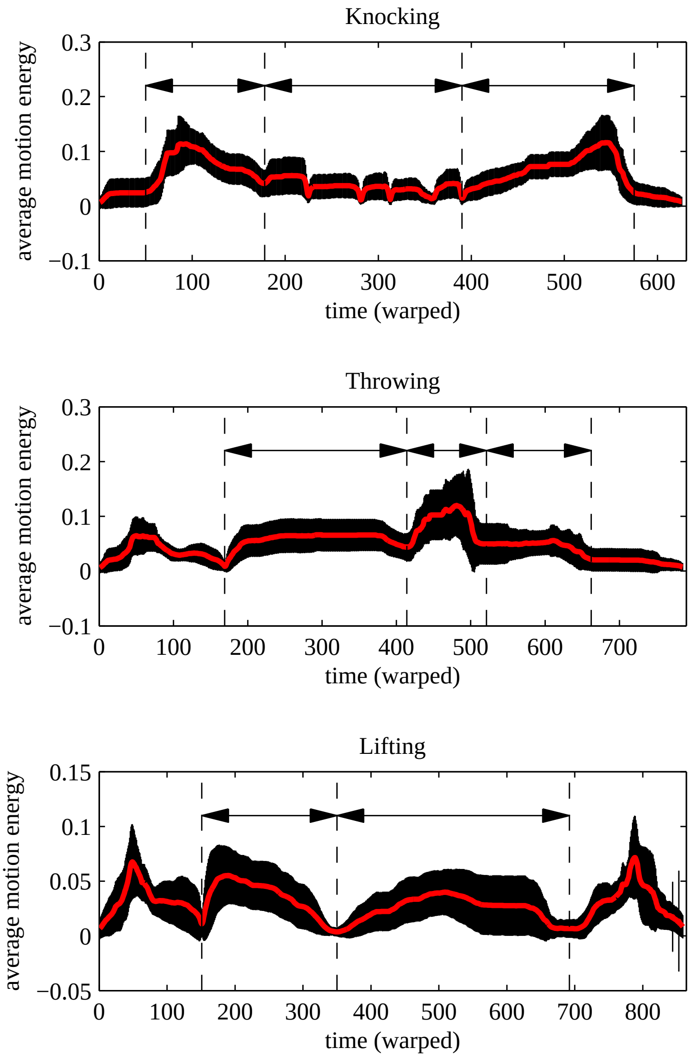

**Figure 4.9:** Plot of the master descriptions for *Knocking*, *Throwing* and *Lifting* motions. Each plot depicts the mean energy graph of approximately 1200 aligned motion sequences. The black regions indicate the variance across the aligned signals at each time frame. The defined primitive regions are indicated by arrows.

### Defining motion primitives

Once all the actions have been aligned, I need to visualise the result in an appropriate manner. The primary goal is to make it easy for a human to choose which portions of the motion should be picked as motion primitives. While working with the energy-based definitions in Section 4.3.1, we saw that energy graphs are a very useful way of visualising dynamic data. The notion of a motion energy graph is very similar to the common use of waveforms reflecting sound pressure levels and sound intensity in audio editors (see Figure 4.10). I visualise the master description as the average energy graph based on all the aligned motions. In order to capture some of the large variation in the data, I also show the maximum and minimum values in the form of a variance region. Figure 4.9 shows the resulting master descriptions for *Knocking*, *Throwing* and *Lifting*.

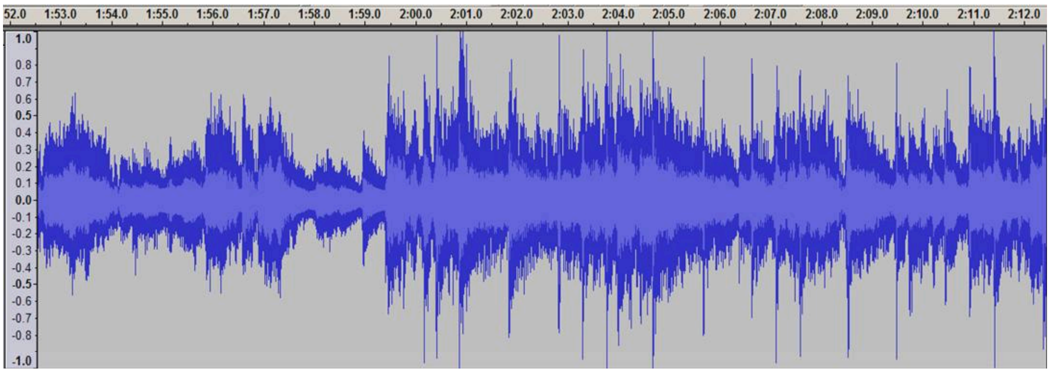

**Figure 4.10:** Sound intensity is a primary cue used in any standard music editor.

The characteristics of each action are clearly discernible. For example, the individual knocks have been preserved in the master description of *Knocking*. This allows for a straightforward definition of motion primitives as a range of frames in the master description graph. Note that each frame in the combined graph corresponds to unique frames in all of the original sequences. Because the original sequences may be locally stretched, however, it is not in general the case that every frame in an original sequence corresponds to a unique frame in the combined graph. Instead, every frame in an original sequence corresponds to a connected *range* of frames in the master description. For the purpose of this dissertation, I am using the following motion primitive boundaries as depicted in Figure 4.9:

- **Knocking**
  - 50 - 178: raising the arm
  - 178 - 390 : repeated knocking
  - 390 - 575 : lowering the arm
- **Throwing**
  - 169 - 414 : flexing the arm before throwing
  - 414 - 521 : throwing motion
  - 521 - 662 : recoil motion
- **Lifting**
  - 151 - 350 : lifting the object
  - 350 - 692 : putting it back

In order to extract these motion primitives from new motion examples, we can use the following algorithm (see also Figure 4.11 for an illustration):

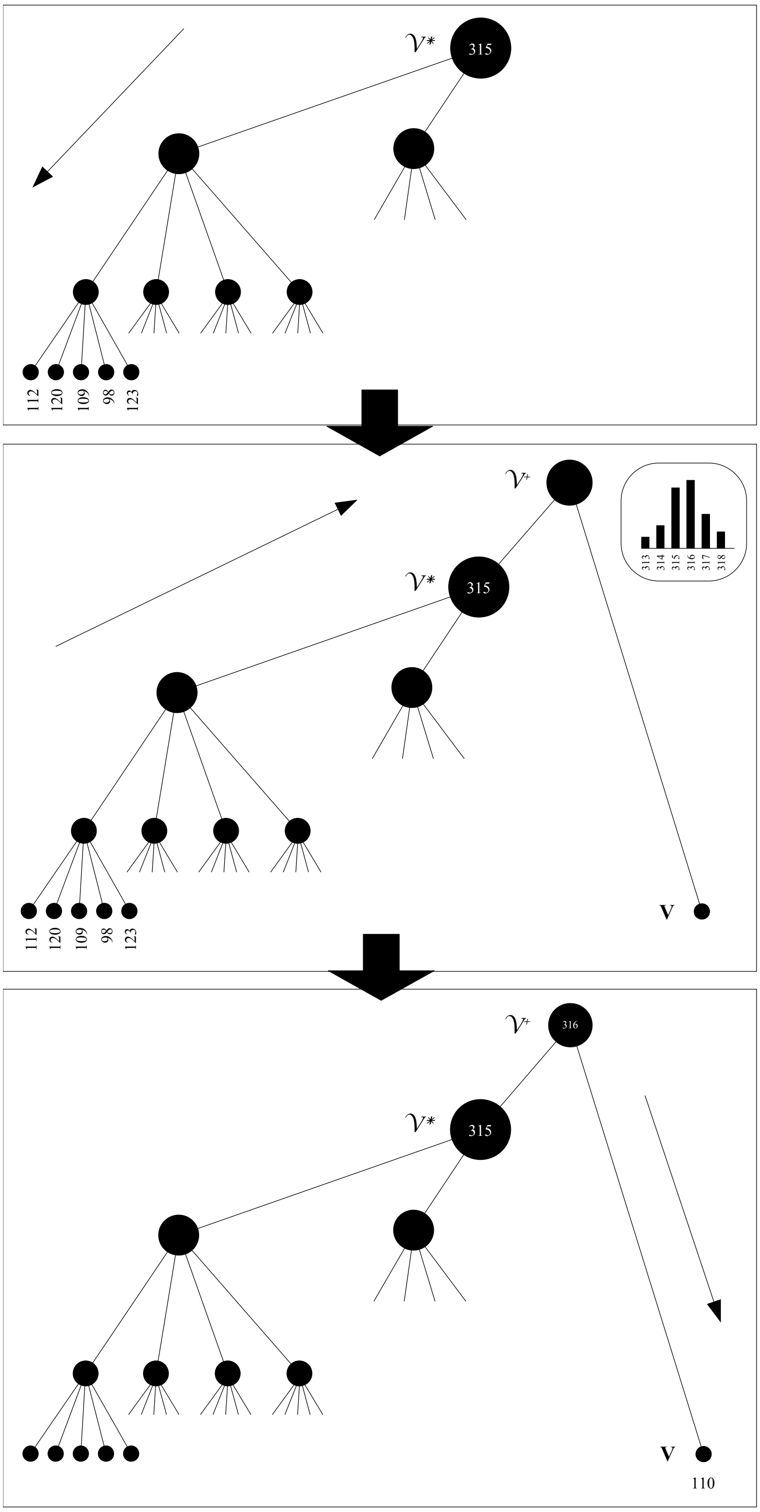

**Figure 4.11:** Propagating the primitive boundaries to a new sequence $\boldsymbol{V}$. The new sequence is first aligned with the full sequence set $\mathcal{V}^*$ to form $\mathcal{V}^+$.

1. For each primitive boundary, uniquely identify the frames in the original sequences which correspond to it (Figure 4.11 top).
2. Align the new motion $\boldsymbol{V}$ to the fully aligned set $\mathcal{V}^*$. This gives rise to a new set $\mathcal{V}^+$.
3. Set $n = 1$.
4. For the $n$th primitive boundary, build a histogram of frames in $\mathcal{V}^+$ which correspond to the primitive boundary (Figure 4.11 centre).
5. Pick the median frame as the primitive boundary in $\mathcal{V}^+$. Uniquely identify the frame in $\boldsymbol{V}$ which corresponds to the primitive boundary (Figure 4.11 bottom).
6. If there are more boundaries, increase $n$ and goto 3.

The expensive step of this algorithm is the alignment in step 2. In its simplest implementation the complexity of DTW is $O(T^2)$ where $T$ denotes the length of the sequence. As this implementation performs multiple sequence alignment, a distance formulation as given in Equation 4.16 is linear in the number of sequences $V$ and hence leads to a complexity of $O(VT^2)$. If performance is important, two optimisations can be used. The first optimisation concerns the number of sequences stored with the master description. It is not necessary to store all $V$ sequences which were used to compute the alignment. If only a constant number of sequences are retained after the master description has been computed, each evaluation of Equation 4.16 becomes constant time. A further optimisation comes from approximating the calculation of the full DTW matrix in linear time (*e.g.* see FastDTW by Stan *et al.* [SC07]). This leaves us with a time complexity of $O(T)$ for step 2. The remaining steps can be carried out in constant time by precomputing and storing the primitive boundaries at the root node of the master description tree. These optimisations were not implemented for this work.

### Evaluation

Examples for aligning new motions to previously defined master descriptions are shown in Figure 4.12. Note that due to the multiple alignments the master descriptions are generally significantly longer than individual motions. Aligning a new motion to the master description computes a warping function which maps between the two sequences. For these examples I am using the master descriptions as shown in Figure 4.9 and align three random sequences and use the above algorithm to propagate the previously defined primitive boundaries to the new sequences. The energy plots indicate that the alignments are meaningful and the boundaries have been mapped correctly to the new motion.

Because all Glasgow motion sequences are available at training time, I can take the primitives directly from the master descriptions. In detail, I computed a master description for

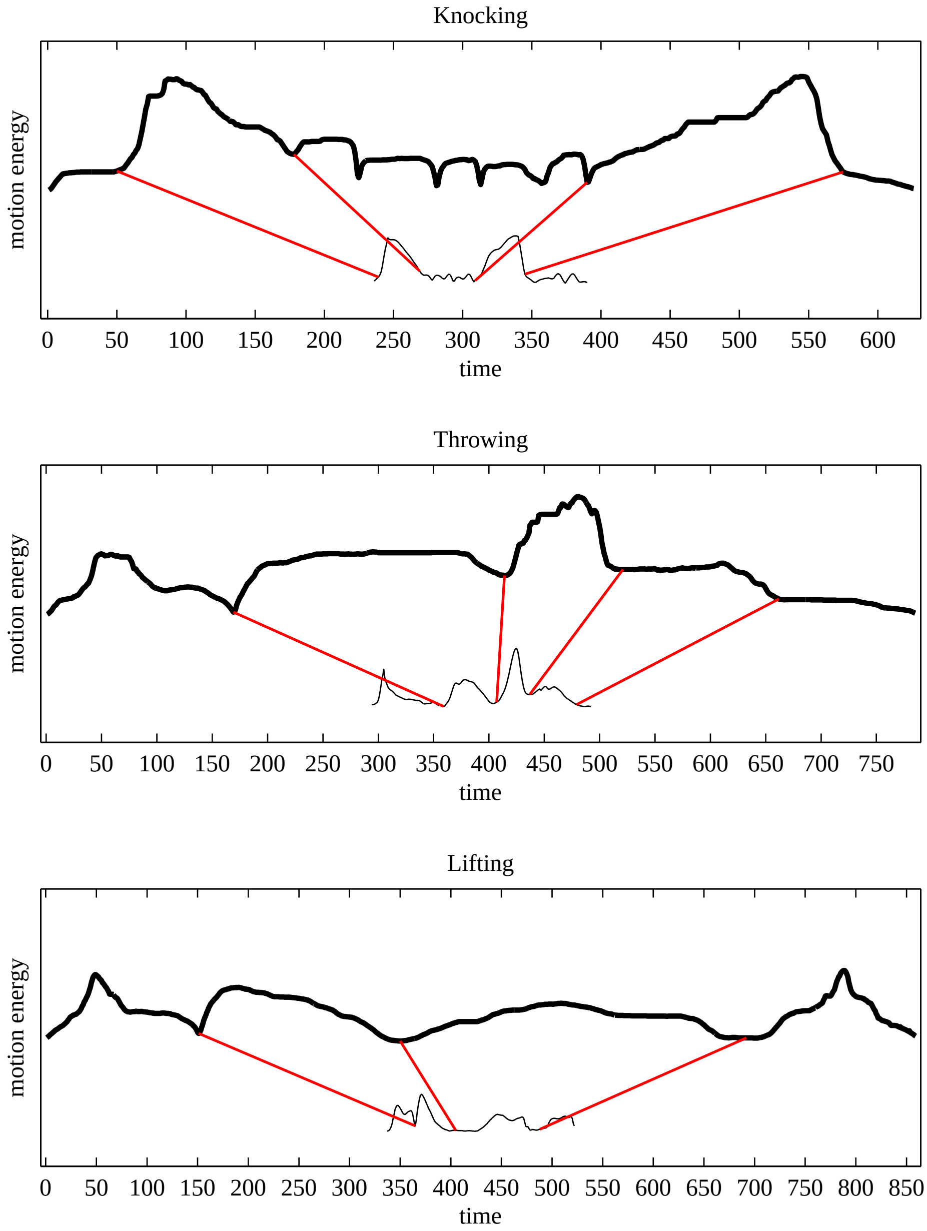

**Figure 4.12:** Example alignments for new *Knocking*, *Throwing* and *Lifting* motions (bottom) to the previously defined master descriptions (top). The red lines indicate where primitive boundaries have been aligned.

each action category by aligning all samples in the Glasgow corpus. Defining the primitives as shown in Figure 4.9 then yields primitives for each of these samples. Those are the primitives I use in future chapters. Additional evaluation statistics for these primitives are available in Appendix D.

### 4.3.3 Cyclic motions

Many real-world tasks can be thought of as cyclic in nature. Walking, reading, eating, hovering, brushing teeth all have more or less repetitive elements. Clearly, sequence

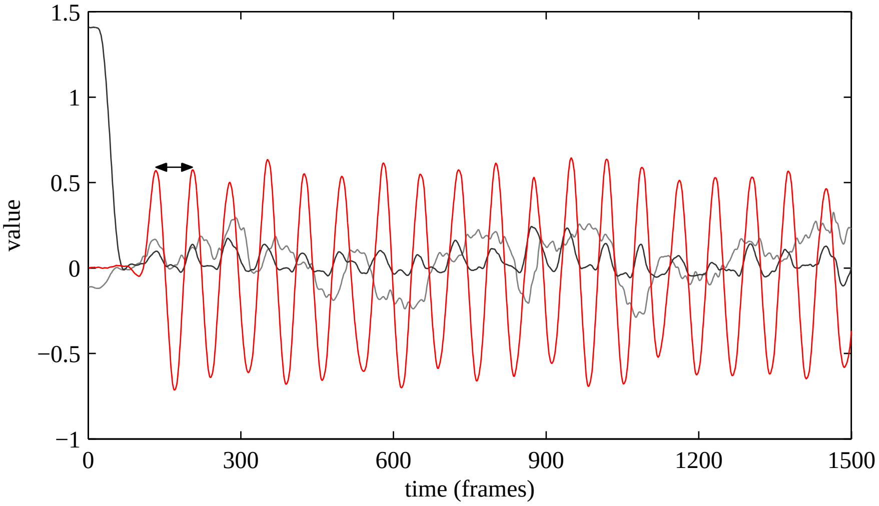

**Figure 4.13:** PCA-transformed *Walking* sequence. The walking cycles are clearly discernible once the signal is projected onto the first principal component (red). A single cycle has been indicated with an arrow.

alignment is not meaningful for cyclic motions if the motions to be aligned exhibit different numbers of cycles. Instead, I first segment cyclic motions into individual cycles. If necessary, I can then extract motion primitives from each cycle individually. I have processed the *Walking* motions from the Glasgow corpus in this way.

Various methods exist to segment cyclic motions into cycles. I could, for example, use a state-based method like cyclic left-to-right HMMs to model the actions (see Section 4.2.2). As walking motions are not very complex in structure, I will use PCA — a simpler method which has the added benefit of also allowing easy segmentation in real-time.

For this approach I am representing the motion as a time series of the lower body joint position and speed information. I introduced this representation as observation variables for the action HMMs at the beginning of this chapter (see Section 4.2.2 and Table 4.1). PCA is a powerful statistical method to reduce the dimensionality of this high-dimensional signal one frame at a time. At the same time the goal of the PCA transform is to preserve as much variation in the data in as few dimensions as possible (see Section 3.5). The main cycles of a cyclic motion such as *Walking* usually account for the largest amount of signal variation. Computing the PCA transform of such a motion is hence likely to expose those main cycles in the first dimension. In order to test this hypothesis I took a random sample of 100 *Walking* motions and computed their PCA transform. I then applied this transform to a new *Walking* sample. Figure 4.13 depicts the first three dimensions of this transformed sample. The first dimension (highlighted in red) follows a clear sinusoid. In this case the period is around 60 frames (1 second). This is the cycle period which the subject took for two consecutive steps. I can robustly extract the cycle boundaries as the local maxima of the sinusoid.

After extracting the cycles I can treat each individual cycle like any non-cyclic action and analyse it in terms of motion primitives. Several authors, for example, have analysed gait in terms of half-cycles or half-steps [VSM00, KS05]. This analysis mainly benefits the modelling and synthesis of human gait. Because walking is such a simple and symmetric motion and in order to simplify the analysis, I treat every cycle as a motion primitive without any further subdivisions.

### 4.3.4 Related work

This section has dealt with the definition and extraction of motion primitives in the context of extracting emotional features from body movements. In her work on emotion recognition from expressive body movements Castellano used dynamic features which describe submotion characteristics like initial, final and main motion peaks [Cas08]. She argues that emotional expression is reflected to a great extent in the timing of the motion rather than in absolute or average measures. In the case of everyday actions, these kinds of dynamic features are generalised by the notion of motion primitives which can capture the dynamics in terms of any number of subelements of an action.

Motion primitives are a very common method of dealing with the complexity of human movements. Indeed, facial movements are very often modeled as series and superpositions of basic actions as in the Facial Action Coding System developed by Ekman and Friesen [EF78] which is very popular with psychologists and has become the de-facto standard for analysing facial expressions in affective computing. Another and slightly lower level facial action parameterisation is provided by the Facial Animation Parameters (FAPs) defined in the MPEG-4 standard to allow the animation of synthetic face models [Ost02].

Unsurprisingly, psychology and animation have been two very active areas for structuring human movements into primitive elements. The psychologist Peter Bull developed a system called the Body Movement Scoring System to transcribe his observations without including social meaning [Bul87]. For a similar purpose the anthropologist Ray Birdwhistell developed a transcription system which is only based on physical appearance and anatomy rather than culture-specific interpretations [Bir70]. My work takes inspiration from Birdwhistell's observation that complex motions can be broken down into an ordered system of isolable elements which he called kinemes. He stressed the closeness to the linguistic concept of phonemes. This conceptual closeness means that many computational approaches first developed for speech processing later found use in the analysis of sign language (*e.g.* see [SNH98]) and I have demonstrated how they can be used for the more general task of action understanding. In other computational systems motion primitives have been used for tracking and recognising human skills, activities and gestures [GG04, Vec02]. In these contexts motion primitives have been labeled in many different ways. While Green called them dynemes, Del Vecchio *et al.* name them movemes. Most of the time, these kinds of primitives are short motion fragments which are combined sequentially in the way that I have in this section.

Finally, motion primitives have been used for *producing* motions in computer animation and robotics. Indeed, my initial cluster-based primitive definitions are based on robotics work by Fod *et al.* [FMJ02]. They propose the use of perceptuo-motor primitives which bias the perceptual system in terms of the set of motor behaviours which a robot can execute. They also point out that there is a neuro-scientific basis for motion primitives and that in fact vertebrate motor control seems to be achieved through the flexible combination of primitive motions [HG04, TS00]. This notion has also been utilised by animators, especially in the field of data-driven motion synthesis [LE06, KS05].

The animation community has also been interested in aligning motion sequences. This idea has been fundamental to my second method of defining motion primitives. Using an approach similar to mine, Heloir *et al.* use temporal alignment to represent gestures in a style-independent way [HCGM06]. While their motion data representation is very similar to mine, their approach is aiming for smooth alignments as needed for animations. Also, few solutions have been developed for aligning more than two sequences at a time. Turning to other fields, multiple sequence alignment without smoothness constraints can be found in the field of computational genetics. In particular, the alignment of multiple DNA sequences through the popular CLUSTAL family of algorithms [HS88, THG94] has largely motivated my approach. Because DNA sequences are one-dimensional, I had to generalise the commonly-used algorithms to multi-dimensional motion signals as outlined in Section 4.3.2.

# Chapter 5

# Emotion recognition from isolated motions

In this chapter I am going to introduce a framework for detecting emotions from isolated everyday body motions. It is important to realise that many factors, including emotional state, have an effect on the dynamics of body movements. I am describing a machine learning framework which can account for some of them with the goal of separating different emotion classes as far as possible to aid classification. A number of experiments at the end of this chapter will evaluate, among other things, how the modelling of these different factors can influence the recognition of emotions.

## 5.1 Framework

At the heart of my emotion recognition approach lies the observation that an observed motion signal is the result of a number of factors which interact in complex ways. Factors which have been studied in past research have been gender, height, age, weight and

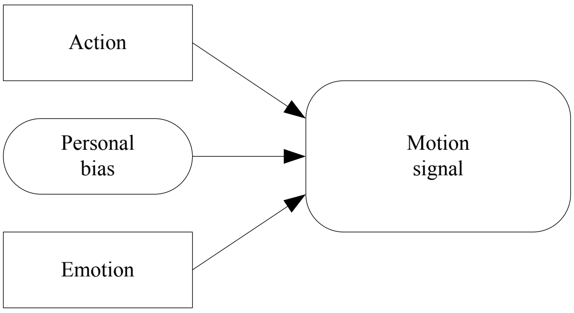

**Figure 5.1:** Three factors are affecting the final observable motion signal. Boxed signals are categorical, rounded signals are continuous.

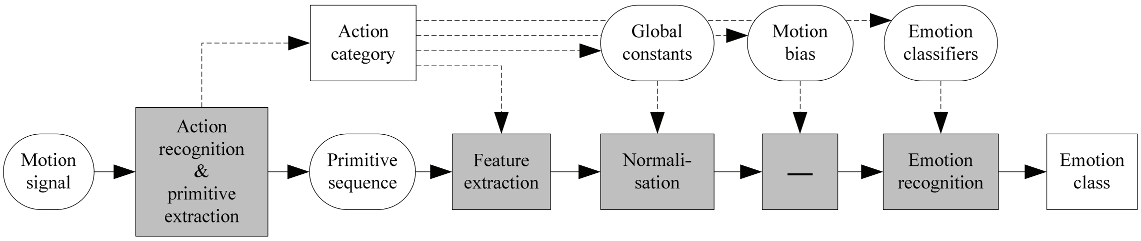

**Figure 5.2:** Data processing pipeline for emotion recognition from isolated motions. Grey components represent algorithms described in this previous chapters. They produce the data shown as white components.

emotion. At a more fundamental level, the motion signal will also be influenced by the action performed. For the purpose of this dissertation, I distinguish between three factors (see Figure 5.1 reproduced from Chapter 1):

1. action
2. personal motion idiosyncrasies
3. emotion.

I will regard these factors as largely independent of each other. This assumption is strongly supported by Wallbott's findings [Wal98]. The goal is to develop a good understanding of these factors and how they influence the motion signal. This will allow me to build effective models including a model for the *emotional* influences on the motion signal. In the emotion recognition framework, I start from an unclassified motion signal. From that, I will separate the influences of action and idiosyncrasies before attempting to detect the emotion. Figure 5.2 summarises the recognition process. I first classify the signal into one of $n$ action classes. At this stage it is also possible to break down the signal into primitives in order to model the emotional influences at a finer scale. I discussed this topic extensively in Section 4.3. Next, I extract emotion-communicating features from the signal. I introduced this concept for the Cambridge corpus in Section 3.4 and will return to it in Section 5.2. The next step is normalisation and is aimed at removing any biases introduced by the feature representation. I will talk more about this step when discussing SVM-based classification in Section 5.5.1. After this stage, the features are still confounded by person-dependent factors. Therefore, before feeding the features to an emotion classifier, I remove the bias introduced into the signal by the personal idiosyncrasies. I will discuss this process in detail in Section 5.3.

There have been several attempts, both by the animation and recognition communities, to build models of body motion which are able to incorporate a number of different factors. One of the most popular abstractions distinguishes between *content* (action) and *style* (*e.g.* gender, emotion *etc.*) [TF00]. Notably, Brand and Hertzmann presented an HMM-based model they called Style Machines [BH00]. Being interested in animation and data generation, their models are generative and designed to produce convincing animations rather than distinguishing labelled styles from a corpus. Bobick and Wilson describe Parametric HMMs which are able to incorporate simple distance parameters in gestures indicating quantities [WB99]. Even though they suggest that the method may be applicable to human motion data such as gait, the most popular approaches to identity recognition from human gait have tended to use explicit feature extraction rather than HMM parameters to model the style differences [Gri02, JB01, NCN$^+$99, UF04]. Furthermore, many approaches to discriminate emotion in the body, face and voice have been discriminative rather than generative like HMMs [KKVB$^+$05, CVC07, DSBB04, MK03, PR00, BLFM03, FT05]. My recognition framework is also based on a discriminative emotion model rather than a generative one. In Section 5.5 I will make use of Support Vector Machines to classify feature vectors into emotion classes.

## 5.2 Feature definitions

In both generative and discriminative approaches the choice of features is paramount to achieving good recognition results. In Section 3.4 I described a set of dynamics-based features which I showed to be appropriate for capturing emotional differences in the Cambridge corpus of expressive body motions. Recent psychological studies have shown that dynamics-based movement descriptions are also effective for explaining the perceived difference in everyday movement scenarios [ATD07, PPBS01]. In particular, the activation dimension in a continuous emotion space has been shown to be strongly correlated with movement speed in certain actions and activities [PPS01, Pat02]. These findings supplement other studies based on more archetypal motions which also stress the importance of dynamic cues [ADGY04, CVC07, CMV04]. I will therefore largely make use of the features described in Section 3.4 to discriminate between emotions in everyday actions. The features I presented before were calculated based on a sliding window and amounted to statistics of

- joint positions (x, y, z)
- joint speed
- joint acceleration
- joint jerk.

**Table 5.1:** Set of simple features defined per joint in the Glasgow corpus.

| Channel | Statistical features computed | # Features |
|---|---|---|
| posture (x-position) | median, min, max, std deviation | 4 |
| posture (y-position) | median, min, max, std deviation | 4 |
| posture (z-position) | median, min, max, std deviation | 4 |
| speed | median, min, max, std deviation | 4 |
| acceleration | median, min, max, std deviation | 4 |
| jerk | median, min, max, std deviation | 4 |
| | | **24** |

**Table 5.2:** Set of simple features defined for the head in the Glasgow corpus.

| Channel | Statistical features computed | # Features |
|---|---|---|
| posture (pitch) | median, min, max, std deviation | 4 |
| posture (roll) | median, min, max, std deviation | 4 |
| | | **8** |

### 5.2.1 Simple features

According to my models in Section 4.2.2, each action is defined in terms of a set of joints $\mathcal{J}_i$. Therefore, for each action class $c$ I derive emotion features based on the set of joints $\mathcal{J}_c$. In detail, the feature vector derived for each joint individually consists of 24 dimensions as shown in Table 5.1. As before, the head is treated somewhat separately and gives rise to an 8-dimensional feature vector (Table 5.2).

For *Knocking* motions, for example, this means that the above statistical features are derived from the right arm joints (elbow and hand) and the head giving rise to 56 dimensions. Because these features are derived from the motion of single joints I refer to them as *simple* features.

### 5.2.2 Derived features

In order to demonstrate to what extent feature extraction can benefit from detailed motion modelling I also define a number of additional, more complex features. Three additional features make use of my increased ability to model activities. Inspired by energy-based features common in speech processing [Fur86, Rab02] I am combining the motion of multiple joints to calculate the overall motion energy of a whole body region (see Equation 4.8). As new derived features I then calculate

1. mean and median motion energy over a whole activity

2. the optimal segmentation threshold $\tau_{\text{opt}}$ as defined in Section 4.3.1.

Furthermore, I define two features based on motion primitives. Those are

1. length of the primitive in frames
2. phase difference between joints.

I am including the latter quantity with a reference to work by Bruderlin *et al.* [ABC96] which I discussed in Section 2.2.2. They suggested that phase differences may be indicative of emotions, which was later confirmed by Pollick *et al.* and Paterson [PPBS01, Pat02]. The phase difference is calculated based on the occurrence of the maximum speed of one joint compared to another one and can therefore be positive or negative. In practice, I am calculating phase differences between the motion of elbow and wrist joints, as well as between knee and foot joints.

### 5.2.3 Action-global *vs.* primitive-local features

In Chapter 3 the motion segments $m$ over which the features are calculated were defined through a sliding-window. I am now using the improved abilities to model complex actions when deriving emotion features. In particular, features can be either derived from a whole action or from individual primitives. In the remainder of this thesis I will refer to these two different kinds of features as *action-global* and *primitive-local* respectively.

In either case, the goal is to account for and discard as much variation as possible in the motion dynamics which comes from the sheer nature of the action. Clearly, the dynamics of the arm motion during *Lifting*, *Knocking* and *Throwing* actions are different by the inherent nature of the actions. Training emotion models depending on which action is being performed enables me to account for this and hence improve over sliding window-based classification. However, I need not stop at the action level. The flexing and throwing segments of a *Throwing* action could impose different biases to the observed dynamics due to the different functions and constraints by which they are governed. Assuming that we can model a more complex action through two motion primitives $m_1$ and $m_2$, deriving features from $m_1$ and $m_2$ independently may therefore enable me to remove even more action-based influence from the dynamics-based features.

I will write an action-global feature vector as $\boldsymbol{\phi}_{\text{Knock},p}$ where $p$ denotes the person. Feature vectors derived from motion primitives $m_1$ and $m_2$ are written as $\boldsymbol{\phi}_{m_1,p}$ and $\boldsymbol{\phi}_{m_2,p}$. $\boldsymbol{\phi}_{\text{Knock},p}$ can also be understood as a feature vector where a single primitive is defined as a whole *Knocking* action.

## 5.3 Personal motion bias

In Section 3.5 we saw that there are individual differences in the expression of emotions in affective gestures. In particular, we saw that the parameters of the feature distributions for different emotions can vary between individuals. These differences are certainly not unexpected. While humans themselves are very well adapted to distinguishing emotions from others' expressions, it can take a long time before we get attuned to the very subtle differences between *individual* expressions of emotions. We find it easier to interpret the emotional signs communicated by a good friend than those of a person who we have only known for a brief period of time. Individual differences in the expression of emotions have recently been investigated by several psychologists. Multiple studies conclude that facial expressions of the same emotion can differ between people because of cultural influences [Mat93, TM08, SWMK88] and individual style [SC01, WMH08]. Similar differences have been reported for body expressions [Cou04, Mei89, Ste92, Wal98]. All these observations lead to one conclusion: While emotional expression can be regarded as universal to some extent [Ekm94], there are individual differences which are important for high-accuracy inferences from emotional displays. Those seem especially pronounced if we consider *dynamic* displays of emotion rather than static images. This section will explore how individual differences occur in the dynamics of everyday actions, how they can be modelled and to what extent they are consistent across different action categories.

### 5.3.1 Modelling individual differences

Despite my observations regarding the human individuality of emotional expressions, many researchers dealing with the subject of emotion recognition treat those differences as marginal. Indeed, some researchers have aimed to produce systems which can detect emotions based on universal and person-independent features. I believe that ignoring individuality of expression from a recognition framework could limit its applicability and real-world success. Indeed, Picard stresses the fact that building universal affect-sensitive systems should start from a focus on the individual [Pic97]. I am therefore modelling personal movement style explicitly. Figure 5.2 showed the subtraction of motion bias as a separate building block of the data processing pipeline.

Many machine learning approaches dealing with human-generated data use some kind of normalisation procedure to account for the variation originating from the fact that the data is generated by different individuals. Once more I turn to research in speech recognition to find evidence for the importance of this kind of normalisation. Besides the modelling of environmental factors, speaker normalisation is discussed by many authors as one of the major factors for improving recognition accuracy [SAE07, GGB06, ZWFW97]. Numerous techniques for speaker normalisation have been developed, many of them focussing on physical differences between speakers such as vocal tract length or base frequency. Indeed, speaker modelling is the basis for the *recognition* of individuals from samples of their speech (*e.g.* see [Cam97, RQD00]). More recently, the same concept has lead researchers to investigate the recognition of individuals from differences in their body movements [LG02, KSR$^+$04].

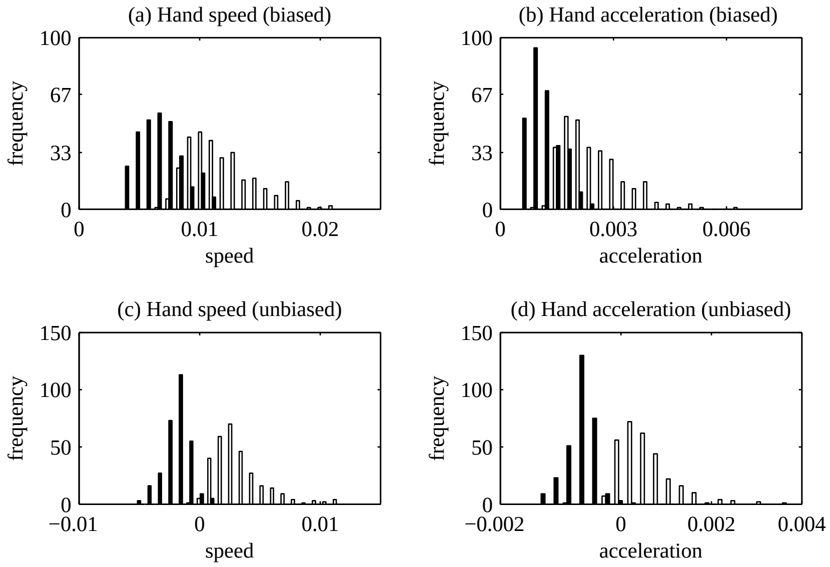

**Figure 5.3:** Biased and unbiased feature distributions for sad knocks (black) and angry knocks (white).

Recently, normalisation with respect to individual subjects has also been proposed in approaches to recognise emotions from various modalities, including speech [SAE07], physiological signals [NALF04] and the body [Cas08]. In particular, Castellano *et al.* mention that normalisation of features derived from body movements is important for achieving good classification performance across subjects [Cas08, CVC07]. Caridakis *et al.* describe a neural network-based approach to adapting their classifiers to user differences [CKK08]. I am building on these insights to model individual differences and hence improve the accuracy of emotion recognition.

### Individual movement bias in the Glasgow corpus

Within the scope of this dissertation, I am interested in how the set of feature vectors $\boldsymbol{\phi}$ are affected by individual differences. I obtain a quantitative view of the feature distributions by creating histograms of individual features and emotion classes. Figure 5.3 (top) shows the feature values for hand speed and acceleration respectively as derived from *sad* and *angry Knocking* motions. As in Section 3.5, the feature distributions are clearly discernible. However, from a pattern classification standpoint, they show a relatively low $\frac{\text{between-class-scatter}}{\text{within-class-scatter}}$ ratio. The distribution of sad samples overlaps heavily with that of angry samples. This exemplifies the problem of individual movement bias. While the *angry* knocks appear generally more energetic than *sad* ones, there is a large spread in the absolute values observed from different individuals. This leads to the tails of the observations overlapping heavily, thus making a global classification on the basis of unnormalised features sub-optimal.

In order to improve the scatter ratio, I am making the assumption that a person's motion idiosyncrasies influence his/her movements in a consistent fashion — after all they are expected to be governed by gender, physical build and other constant or semi-constant factors. I therefore model individual motion bias as a vector $\bar{\boldsymbol{\phi}}_{m,p}$ which I call motion signature. For a person $p$ and motion primitive $m$ I am thinking of the motion signature as a prior probability distribution parameterised by a mean $\mu(\bar{\boldsymbol{\phi}}_{m,p})$ and variance $\mathrm{var}(\bar{\boldsymbol{\phi}}_{m,p})$. The distribution captures the prior belief about the features observed for person $p$ displaying motion primitive $m$. As described before, the primitive index $m$ captures both the action category and particular primitive within that action category. Since I am assuming motion bias and emotional influences to be independent, I can calculate an unbiased feature vector $\hat{\boldsymbol{\phi}}_{m,p}$ by normalising with respect to the bias parameters:

$$\hat{\boldsymbol{\phi}}_{m,p} = \frac{\boldsymbol{\phi}_{m,p} - \mu(\bar{\boldsymbol{\phi}}_{m,p})}{\mathrm{var}(\bar{\boldsymbol{\phi}}_{m,p})} \tag{5.1}$$

In the remainder of this dissertation I will refer to $\hat{\boldsymbol{\phi}}_{m,p}$ as **unbiased feature vector** while $\boldsymbol{\phi}_{m,p}$ denotes the original, **biased feature vector**. Note that the bias is a unique signature computed per person and motion primitive. An important problem is how to estimate $\bar{\boldsymbol{\phi}}_{m,p}$. If a person is not "known", *i.e.* the system has no history of his/her movements, it may be necessary to take an a priori guess, maybe conditioned on gender or other cues. A brief exploration of the effects of gender and age on motion bias is given in Section 5.3.2. However, if we have observed a number of movements from the person in the past, it is possible to compute $\bar{\boldsymbol{\phi}}_{m,p}$ from all the observed motions. I therefore compute the motion bias $\bar{\boldsymbol{\phi}}_{m,p}$ as the mean and standard deviation over all the motions primitives of type $m$ for person $p$ in the database. This operation entirely ignores any emotion-specific information in the database as different emotion classes are represented at equal frequencies. Fig. 5.3 (bottom) illustrates how this normalisation improves the between-class variability for the two previously considered emotion classes considerably. In Section 5.5.2 I will give a quantitative account of the improvements in classification accuracy achieved when taking movement bias into consideration. First, however, I am going to explore in more detail how personal motion signatures $\bar{\boldsymbol{\phi}}_{m_1,p}$ and $\bar{\boldsymbol{\phi}}_{m_2,p}$ are related for motion primitives extracted from different action categories.

### 5.3.2 Predicting motion bias — user adaptation

Adding the personal motion signature into the data processing pipeline clearly improves the accuracy of the models (see Figure 5.3). However, from a practical standpoint, this kind of person-dependent model clearly comes at a cost. $\bar{\boldsymbol{\phi}}_{m,p}$ needs to be estimated from data supplied by person $p$. In particular, because $\bar{\boldsymbol{\phi}}_{m,p}$ is formulated in terms of a particular motion primitive $m$, it is necessary to observe a range of motions from each relevant action category. For certain scenarios this might not be a problem. It could, however, make adding new motion categories difficult and would mean that a new person would need to *train* the system for a period of time before he or she could expect accurate recognition results.

One intriguing possibility is to try and find a relationship between $\hat{\boldsymbol{\phi}}_{m_1,p}$ and $\hat{\boldsymbol{\phi}}_{m_2,p}$ which is consistent across many subjects $p$ and where $m_1$ and $m_2$ may come from different action categories. In effect, I would like to define a functional mapping $\hat{\boldsymbol{\phi}}_{m_2,p} = B(\hat{\boldsymbol{\phi}}_{m_1,p})$. I call $B$ an adaptation function. The definition of such an adaptation function would allow the generalisation of personal motion bias from one motion to another. If $m_1$ comes from a *Walking* sequence and $m_2$ is a *Knocking* primitive, it would be possible to predict what a certain person's knocking looks like on average from having observed him or her walking. This approach has the distinct advantage that some types of motions like *Walking* are much more common than others, so gathering enough data for estimating $\hat{\boldsymbol{\phi}}_{\text{Walk},p}$ reliably might be relatively easy.

### Adaptation

Indeed, it may not be necessary to derive $\hat{\boldsymbol{\phi}}_{m_1,p}$ from naturally observed data. As it potentially enables the derivation of motion signatures for many other action categories, it may be practical to record a few baseline calibration motions explicitly from which the motion signatures for all the other action categories of interest can be estimated. This approach is very popular in speech recognition systems which often include speaker adaptation steps to improve the recognition results. A common adaptation mechanism includes a short number of sentences which the user needs to read. This allows the system to build a speaker model which is generalisable to words and sentences beyond the ones read during the initialisation.

Given the relatively limited vocabulary of actions in the Glasgow corpus, I conducted a proof-of-concept experiment which is based on the notion of user adaptation based on explicitly supplied calibration data. As in the speech example, I am going to work from a dedicated calibration motion which a user would be instructed to repeat several times during an initialisation phase. In the Glasgow corpus, the best fit for such a motion is the initial *TDown* motion which the subjects are instructed to display before every motion recording. As it is very simple, it is the kind of motion which would be practical to use as a calibration motion. Yet, it contains enough dynamic elements to allow the estimation of all the dynamic parameters which make up $\hat{\boldsymbol{\phi}}_{m,p}$. Indeed, research by Castellano *et al.* has shown that a great range of dynamic subtleties can be expressed in simple movements very similar to the *TDown* motion [CVC07].

### Feature correlations

One way in which the speech community has achieved speaker adaptation is through the exploitation of correlations between speech units [CD97]. In the following experiment I am investigating the correlations between feature biases as derived from *TDown* and the upper body actions *Knocking*, *Throwing* and *Lifting*. I picked the following feature set for this experiment to include a variety of measures for multiple joints, as well as static and dynamic information:

- maximum vertical hand and elbow distance (2 dimensions)
- median hand and elbow speed (2 dimensions)
- median hand and elbow acceleration (2 dimensions)
- median hand and elbow jerk (2 dimensions)
- median head posture (2 dimensions).

First, I calculated the set of features $\boldsymbol{\phi}_{\text{TDown},p}$ for all samples in the Glasgow corpus. I then computed $\hat{\boldsymbol{\phi}}_{\text{TDown},p}$ as the corresponding medians and standard deviations for each of the 30 subjects. I am thinking of each of the elements of the 10-dimensional median and standard deviation vectors as a variable which takes on a different value for different subjects. I call these variables *adaptation variables.* Similarly, I calculated $\hat{\boldsymbol{\phi}}_{\text{Knock},p}$, $\hat{\boldsymbol{\phi}}_{\text{Throw},p}$ and $\hat{\boldsymbol{\phi}}_{\text{Lift},p}$. I am also thinking of the feature vector elements as variables which I call *bias variables.* As I am trying to infer the bias variables from the adaptation variables, we can think of the adaptation variables as *independent* while the bias variables are the set of *dependent variables.* In order to visualise the relationship between independent and dependent variables, I plot some of the bias variables computed from *Knocking* actions against the according adaptation variables in the form of a scatter plot in Figure 5.4. These plots strongly suggest a linear relationship between the adaptation and signature features.

In order to quantify this relationship, I carried out a correlation analysis. I calculate the Pearson correlation coefficient $\rho_{\phi_a,\phi_b}$ for all feature pairs $(\phi_a, \phi_b)$ from the adaptation and bias data respectively. With gender and age I am adding an extra two independent variables which could be used as a basis for adaptation. I am also computing correlations based on the added age and gender variables. Figure 5.5 (top) shows a visual representation of $|\rho_{\phi_a,\phi_b}|$ for the three different action categories.

My final goal is to establish which of the correlations are statistically significant. I form the null-hypothesis that a feature pair is uncorrelated and use Student's t-distribution for a transformation of the previously derived correlation values. This gives me a *p*-value, $p_{\phi_a,\phi_b}$, indicating how likely the null-hypothesis is for adaptation feature $\phi_a$ and bias feature $\phi_b$, given the data samples. Figure 5.5 (bottom) shows $p_{\phi_a,\phi_b}$ thresholded at different levels $\tau_p$.

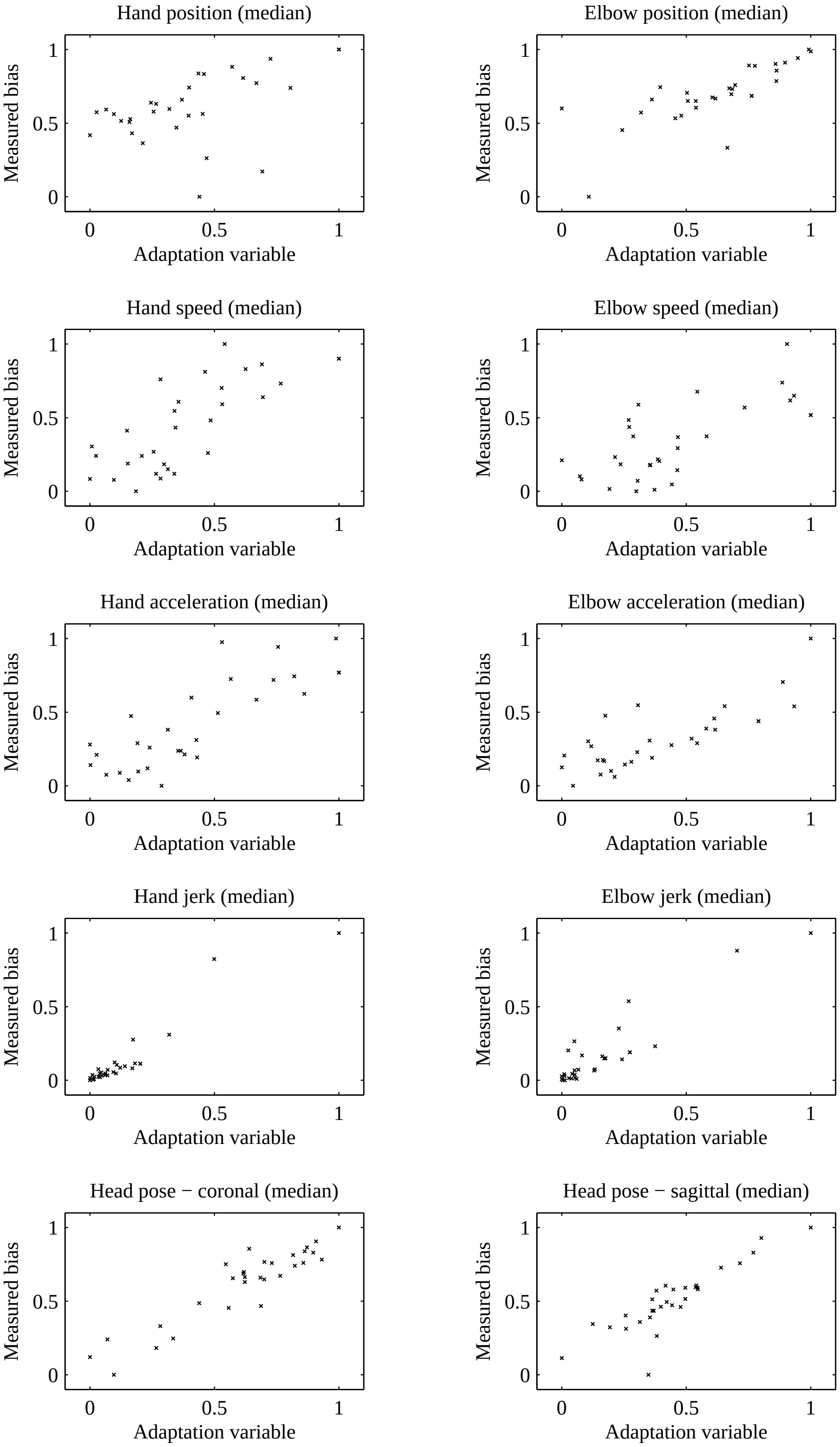

**Figure 5.4:** Plots of features derived from the *TDown* adaptation motions versus the motion bias calculated from all *Knocking* motions in the Glasgow corpus. Each data point is derived from one of the 30 subjects in the corpus. Values were normalised to the interval $[0, 1]$.

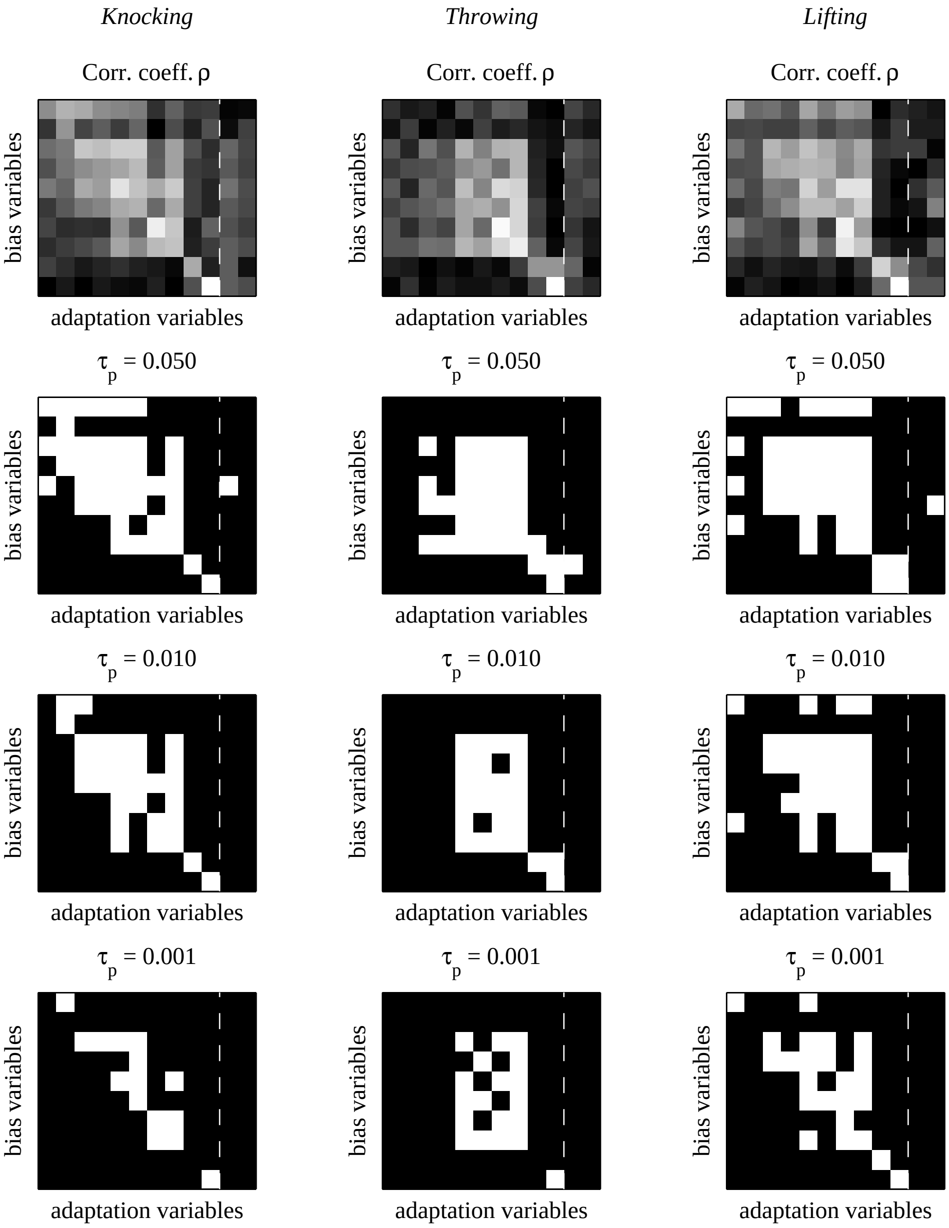

**Figure 5.5:** Pairwise correlations between adaptation variables and motion bias variables. Elements $(i,i)$ in the visualised matrix correspond to the *same* features (e.g. hand speed). White elements denote a significant correlation between the corresponding variables at the according significance threshold $\tau_p$. The two additional adaptation variables separated by a dashed line are age and gender.

### Discussion

We can see very clear correlations between the features of the calibration motion *TDown* and other actions. Not surprisingly, the correlations are particularly pronounced along the diagonal, suggesting that the same variables (such as median hand speed) in *TDown* and other actions are correlated. However, we can also see other features off the diagonal being highly correlated. Interestingly, age and gender do not seem to have as strong

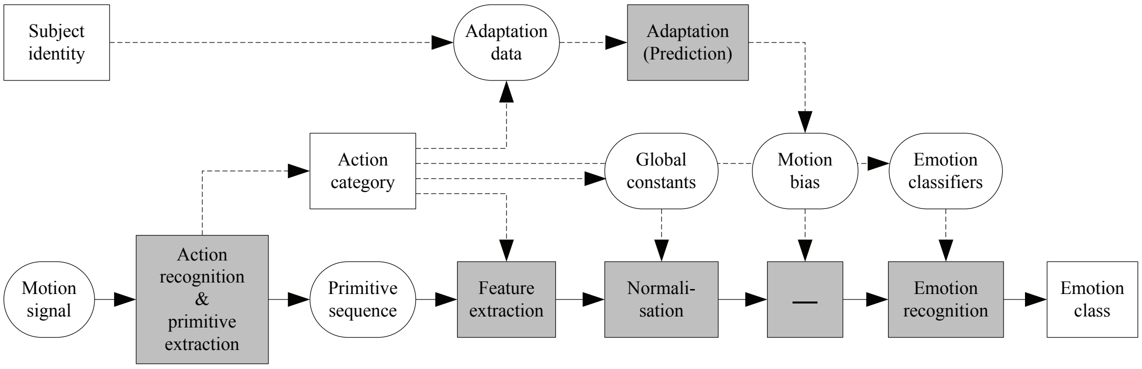

**Figure 5.6:** Data processing pipeline augmented to incorporate adaptation based on explicitly collected adaptation data. The adaptation data to use is chosen according to the action category and primitive.

an effect. This suggests that age and gender are not very good predictors as far as the dynamics of human upper body motions are concerned. On the other hand, there are many features which are significantly correlated at the $p = 0.001$ level which suggests that adaptation on the basis of a calibration motion might be a practical possibility.

Given the strong correlations, a natural candidate to define an adaptation function $B$ is linear regression. This method has been very popular in the speech recognition community [LW95]. Figure 5.6 shows how my data processing pipeline can be extended to allow for $B$ to predict the motion bias for a particular action category and primitive. In Section 5.5.2 I will make use of this technique to evaluate how effectively user models can be built based on such an adaptation function. In particular, I will compare how well emotion classification performs using $B$ compared to a non-personalised baseline model and a fully person-dependent model which makes use of the "true" motion bias. First, however, I will take a closer look at the role of the different features in emotion recognition.

## 5.4 Feature analysis

In the previous sections and chapters I have made use of rather ad-hoc choices for the subset of emotion features to use. It is, however, not obvious which features will be good predictors for different emotion classes. A closer analysis of the importance of features in this section serves two purposes. Firstly, it will give us important insights into which qualities of body motions encode emotional meaning. Secondly, it will allow me to reduce the feature set effectively by removing features which are irrelevant or noisy. In certain scenarios, this has been argued to improve the performance of machine learning. I will again focus on the set of upper body actions, but will give due consideration to a variety of different factors. Firstly, I will investigate the usefulness of different features for different action categories. Furthermore, I categorise every feature along a number of dimensions.

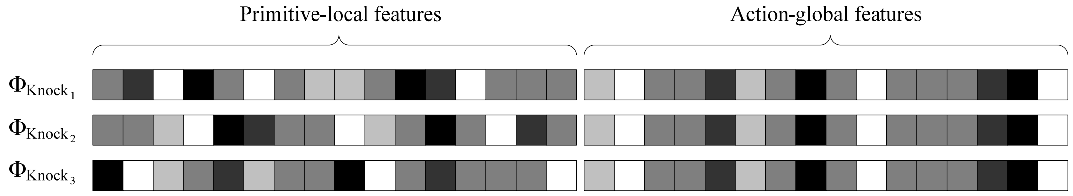

**Figure 5.7:** Schematic representation of feature vectors derived from the three primitives extracted from the same *Knocking* sample. Shaded elements represent feature values of different dimensions.

For each of the dimensions below I can later determine which values (given in brackets) provide particularly discriminative information.

- statistical measures used to derive features (median/standard deviation/min/max)
- originating joints of features (elbow/hand/head)
- primitive-based *vs.* action-global features (primitive-local/action-global)
- simple statistical *vs.* complex derived features (simple/derived).

I will base my analysis on a feature subset selection framework. In particular, I compute the full set of features and collect them in a large feature vector $\boldsymbol{\phi}_{m,p}$. There is one such vector per motion primitive $m$. In order to be able to evaluate the relative merit of primitive-based *vs.* global features, each vector contains both the features derived from the primitive *and* the features derived from the whole action globally (see Figure 5.7). I will then perform a feature selection routine which chooses dimensions of the feature vectors which are highly predictive of the associated emotion classes. I can then analyse the selected subsets with regards to the different factors outlined above. For example, we can see how important primitive-based features are compared to action-global features by what proportion of each was picked by the selection algorithm. In the following analysis I am using features which have been *unbiased* as described in Section 5.3.1.

### 5.4.1 Feature subset selection

Feature subset selection is a topic in its own right in the pattern recognition community. The goal here is not to investigate the relative merit of the many algorithms available, but to choose one which can be expected to supply useful information for the problem at hand. In particular, I want to establish which features from the large and redundant feature vectors $\boldsymbol{\phi}$ carry the most information about the emotional classes. Within a machine learning framework, the features are normally evaluated in terms of the performance they yield when used in conjunction with a particular machine learning algorithm — an approach known as feature *wrappers* [KJ97]. Although I will ultimately use the features in a machine leaning framework, the results from this section should give us insights going beyond a particular machine learning application. In order to ensure the generality of the results, I will therefore use a *filter* technique which evaluates features according to heuristics based on general characteristics of the data itself [JKP94].

Although filters are a relatively old technique, they are the optimal choice here as many of the more recently developed techniques such as boosting-based methods [Das01, TV04, RL05] benefit from the considerable progress in the field of pattern classification and are thus inherently bound to classifiers. Other popular techniques operate in the principal or independent component space [LEAP05, GCP07] and thus achieve feature reduction by combining multiple features. This makes it harder to draw conclusions about the significance of observable properties such as hand speed. Finally, being a well-studied technique, implementations for various filter methods are readily available for experimentation in knowledge analysis packages such as WEKA [HFH$^+$09].

Turning to the structure of the data at hand, one of the problems is that many of the features are likely to be correlated with each other. For my analysis, however, it would be beneficial to derive the *most parsimonious* set of features which shows a small amount of inter-correlation. This way we can be sure that each selected feature actually captures some significant and unique information about the expression of emotions. In other words, I am looking for a set of features which are highly correlated with the emotional class, but uncorrelated with each other.

One algorithm which solves this problem and whose implementation has been discussed in the literature is a correlation-based approach by Hall *et al.* [HS97]. In order to evaluate the correlation between features and the emotion classes the approach makes use of the information-theoretic concept of information gain. The merit of a certain subset $\Phi$ of $k$ features can be quantified as the quotient of the mean class-feature correlations $\rho_{c,\phi}$ and the mean feature-feature intercorrelation $\rho_{\phi,\phi}$:

$$Merit(\Phi) = \frac{k\rho_{c,\phi}}{\sqrt{k + k(k-1)\rho_{\phi,\phi}}} \tag{5.2}$$

I use the $Merit$ function as a heuristic to direct a Hillclimbing algorithm to find a good feature subset $\hat{\Phi}$. This is a common approach used to cut the search complexity in feature subset selection [HS97, RK90]. Starting with no features, new features are iteratively added and their $Merit$ scores stored in a priority queue. The queue enables backtracking in a best-first fashion in case the new sets do not amount to any improvements in $Merit$. At every iteration the feature set with the biggest $Merit$ is chosen from the head of the queue and expanded into new sets by adding new potential features. The algorithm stops after not having improved $Merit$ for five iterations, yielding the subset with the largest $Merit$ seen up to this point, $\hat{\Phi}$.

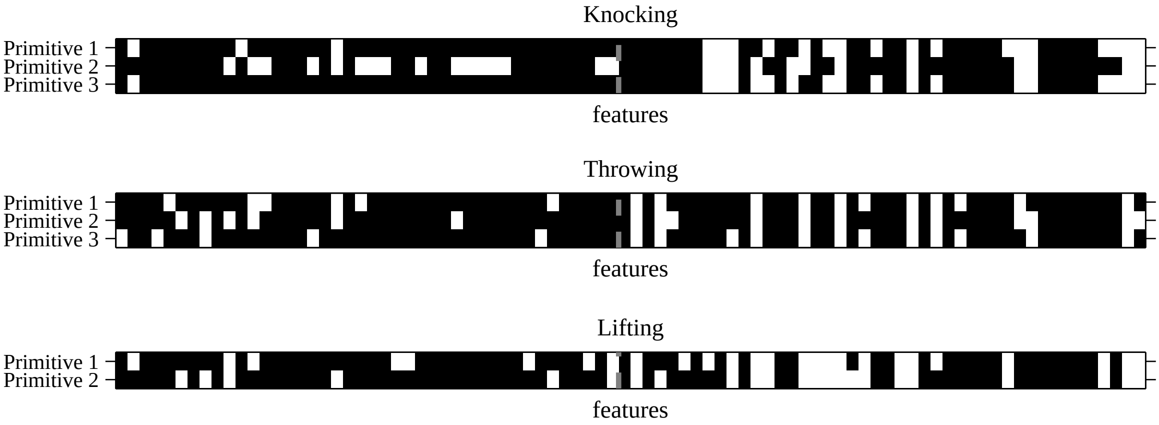

**Figure 5.8:** Results of the feature subset selection procedure for *Knocking*, *Throwing* and *Lifting* actions. Each matrix element $(i, j)$ represents feature $\phi_j$ derived from the primitive $\phi_i$. White elements represent features which were selected by the algorithm as informative. Primitive-local features (left) and action-global features (right) are separated by a grey dashed line.

### 5.4.2 Results

Figure 5.8 visualises the subset selection results. White squares symbolise features which have been selected, black squares denote unselected features. See Appendix E for a detailed list of every feature. Remember that the feature vectors contain both primitive-local and action-global features. These portions are visualised in the left and right part of the figure respectively. Appendix E also lists the results of a variation of the experiment where only action-global features were considered during the selection process.

Overall, the routine selected 24% of the features as predictive. In the remainder of the section I will present the distributions of selected features in more detail. In particular, I will analyse the selected features along the dimensions listed on page 116. For each dimension I will visualise the results in the form of bar charts. As features can be of one of a number of categories making up each dimension, the plotted quantities represent the fraction of features picked from a certain category. That is, for a feature category $\kappa$, the plotted quantities represent

$$\frac{\text{number of features of category } \kappa \text{ picked by the algorithm}}{\text{total number of features of category } \kappa} \tag{5.3}$$

Note that the results given here were derived by considering all emotions simultaneously. Appendix E also gives results for a variation of the experiment where features were selected which are particularly informative for distinguishing *one particular* emotion from all the others. The features selected there are hence informative for distinguishing one particular emotion.

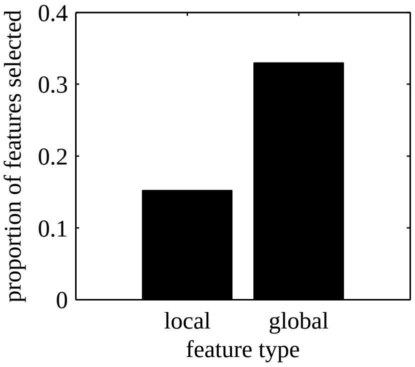

**Figure 5.9:** Proportion of selected features: primitive-local versus action-global.

### Primitive-local versus action-global features

Figure 5.9 confirms the impression we get when looking at Figure 5.8: the algorithm strongly favours global features over features derived from the motion primitives. While roughly 1/3 of the global features were picked, only 15% of the primitive-based features are deemed informative.

This is an important result. It gives us information to evaluate my earlier claim that detailed motion understanding at the primitive level will in fact help the recognition of emotions. Much of the emotional information seems to be readily available at the action level. Action-level features might also be preferred because they are likely to be less noisy as they are computed over a longer time span. However, the importance of primitive-based features should not be totally dismissed. Roughly 1/3 of all selected features are primitive-local and therefore add some significant information to emotion recognition. Indeed, in the final section of this chapter I will show that emotion recognition results can be improved significantly by adding primitive-based features to a machine learning-based framework.

### Statistical measures

The vast majority of features is defined in terms of simple statistical quantities over either a motion primitive or the whole action. Figure 5.10(a) illustrates that by far the least informative of the statistical measure is the sample minimum. Minimum-based features were only picked 10% of the time while the most informative, median-based features were picked in over 35% of the cases. The variation of the signal seems to be informative as well as standard deviation-based features were picked in nearly 25% of the cases while maximum-based features were selected in one out of five cases.

This result confirms the common view that median and standard deviation are robust and informative measures to describe a time series of measurements. They are clearly favoured by the feature selection. Minimum and maximum values capture the extrema of

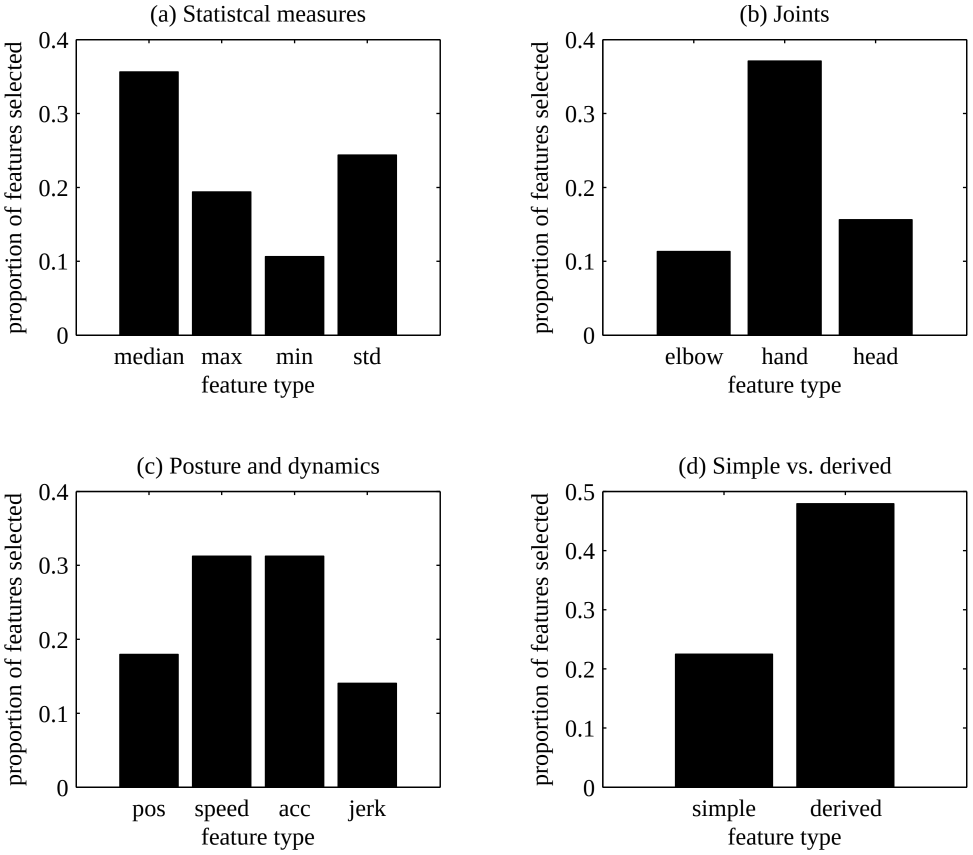

**Figure 5.10:** Proportion of selected features: results grouped according to four different dimensions.

a time series. These measures are more likely to be subject to noise which may inhibit a strong correlation with the emotional labels.

### Joints

Three joints were used to derive the statistical features. In particular the vectors contain position, speed, acceleration and jerk information for the right hand and elbow, as well as posture information for the head. Figure 5.10(b) shows that by far the most informative of the three joints is the hand, with 37% of the features selected as informative. Much less additional information seems to come from the elbow (11%) and head (16%).

One reason for the low significance of the elbow might be its naturally large correlation with the hand motions. It is also natural for the end-effector to exhibit the largest amount of movement, which further favours the hand-based statistics over the elbow-based values. The fact that head posture holds important information in everyday scenarios is encouraging. The importance of head posture and head gestures for emotion recognition has recently been reported by El Kaliouby and Robinson [KR04].

### Posture and dynamics

The features capture a range of static and dynamic information. Posture features, as captured in terms of the joints' position data are selected 18% of the time (see Figure 5.10(c)). More informative than the static information seem speed and acceleration which are selected with a frequency of 31%. Importantly, speed and acceleration supply features which are both highly informative in their own right. Both therefore seem to supply complementing information for emotion recognition. A 14% selection rate means that jerk-related information is nearly as important as static information.

### Simple versus derived features

Figure 5.10(d) shows that a large proportion of 48% from the derived features (such as joint phase difference and energy statistics) has been selected. This is in contrast to the smaller fraction of 23% of simple dynamics-based features which the algorithm selected as informative.

These numbers have to be interpreted with some care. First of all, there are significantly less derived features to choose from. Secondly, while the simple features can be expected to be highly inter-correlated from the way that they are defined, the derived features are more diverse in nature. Both of these factors can explain why the derived features seem to be favoured so significantly. Nevertheless, this graph clearly illustrates that deriving more complex features from a number of simple measurements can add significant information relevant to emotion recognition.

### Activity

The above graphs and analyses give us useful insights of a general nature and grouped by various factors of interest. What they do not capture, however, is how individual motion structures influence which features are selected as informative. From Figure 5.8 we can see that not the same features are informative for all action categories. Also, the total number of features selected can vary between action categories.

One of the most striking action-specific observations is that in the second *Knocking* primitive ("repeated knocking"), primitive-local features seem to carry a lot of information relevant to emotion. As the only primitive, the second *Knocking* primitive draws more emotion-specific information from primitive-local than from action-global features.

## 5.5 Emotion Recognition

We are now ready to look at the emotion recognition framework developed for this research. The developed classifiers are feature vector-based and use the statistical and

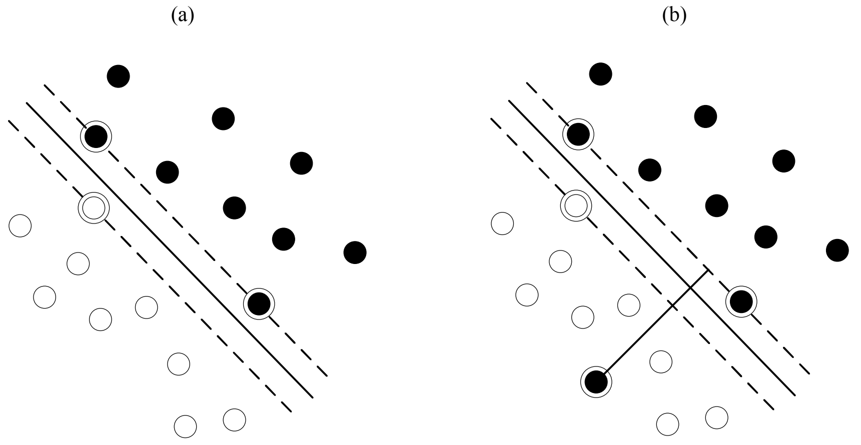

**Figure 5.11:** SVMs define a maximum margin for (a) separated training data; (b) unseparated training data. Support vectors are highlighted through an extra circle.

derived features $\boldsymbol{\phi}$ discussed in the previous section. The high-dimensional feature vectors supplied to the classifiers capture the motion dynamics for a whole action and for a certain primitive. Based on these measurements, the classifier is expected to supply an inferred emotion label for a *whole* action. It therefore needs to be able to integrate information from multiple primitives. Features will generally be unbiased before supplying them to the learning and classification algorithms. However, in order to illustrate the benefit of modelling individual differences I will also evaluate the classification performance for biased feature vectors.

In the remainder of this section I am going to present in detail Support Vector Machines as the classification technique used for this research. I will also present a detailed evaluation of the classification performance. The focus of the evaluation is the ability of emotion classifiers to generalise beyond previously unseen subjects. I will also evaluate how influential different aspects such as action category, choice of features, use of motion bias and classifier complexity are with respect to emotion classification.

### 5.5.1 Discriminative classification using SVMs

I am using Support Vector Machines (SVMs) to classify motion features into emotional classes. As opposed to other classification techniques such as Neural Networks, SVMs are inherently linear classifiers. They define decision boundaries which take the form of lines in 2D and hyperplanes in higher-dimensional feature spaces. In their primitive form SVMs are binary classifiers defining a single decision boundary between two classes. Samples on one side of the boundary are classified as class A, while samples on the other side are associated with class B. I will describe how I generalise SVMs to more than two emotion classes after covering the basic SVM principles below.

SVMs are trained on a number of labelled training samples. The training formalism is based on statistical learning theory and in contrast to previously seen generative methods aims to be as discriminative as possible. The derived decision boundary is therefore defined as the one which maximises the margin (separation) between the samples from both classes [Vap00]. For a well-separated set of samples the margin is defined as the sum of the distances between the decision boundary and the points closest to it from either class (Figure 5.11(a)). These closest points, or support vectors, play an important role as they provide all the evidence for defining the decision boundary. In most real-world datasets, however, classes are not perfectly separated (Figure 5.11(b)). In this case samples on the "wrong" side of the boundary also become support vectors, but their ability to influence the decision surface is limited by a regularisation constant (normally denoted as $C$). There is overwhelming evidence that the maximum-margin property of SVMs leads to an excellent generalisation ability beyond the training examples [Bur98], even if the data is not well-separated. As mentioned before, generalisation between subjects is one of my main requirements for the developed classifier. As other popular classification techniques tend to suffer from overfitting, especially when the dimensionality of the feature space is high, SVMs provide a good solution for the classification problem at hand.

### SVM kernels

SVMs seem very limited in their discriminatory power by the restriction to define linear boundaries. Indeed, in Section 3.5 we saw that increasing the boundary complexity beyond linear can improve the discriminatory power of a classifier considerably. In order to relax the restriction of linearity, SVMs make use of a more general technique known as the kernel method. While a set of points from different classes may not be linearly separable in their original feature space, they might be linearly separable after projecting them to some higher-dimensional space using a mapping function $H$. Formally, given two feature vectors $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ in the original space, a kernel $K$ is an expression for the dot product of $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ in some higher-dimensional space, defined by $H$:

$$K(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) = H(\boldsymbol{\phi}_1) \cdot H(\boldsymbol{\phi}_1) \tag{5.4}$$

SVM training and classification can be expressed solely in terms of dot-products between data points. For example, finding the maximum margin decision boundary can be shown to be equivalent to maximising the following function with sample vectors $\boldsymbol{\phi}$, class labels $e$ and optimisation parameters $\alpha$ [Bur98]:

$$L_D \equiv \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j e_i e_j \boldsymbol{\phi}_i \cdot \boldsymbol{\phi}_j \tag{5.5}$$

In order to find a non-linear boundary, the dot-product in Equation 5.5 can be replaced by the corresponding kernel function.

$$\begin{aligned} L_D &\equiv \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j e_i e_j H(\boldsymbol{\phi}_i)\cdot H(\boldsymbol{\phi}_j) && (5.6)\\ &\equiv \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j e_i e_j K(\boldsymbol{\phi}_i, \boldsymbol{\phi}_j) && (5.7)\end{aligned}$$

Note how the use of a kernel function $K$ allows me to work in a higher-dimensional feature space without needing to supply a concrete definition of $H$. If SVMs use the kernel method when finding a linear boundary, the derived boundaries will usually be non-linear when projected back into the original feature space. In fact, the derived boundaries can be almost arbitrarily complex. In this dissertation I am using two families of kernel functions which give rise to well-studied boundaries:

$$\begin{aligned} K(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) &= (\boldsymbol{\phi}_1 \cdot \boldsymbol{\phi}_2 + 1)^p && (5.8)\\ K(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) &= e^{-||\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2||^2/2\sigma^2} && (5.9)\end{aligned}$$

Equation 5.9 gives rise to a polynomial decision boundary of degree $p$. I will use boundaries of various complexities with $p$ varying between 1 and 4. Equation 5.9 results in complex decision boundaries equivalent to a Gaussian Radial Basis Function (RBF)-based classifier [Vap00, Bur98]. The resulting decision boundaries can be visualised as the isosurfaces of Gaussian mixtures in the original feature space. While the SVM finds an optimal linear decision boundary in the projected space, the parameters of the resulting boundaries in feature space are affected in complex ways. Most notably, the number and centres of the Gaussian RBFs, as well as a number of other parameters are determined automatically through SVM learning, making the kernel method very powerful. Additionally, the derived solutions are guaranteed to be globally optimal. Using the kernel method allows me to evaluate the relative merits of decision boundaries of varying complexities in Section 5.5.2.

### Feature normalisation

It is usually regarded as good practice to normalise features which would otherwise have very different orders of magnitude. This is especially important because I am dealing with position, speed, acceleration and jerk data. Differences between position and derivative data can easily reach five orders of magnitude. Before subjecting data to SVMs, I therefore first normalise each feature by subtracting the sample mean and dividing by the standard deviation as computed over all samples from all subjects. Note that this normalisation happens *before* subtracting any motion bias. I included this normalisation step explicitly in the recognition pipeline pictured in Figures 5.2 & 5.6.

### Combined classification

Two levels of combination need to be achieved on top of the primitive binary SVMs:

1. emotion level: combine multiple binary SVMs to create a 4-emotion classifier
2. action level: combine classifications based on multiple primitives to obtain a classification for the whole action.

Multiple ways of generalising primitive binary SVMs to multi-class SVMs have been suggested in the past [HL02]. While a generalisation of the SVM formalism to multiple classes exist, these are more computationally complex and have not been as widely studied and evaluated. Also, in most cases using simpler voting schemes performs as well or even better in practice [HL02, CS02]. I therefore use a voting scheme based on binary SVMs, each trained on two emotion classes at a time. In the combined classification task, each binary SVM contributes a single vote for one class. The class with the most votes is returned as the final classification result. This combination scheme also generalises well to the second problem of combining multiple primitives.

More formally, I am training a family of SVMs $M^m_{e_1,e_2}$. The classifier $M^m_{e_1,e_2}$ aims to find the maximum margin between emotion classes $e_1$ and $e_2$ for motion primitives of type $m$. $e_1$ and $e_2$ are drawn from a set of emotions $\mathcal{E}$. Once these binary classifiers have been trained, I classify a sequence of motion primitives $m_1 \dots m_n$ into one of $|\mathcal{E}|$ emotions as follows:

1. Set primitive index $i = 1$ and set the $|\mathcal{E}|$ elements of the vote vector **votes** to 0.
2. Apply all pairwise SVMs $M^{m_i}_{e_1,e_2}$. Add the votes for each emotion class $e$ to $\mathbf{votes}(e)$.
3. If $i < n$ then $i = i + 1$, goto 2.
4. Classify the whole motion by the majority vote in **votes**, *i.e.* return emotion class $\arg\max_e \mathbf{votes}(e)$. Resolve ties by choosing a class with maximum vote count at random.

## 5.5.2 Experimental evaluation

In this section I will examine the performance of my emotion recognition framework in detail. In Section 3.5 we noticed that emotional body expressions can be rather person-specific. In order to produce meaningful quantitative results which can predict the performance on new subjects, it is important to construct the experiments carefully. The classification performance will be evaluated as recognition rates on a test set after the classifiers have been trained on a separate training set. It is very common to obtain good results when data samples from the same subject are included in both training and test sets. This allows the classifier to adjust to the individual expression of the subject which improves recognition rates. I am, however, interested in estimating the performance of the technique for a wider population, and beyond the 30 subjects contained in the Glasgow corpus. It is therefore important only to test on subjects whose data has not been included in the training set. In order to maximise the amount of evaluation data while minimising the amount of bias, I am using 10-fold cross validation (10F-CV) for all experiments. At any one time classifiers will be trained on data from 27 subjects and tested on the data from the remaining 3 subjects. Results from 10 of these folds are combined to give the final recognition rates. Folds are created such that each subject is included in the test set exactly once.

Within each fold, SVM kernel parameters (regularisation parameter and RBF kernel width) are optimised through a logarithmic grid search. In order to minimise the chance for overfitting and sacrificing generalisation performance [Sch05b], the parameters are estimated through 3-fold cross validation using data from the 27 training subjects. That is, for each of the evaluated parameter combinations, an SVM is trained on (3 groups of) 18 subjects' samples and evaluated on the remaining 9 subjects. The best performing parameter combination is picked for the outer 10F-CV. All experiments were carried out with custom Matlab code and an integration of Joachim's SVMlight implementation of binary SVMs [Joa99].

In order to enable a post-hoc statistical analysis of the results I am recording results in 2 ways:

1. Average recognition performance in the form of confusion matrices.
2. Recognition rates for each individual.

It is important to realise that this procedure only tests for the generalisation ability of the SVM-based *emotion* classifiers. We saw in Figure 5.6 that motion bias is a factor which is estimated in a separate adaptation step. In experiments 1 and 3 I am assuming that the motion signature of a person was accurately estimated before attempting emotion recognition and hence perform 10F-CV using unbiased features.

### Experiment 1: Action categories and feature sets

Experiment 1 considers the recognition accuracy for the four different action categories in the Glasgow corpus: *Knocking*, *Throwing*, *Lifting*, *Walking*. The goal is to find evidence for potential differences in emotion recognition performance between different action categories. The feature vectors are extracted based on the motion primitives defined in Section 4.3.2. I am also examining four different feature sets:

1. action-global: a feature vector derived from the whole action
2. primitive-local: a number of feature vectors, each derived from a single primitives only
3. global and local: a number of feature vectors derived from the motion primitives, each concatenated with the global features for the whole action as shown in Figure 5.7.
4. subset-selected: a feature vector as defined through the optimal subsets found in Section 5.4.2.

Note that for the latter two feature vector definitions, the global features are included repeatedly in the combined feature vectors for different motion primitives of the same action category. I am performing 10F-CV for all $4 \times 4$ motion category and feature definition combinations. All features are *unbiased* before supplying them to SVM training and classification.

### Experiment 2: Classifier complexity

Experiment 2 considers the effect of classifier complexity on the recognition performance. Our observations in Section 3.5 suggested that the complex feature spaces arising from emotional body motions can benefit from classifiers that are able to define more complex decision boundaries. SVMs allow me to experiment with different complexities easily by changing the kernel function $K$. I am running 10F-CV on all upper-body motions (*Knocking*, *Throwing*, *Lifting*) with the following kernel functions:

1. $K(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) = \boldsymbol{\phi}_1 \cdot \boldsymbol{\phi}_2 + 1$ (linear)
2. $K(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) = (\boldsymbol{\phi}_1 \cdot \boldsymbol{\phi}_2 + 1)^2$ (quadratic polynomial)
3. $K(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) = (\boldsymbol{\phi}_1 \cdot \boldsymbol{\phi}_2 + 1)^3$ (cubic polynomial)
4. $K(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) = (\boldsymbol{\phi}_1 \cdot \boldsymbol{\phi}_2 + 1)^4$ (quartic polynomial)
5. $K(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) = e^{-||\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2||^2/2\sigma^2}$ (RBF)

In the latter condition I need to set the RBF width parameter $\sigma$. As is common practice, I am using a logarithmic grid search to find the value of $\hat{\sigma}$ which maximises recognition performance on a representative subset of the data. For my data $\hat{\sigma} = 2.0$. As the feature vector I am using the set of unbiased action-global features.

**Experiment 3: Motion bias modelling**

Experiment 3 looks in detail at the effect of motion bias modelling on emotion recognition. In Section 5.3.1 we saw how the feature distributions are affected by the inclusion of motion bias in the feature derivations. I will examine the influence of this factor on all actions in the corpus in detail. I am performing 10F-CV in the following conditions:

1. biased features: no motion bias is subtracted from the features
2. predicted bias: motion bias as estimated from *TDown* motions is subtracted
3. unbiased features: motion bias as computed on all *Knocking* motions of a subject is subtracted

The latter condition tests for the significance of the correlation effects found in Section 5.3.2. In particular, we saw strong correlations between corresponding features (features on the diagonal) between the *TDown* and *Knocking* motions. I therefore train a linear regression model for each of the *Knocking* features based on the corresponding *TDown* feature. During the 10F-CV the regression coefficients are estimated based on the 27 subjects used for SVM training. The *TDown* features of the test subjects are then used to predict their *Knocking* bias. During the classification the normalised features $\hat{\boldsymbol{\phi}}$ are then computed using this estimated bias. Because the plots suggest that there might be some non-Gaussian outliers, I am using iteratively reweighted least squares to robustly estimate model coefficients [HW77]. I am only using the action-global features for this experiment.

In addition, Appendix F gives a stand-alone evaluation of this regression-based bias prediction approach in terms of root mean squared error between true and predicted bias. It shows that prediction accuracy varies between different action categories. While it is best for *Knocking* actions, prediction for *Walking* is particularly poor.

### 5.5.3 Results

In this section I am presenting the results for Experiments 1 – 3. All experiments give rise to confusion matrices which are presented where appropriate. In particular, Tables 5.3, 5.3 and 5.6 report a confusion matrix for each experimental condition. Each confusion matrix reports the classification performance for all emotions in the order *neutral*, *happy*, *angry* and *sad*. Correct recognition rates are found on the diagonals while the remaining entries represent the rates of incorrect classifications. I will also give detailed results arising from the statistical significance analysis in Experiments 1 and 3. A detailed discussion and analysis of the results is given in Section 5.6.

**Table 5.3:** Emotion recognition results by action and feature set. Individual confusion matrices list emotions in the order *neutral*, *happy*, *angry*, *sad*.

| | global features | | | | local features | | | | global and local features | | | | subset selected features | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Knocking* | **0.78** | 0.15 | 0.00 | 0.07 | **0.79** | 0.14 | 0.00 | 0.07 | **0.80** | 0.13 | 0.00 | 0.07 | **0.78** | 0.16 | 0.00 | 0.06 |
| | 0.17 | **0.75** | 0.07 | 0.01 | 0.22 | **0.68** | 0.07 | 0.03 | 0.15 | **0.77** | 0.05 | 0.02 | 0.16 | **0.79** | 0.04 | 0.01 |
| | 0.01 | 0.10 | **0.89** | 0.00 | 0.03 | 0.12 | **0.85** | 0.00 | 0.01 | 0.06 | **0.93** | 0.00 | 0.01 | 0.06 | **0.93** | 0.00 |
| | 0.08 | 0.01 | 0.00 | **0.91** | 0.14 | 0.01 | 0.00 | **0.85** | 0.07 | 0.01 | 0.00 | **0.92** | 0.08 | 0.02 | 0.00 | **0.90** |
| | **average rate: 0.83** | | | | **average rate: 0.79** | | | | **average rate: 0.86** | | | | **average rate: 0.85** | | | |
| *Throwing* | **0.49** | 0.20 | 0.00 | 0.31 | **0.63** | 0.14 | 0.00 | 0.23 | **0.52** | 0.25 | 0.00 | 0.23 | 0.63 | 0.20 | 0.00 | 0.17 |
| | 0.32 | **0.58** | 0.02 | 0.08 | 0.38 | **0.49** | 0.01 | 0.12 | 0.30 | **0.65** | 0.00 | 0.05 | 0.30 | **0.60** | 0.01 | 0.09 |
| | 0.01 | 0.01 | **0.97** | 0.01 | 0.01 | 0.05 | **0.94** | 0.00 | 0.01 | 0.01 | **0.97** | 0.01 | 0.00 | 0.01 | **0.98** | 0.01 |
| | 0.36 | 0.05 | 0.00 | **0.59** | 0.36 | 0.01 | 0.00 | **0.62** | 0.25 | 0.05 | 0.00 | **0.70** | 0.27 | 0.03 | 0.00 | **0.70** |
| | **average rate: 0.66** | | | | **average rate: 0.67** | | | | **average rate: 0.71** | | | | **average rate: 0.73** | | | |
| *Lifting* | **0.75** | 0.13 | 0.01 | 0.11 | **0.72** | 0.14 | 0.01 | 0.13 | **0.74** | 0.12 | 0.00 | 0.14 | **0.75** | 0.12 | 0.00 | 0.13 |
| | 0.16 | **0.70** | 0.13 | 0.01 | 0.16 | **0.69** | 0.13 | 0.01 | 0.14 | **0.71** | 0.13 | 0.02 | 0.14 | **0.78** | 0.07 | 0.01 |
| | 0.01 | 0.16 | **0.83** | 0.00 | 0.05 | 0.28 | **0.66** | 0.01 | 0.01 | 0.16 | **0.82** | 0.00 | 0.01 | 0.10 | **0.89** | 0.00 |
| | 0.20 | 0.01 | 0.00 | **0.79** | 0.23 | 0.01 | 0.00 | **0.76** | 0.18 | 0.01 | 0.00 | **0.81** | 0.14 | 0.01 | 0.00 | **0.85** |
| | **average rate: 0.77** | | | | **average rate: 0.71** | | | | **average rate: 0.77** | | | | **average rate: 0.82** | | | |
| *Walking* | **0.70** | 0.18 | 0.01 | 0.11 | | | | | | | | | **0.79** | 0.15 | 0.01 | 0.05 |
| | 0.12 | **0.77** | 0.11 | 0.00 | | | | | | | | | 0.12 | **0.80** | 0.08 | 0.00 |
| | 0.01 | 0.12 | **0.87** | 0.00 | | | | | | | | | 0.01 | 0.09 | **0.90** | 0.00 |
| | 0.15 | 0.01 | 0.00 | **0.84** | | | | | | | | | 0.06 | 0.01 | 0.00 | **0.93** |
| | **average rate: 0.79** | | | | | | | | | | | | **average rate: 0.85** | | | |

**Table 5.4:** Significance levels for difference of mean recognition rates using action-global features, primitive-local features, local and global features and an optimal feature subset. Statistically significant differences are highlighted in bold.

| | *Knocking* | *Throwing* | *Lifting* | *Walking* |
|---|---|---|---|---|
| global *vs*. local | **0.0020** | 0.5270 | **0.0001** | - |
| global *vs*. global-local | 0.0050 | **0.0020** | 0.9690 | - |
| global *vs*. subset | 0.0970 | **0.0001** | **0.0010** | **0.0001** |
| local *vs*. global-local | **0.0001** | 0.0320 | **0.0001** | - |
| local *vs*. subset | **0.0001** | 0.0030 | **0.0001** | - |
| global-local *vs*. subset | 0.4070 | 0.2330 | **0.0010** | - |

### Experiment 1: Emotion recognition from different action categories

Table 5.3 shows the confusion matrices for all four action categories and four feature sets. Note that the *Walking* cycles are not broken down further into primitives. There are therefore no primitive-based results for *Walking*. We can observe a fairly large spread of recognition rates between 66% for *Throwing* motions using global features to 86% for *Knocking* motions using a full set of action-global and primitive-local features. The overall recognition performance for *Knocking* and *Walking* samples is better than for *Lifting* and *Throwing* samples.

It is also clear that the different feature sets give rise to different recognition rates. For example for *Throwing* actions, we have an average recognition rate of 66% from primitive-local features while a selected subset of features boosts the rate to 73%. In order to judge which of the effects are statistically significant I am using the fact that the experiments produced repeated measurements of recognition performance for the same subject under different conditions.

A sound answer can be obtained from a repeated-measures analysis of variance. In fact, this shows a highly significant effect of the feature subsets on the mean recognition performance ($p < 0.0001$). To investigate the difference between individual conditions further, I conducted a post-hoc analysis using repeated-measures t-tests testing for the difference in sample means. Table 5.4 shows the pairwise significance values from the t-tests. Bold entries highlight significant differences between conditions. The significance threshold was adjusted to $0.05/19 = 0.0026$ to account for multiple comparisons. After correction many of the conditions are significantly different. In particular, I find that using local-only features results in significantly worse recognition than using action-global features for both *Knocking* and *Lifting* motions with a decrease in average recognition rates of 4% and 6% respectively. For all action categories combining action-global with primitive-local features improves or at least preserves recognition accuracy. For *Throwing* actions the improvement is significant although it is also consistent for *Knocking* actions. Further

**Table 5.5:** Emotion recognition results by action and Kernel function $K$. Individual confusion matrices list emotions in the order *neutral*, *happy*, *angry*, *sad*.

| *K* | *Knocking* | | | | *Throwing* | | | | *Lifting* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| linear | **0.78** | 0.15 | 0.00 | 0.07 | **0.49** | 0.20 | 0.00 | 0.31 | **0.75** | 0.13 | 0.01 | 0.11 |
| | 0.17 | **0.75** | 0.07 | 0.01 | 0.32 | **0.58** | 0.02 | 0.08 | 0.16 | **0.70** | 0.13 | 0.01 |
| | 0.01 | 0.10 | **0.89** | 0.00 | 0.01 | 0.01 | **0.97** | 0.01 | 0.01 | 0.16 | **0.83** | 0.00 |
| | 0.08 | 0.01 | 0.00 | **0.91** | 0.36 | 0.05 | 0.00 | **0.59** | 0.20 | 0.01 | 0.00 | **0.79** |
| | **average rate: 0.83** | | | | **average rate: 0.66** | | | | **average rate: 0.77** | | | |
| *quadratic* | **0.85** | 0.11 | 0.00 | 0.04 | **0.59** | 0.18 | 0.00 | 0.23 | **0.73** | 0.12 | 0.01 | 0.14 |
| | 0.25 | **0.68** | 0.06 | 0.01 | 0.36 | **0.55** | 0.03 | 0.06 | 0.19 | **0.66** | 0.13 | 0.01 |
| | 0.01 | 0.13 | **0.86** | 0.00 | 0.01 | 0.03 | **0.95** | 0.01 | 0.01 | 0.20 | **0.79** | 0.00 |
| | 0.10 | 0.02 | 0.00 | **0.88** | 0.36 | 0.09 | 0.00 | **0.55** | 0.19 | 0.03 | 0.00 | **0.78** |
| | **average rate: 0.82** | | | | **average rate: 0.66** | | | | **average rate: 0.74** | | | |
| *cubic* | **0.86** | 0.10 | 0.00 | 0.04 | **0.63** | 0.16 | 0.00 | 0.20 | **0.77** | 0.10 | 0.01 | 0.12 |
| | 0.25 | **0.66** | 0.08 | 0.01 | 0.42 | **0.48** | 0.03 | 0.07 | 0.26 | **0.61** | 0.12 | 0.01 |
| | 0.01 | 0.13 | **0.86** | 0.00 | 0.02 | 0.03 | **0.95** | 0.00 | 0.03 | 0.23 | **0.74** | 0.00 |
| | 0.08 | 0.02 | 0.00 | **0.90** | 0.41 | 0.06 | 0.00 | **0.53** | 0.23 | 0.01 | 0.00 | **0.76** |
| | **average rate: 0.82** | | | | **average rate: 0.65** | | | | **average rate: 0.72** | | | |
| *quartic* | **0.85** | 0.11 | 0.00 | 0.04 | **0.64** | 0.18 | 0.00 | 0.18 | **0.78** | 0.11 | 0.01 | 0.11 |
| | 0.28 | **0.64** | 0.07 | 0.01 | 0.39 | **0.49** | 0.03 | 0.09 | 0.28 | **0.58** | 0.13 | 0.02 |
| | 0.01 | 0.13 | **0.85** | 0.01 | 0.02 | 0.03 | **0.94** | 0.01 | 0.02 | 0.23 | **0.75** | 0.00 |
| | 0.07 | 0.03 | 0.00 | **0.90** | 0.35 | 0.10 | 0.01 | **0.54** | 0.23 | 0.02 | 0.01 | **0.74** |
| | **average rate: 0.81** | | | | **average rate: 0.65** | | | | **average rate: 0.71** | | | |
| *RBF* | **0.80** | 0.14 | 0.00 | 0.06 | **0.62** | 0.16 | 0.00 | 0.22 | **0.74** | 0.12 | 0.00 | 0.14 |
| | 0.18 | **0.75** | 0.06 | 0.01 | 0.26 | **0.64** | 0.04 | 0.07 | 0.17 | **0.70** | 0.11 | 0.01 |
| | 0.01 | 0.10 | **0.89** | 0.00 | 0.00 | 0.04 | **0.95** | 0.01 | 0.00 | 0.15 | **0.84** | 0.00 |
| | 0.06 | 0.01 | 0.00 | **0.93** | 0.36 | 0.05 | 0.00 | **0.58** | 0.20 | 0.01 | 0.00 | **0.79** |
| | **average rate: 0.84** | | | | **average rate: 0.70** | | | | **average rate: 0.77** | | | |

selecting strongly correlated features from the global-local vectors improves mean recognition rate in most cases. The improvement is significant for both *Walking* and *Lifting* actions.

### Experiment 2: Classifier complexity

The recognition rates for all motion categories and kernel types are given in Table 5.5. In general we observe that the linear and RBF kernels perform the best. With increasing polynomial complexity, the recognition rates do not improve across all motion categories.

### Experiment 3: Motion bias modelling

Table 5.6 shows the confusion matrices for all four action categories and three treatments of motion bias. It is immediately clear that not modelling individual motion bias (first

**Table 5.6:** Emotion recognition results by action and treatment of motion bias. Individual confusion matrices list emotions in the order *neutral*, *happy*, *angry*, *sad*.

| | biased | | | | bias-adapted | | | | unbiased | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Knocking* | **0.40** | 0.20 | 0.12 | 0.28 | **0.59** | 0.21 | 0.04 | 0.16 | **0.78** | 0.15 | 0.00 | 0.07 |
| | 0.29 | **0.34** | 0.24 | 0.13 | 0.36 | **0.36** | 0.23 | 0.05 | 0.21 | **0.71** | 0.07 | 0.02 |
| | 0.16 | 0.27 | **0.55** | 0.02 | 0.04 | 0.26 | **0.70** | 0.00 | 0.01 | 0.14 | **0.85** | 0.00 |
| | 0.23 | 0.11 | 0.02 | **0.64** | 0.12 | 0.05 | 0.00 | **0.83** | 0.08 | 0.03 | 0.00 | **0.89** |
| | **average rate: 0.48** | | | | **average rate: 0.62** | | | | **average rate: 0.81** | | | |
| *Throwing* | **0.38** | 0.10 | 0.07 | 0.45 | **0.51** | 0.19 | 0.02 | 0.28 | **0.63** | 0.16 | 0.00 | 0.21 |
| | 0.34 | **0.21** | 0.18 | 0.26 | 0.30 | **0.37** | 0.16 | 0.17 | 0.32 | **0.58** | 0.03 | 0.07 |
| | 0.04 | 0.14 | **0.79** | 0.03 | 0.04 | 0.13 | **0.80** | 0.03 | 0.00 | 0.04 | **0.95** | 0.01 |
| | 0.28 | 0.04 | 0.05 | **0.63** | 0.32 | 0.12 | 0.03 | **0.53** | 0.38 | 0.07 | 0.00 | **0.55** |
| | **average rate: 0.51** | | | | **average rate: 0.55** | | | | **average rate: 0.68** | | | |
| *Lifting* | **0.53** | 0.21 | 0.04 | 0.22 | **0.56** | 0.17 | 0.04 | 0.22 | **0.70** | 0.16 | 0.00 | 0.14 |
| | 0.34 | **0.33** | 0.26 | 0.08 | 0.27 | **0.40** | 0.27 | 0.06 | 0.17 | **0.59** | 0.24 | 0.00 |
| | 0.09 | 0.31 | **0.60** | 0.00 | 0.06 | 0.32 | **0.60** | 0.01 | 0.03 | 0.26 | **0.71** | 0.00 |
| | 0.22 | 0.04 | 0.01 | **0.73** | 0.19 | 0.02 | 0.00 | **0.79** | 0.18 | 0.00 | 0.00 | **0.82** |
| | **average rate: 0.54** | | | | **average rate: 0.59** | | | | **average rate: 0.70** | | | |
| *Walking* | **0.41** | 0.29 | 0.12 | 0.18 | **0.37** | 0.33 | 0.12 | 0.18 | **0.69** | 0.19 | 0.00 | 0.12 |
| | 0.12 | **0.47** | 0.29 | 0.12 | 0.11 | **0.48** | 0.29 | 0.12 | 0.15 | **0.70** | 0.13 | 0.02 |
| | 0.07 | 0.17 | **0.69** | 0.06 | 0.07 | 0.17 | **0.70** | 0.06 | 0.02 | 0.11 | **0.86** | 0.01 |
| | 0.21 | 0.08 | 0.03 | **0.68** | 0.24 | 0.06 | 0.03 | **0.67** | 0.14 | 0.02 | 0.00 | **0.84** |
| | **average rate: 0.56** | | | | **average rate: 0.55** | | | | **average rate: 0.77** | | | |

**Table 5.7:** Significance levels for difference of mean recognition rates using biased features and linear regression-based adapted features. The differences for the action *Knocking* are statistically significant at $p = 0.01$.

| Action | p-value |
|---|---|
| *Knocking* | **0.0001** |
| *Throwing* | 0.143 |
| *Lifting* | 0.079 |
| *Walking* | 0.168 |

column of results) is significantly worse than removing the motion bias completely (last column). The interesting question is, however, to what extent predicting the motion bias using regression can account for not computing the motion bias explicitly. This question is answered with the results in the middle column. For *Knocking* motions, bias prediction works particularly well, with mean recognition rate improved by 14%. Improvements for *Throwing* and *Lifting* are still notable, but much smaller at 5% and 4% respectively. For *Walking* the prediction of motion bias is not possible. This is expected, as the prediction is based on the upper body action *TDown* whereas *Walking* is a lower-body action.

In order to see whether the improvements are statistically significant, I am again using the subject-wise recognition rates to perform a repeated-measures t-test between the biased and adapted conditions. Table 5.7 shows the resulting significance values. We see that for *Knocking* the difference is statistically significant. For *Throwing* and *Lifting* the improvements are just outside the region of statistical significance.

## 5.6 Discussion

At a large scale, we notice that emotion recognition performance can vary between different action categories. In Experiment 1 we see the best results for *Knocking* and *Walking* motions. The worse recognition rates are observed for *Throwing* actions. A closer inspection of the confusion matrices reveals that the recognition rate for *sad Throwing* is surprisingly low when compared to the other action categories. I believe that this manifests the large complexity of interactions between the emotion and activity factors. Sad motions usually appear slow and slack, which makes them easy to recognise in other action contexts. In order to throw an object, however, the arm has to move fast enough in order to physically throw an object. It is therefore very likely that the usually slack motions are more constrained for *sad Throwing*, making them very similar to *neutral Throwing* motions. The difference between the recognisability of different emotion classes is a general pattern which emerges from Experiment 1. While *angry* and *sad* movements tend to be recognised relatively well, there often is a higher confusion rate between *neutral* and *happy* motions.

The recognition results for *Knocking* motions allow me to compare between human and machine performance for recognising emotions from isolated everyday actions. In an experiment Pollick *et al.* found that humans can distinguish between five emotions with an accuracy of 59% for point-light and 71% for full video stimuli of human *Knocking* motions [PPBS01]. These figures illustrate that even humans are far from perfect at classifying emotion in everyday actions. My algorithm achieves a rate of between 83% and 85% depending on exact feature definitions. In order to compare these numbers, the figures need to be normalised to account for different correct recognition rates due to chance. These normalised recognition rates are also known as efficiencies $\eta$ [HS01] and can be defined as

$$\eta = \frac{\text{achieved recognition rate}}{\text{recognition rate expected by chance}} \tag{5.10}$$

Expressed in terms of recognition efficiency, humans achieve $\hat{\eta}_{pl} = 2.95$ and $\hat{\eta}_v = 3.55$ for point-light and video stimuli respectively whereas my SVM-based approach can achieve efficiencies $3.32 \leqslant \eta_{ub} \leqslant 3.40$. For biased and predicted-bias features the efficiencies are $\eta_b = 1.92$ and $\eta_{pb} = 2.48$. For the best configuration my approach therefore approaches human capabilities from video data. This is significant as there is arguably much more information in the video data than in the joint position data my algorithms have to work with. It is also worth pointing out that all of the recognition rates are significantly above the chance level of 25%. In particular, even $\eta_b$ is much larger than 1, indicating that the features I extract contain significant emotion information, even if the individual motion bias is not removed before classification.

Modelling individual motion bias is, however, a significant contributor to accurate emotion recognition. This stresses once more the importance of taking subject or user-specific differences into consideration. My definition of a motion signature captures the differences effectively and Experiment 3 confirms that the achieved recognition improvements are significant. Given the relatively large improvement of average recognition performance from 48% to 81% actively modelling user biases will almost certainly make a notable difference in practice. As getting an accurate action-specific bias estimate can be difficult in a real-world setup, I also investigated the prediction of action-specific bias from a set of subject-specific calibration motions. Experiment 3 confirms that using linear regression to predict the bias can bring a significant improvement. It can therefore be used in practical settings to boost recognition performance during early-stage usage of a system. If a real-world system is used for longer time periods by the same person, a more accurate action-specific bias can be calculated based on actually observed occurrences of the action by the user.

In Experiment 1 we saw that the choice of features has a significant impact on the emotion recognition accuracy. Generally, it has been suggested that using uncorrelated sets of features which are highly predictive of the emotion class will yield the best recognition performance. This view was largely confirmed with the feature subsets selected to reflect this heuristic outperforming other feature sets. Experiment 1 also showed, however, that features selected in this way do not necessarily have to yield a statistically significant improvement over more redundant feature sets. This is due to the SVMs' maximum margin property which tends to counteract overfitting which is often seen in other classifiers — a phenomenon known as the *curse of dimensionality*. We also found that although incorporating primitive-based features into the recognition framework can improve recognition rates significantly, the average improvements seen are rather small. Whether this amount of improvement is worth the extra computational effort of extracting motion primitives is application-dependent. It is certainly worth bearing in mind that nearly as good results can be achieved by simply looking at actions globally.

Interestingly, Experiment 2 found that one of the best kernels to use for this classification problem are those of the simplest kind — linear. This suggests the following conclusions. Similarly to the emotionally expressive motions I examined in Section 3.5 the feature space has a rather simple structure which can be captured well by linear class boundaries. In addition to the polynomial family of kernels I also investigated the use of RBF kernels which in all cases performed as well or better than the polynomial kernels. However, a careful optimisation of kernel parameters was necessary in order to avoid overfitting and optimise generalisation beyond the seen training examples. This procedure can be time-consuming, especially since for RBF kernels two parameters need to be optimised simultaneously. However, for *Throwing* motions the RBF kernels outperformed the polynomial kernels by 4%.

Finally, the set of experiments has shown that even in highly controlled environments where actions occur in isolation and subject to strict instructions, emotion recognition cannot be expected to be 100% accurate. This has been confirmed by psychological experiments analysing human ability to recognise emotions from body cues [PPBS01, DTLM96, DSBB04]. We have seen that the display of emotions is a very fuzzy and noisy process, further complicated by issues of personal differences and action context. I have shown, however, that significant progress can be made by modelling these factors explicitly. In the next chapter I will discuss another complicating factor and consider actions occurring in natural sequences rather than perfectly isolated samples.

# Chapter 6

# Emotion recognition from sequential motions

In the previous chapter I made one crucial assumption: we are dealing with isolated actions. In this chapter I am going to present an approach for recognising emotions from body motions which occur in connected sequences. In real-world environments people perform actions in fluid succession, sometimes interleaving parts of actions such as getting up and walking while drinking a cup of coffee. Any of these actions may individually reveal some information about a person's emotional state. In order to make maximum use of the information contained in every single action, an emotion recognition system needs to understand this sequential and parallel nature of actions.

I will build on the frameworks for action and emotion recognition as developed in Chapters 4 and 5. At the heart of the chapter lies the development of a method for segmenting parallel streams of action sequences into isolated actions. It is important to remember at this point that the sequences in the Glasgow corpus all follow the same order of eight actions. I will nevertheless develop a method which could in principle analyse and segment sequences of any number and order of actions. The developed method will, however, support the notion of an action grammar such that prior knowledge of the transition probabilities between actions can be taken into account during segmentation. Evaluation of the segmentation quality on sequences other than complete or segments of the original Glasgow sequences is beyond the scope of this dissertation.

Instead, the bulk of the evaluation will focus on emotion recognition. In a final experiment I therefore evaluate how my emotion recognition technique copes with the sequential data. As the final result of this work I combine recognition results from multiple actions to arrive at a classification for a whole action sequence.

## 6.1 The difference between isolated and sequential data

Sequential actions are not merely isolated actions concatenated together. When humans start executing actions in natural sequences they produce motions which are very different from when they are instructed to perform motions in isolation. Algorithms for analysing natural motions need to be able to cope with this difference. In many practical scenarios requiring sequential pattern recognition it is relatively easy to obtain isolated examples of the patterns to be distinguished. For example, in a speech recognition scenario with limited vocabulary these might be isolated words [MR81a, MR81b, RL85], in a sign language recognition task, these might be isolated signs [AG98, VM97]. In my case, they are isolated samples of the actions to be recognised as presented in Chapter 4. Previous work in the speech and gesture recognition communities shows, however, that training data obtained from isolated samples can be significantly different from real-world sequential data [WH99, HH99]. I am going to discuss two kinds of problem encountered when moving from isolated to sequential motions: coarticulation and more general change of appearance.

### 6.1.1 Coarticulation

One of the first things one notices when looking at sequential actions is how fuzzy the boundaries between actions become. The end of one action blends fluently into another. Remember that in the Glasgow corpus subjects were asked to perform upper body actions *Knocking*, *Lifting* and *Throwing* separated by *Walking* between different locations in the capture area. In many cases upper body actions are started while the subject is still walking. It is virtually impossible to observe a sequence of motions in a natural setting where the person first fully stops at the location of action with arms by the side, then performs the action, lowers the arms, and starts walking again. By splitting the body into independent upper and lower body regions (see Section 4.2.2) my action models are able to represent upper and lower body actions happening in parallel. It is insufficient, however, to assume that sequential actions simply correspond to isolated actions occurring in parallel. Action characteristics become different due to the preceding and following actions — an effect known as coarticulation [HH99].

Coarticulation has been studied extensively for the case of human speech. A particular issue investigated is the noticeable change that occurs when going from slow or temporally isolated speech to fast and connected speech more commonly found in conversations. In faster speech the pronunciation of speech segments is frequently altered based on surrounding speech segments [Gie92]. Sometimes whole segments are not pronounced altogether. These processes are called *assimilation* and *elision*. They are among the major factors which make connected speech recognition difficult and they are also posing a challenge to connected action recognition.

One of the classic explanations for these phenomena has been the "principle of least effort". The principle states that these effects are observed because the vocal apparatus tends to do as little work as possible while still producing recognisable utterances. Many patterns relating to human activity have been analysed in the light of the principle of least effort [Zip49]. Coarticulation effects are clearly observable when comparing the isolated and sequential motions in the Glasgow corpus. All of the Glasgow sequences contain a *Knocking* action, followed by *Walking* to a table behind the subject and a subsequent *Throwing* action of an object on the table. What tends to happen is that while the subject is finishing the *Knocking* action he or she turns the body to start the *Walking* towards the table. This means that the *Knocking* action is interleaved with a turn of the body. While isolated *Knocking* actions always end with a complete lowering of the arms, this does not occur in the sequential motions — the example of an elision. Before the subject reaches the table, he or she will start extending the right hand to pick up the object. This anticipatory movement is usually significantly longer than the arm lifting segment during the isolated *Throwing* actions — an assimilation. I will account for these effects by adjusting action model parameters according to observations from the action sequences (Section 6.2.3)

### 6.1.2 Change of appearance

The coarticulation effects are clearly a source for change of appearance of actions. However, coarticulation may not be the only reason why the appearances of actions change. Subjects were instructed very carefully in the isolated action recordings. For example, this included instructions for the body orientation towards objects with which to interact. This naturally suggests that there should be far less variation in isolated action samples. Another source for variation in the appearance might come from the factor of direct repetition. While many of the isolated samples were recorded as immediate repetitions of each other, the sequential actions are interspersed with other actions. This, again, suggests that isolated actions may appear more uniform than the sequential ones.

Before discussing the significance of this kind of change in appearance, I will present some quantitative evidence. Here I focus on the amount of *variation* observed across different subjects and repetitions of the same action. A very good set of descriptors capturing the dynamics of the actions is the set of features used for emotion recognition (Section 5.2). It comprises position, speed, acceleration and jerk as calculated from the trajectories of different body joints. Based on those I compute the sample standard deviations $\sigma_{I,\phi}$ and $\sigma_{C,\phi}$ over all isolated and connected action samples respectively. For every feature $\phi$, $\sigma_{I,\phi}$ and $\sigma_{C,\phi}$ denote how much $\phi$ varies across different instances of the same action. Finally, I partition the set of features $\Phi$ into $\Phi_C$ and $\Phi_I$ such that

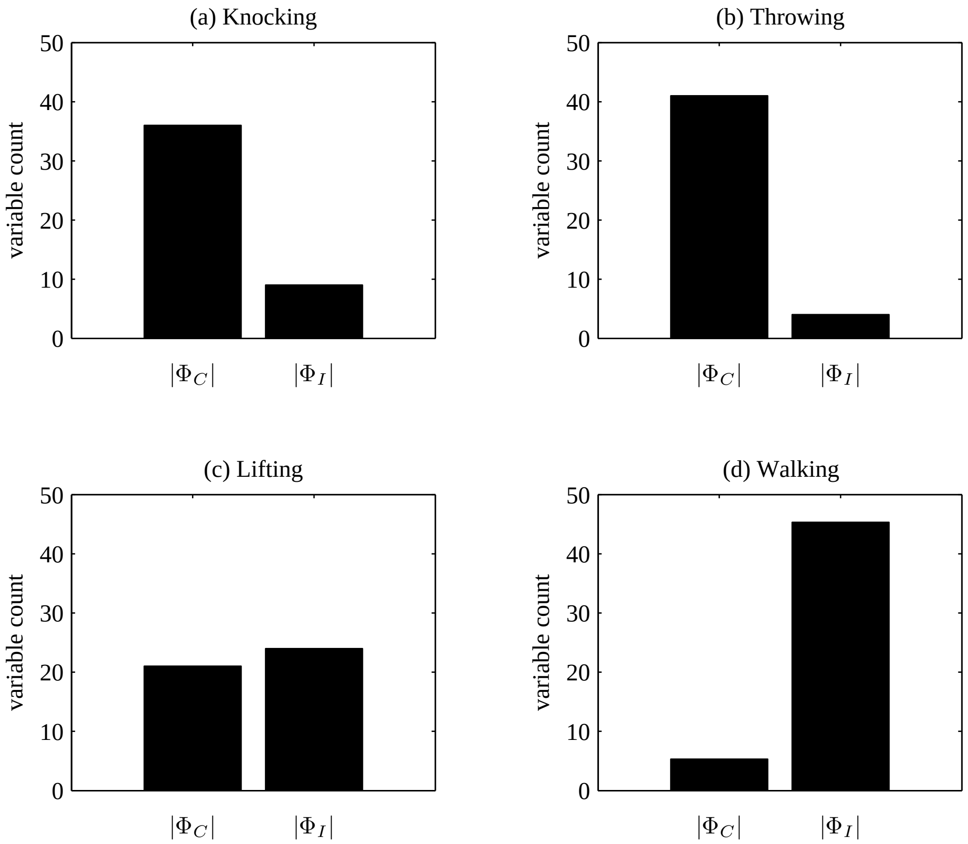

**Figure 6.1:** Difference in variation between isolated and connected actions.

$$\phi \in \Phi_C \quad \Leftrightarrow \quad \sigma_{C,\phi} > \sigma_{I,\phi} \tag{6.1}$$

$$\phi \in \Phi_I \quad \Leftrightarrow \quad \sigma_{C,\phi} \leqslant \sigma_{I,\phi} \tag{6.2}$$

Figure 6.1 shows $|\Phi_C|$ and $|\Phi_I|$ for every action category. If the appearance of isolated and connected cases was similar, each pair of bars would be of roughly equal height. We see, however, that for both *Knocking* and especially *Throwing* actions, the variation is a much larger in the sequential case. This confirms my informal observations which predicted less uniform motions in the sequential case. The variation of *Lifting* motions, on the other hand, appears similar in both conditions. Interestingly, *Walking* motions show a reverse trend. Here samples show less variability in the sequential case. This can be explained by the setup of the recordings. While isolated *Walking* was recorded as long stretches of walking, including turns, examples of sequential *Walking* were very short, straight line walks between two well-defined points. This clearly limited the amount of observable variation.

Differences in appearance are relevant for two reasons. Firstly, my action models have only been trained on isolated data and significant deviation from the observed variations could lead to poorer action recognition performance. In Section 6.2.4 we will see that there is evidence for this and I am proposing to use bootstrapping in order to adjust model parameters accordingly. Secondly, a change of appearance is likely to entail a difference in the appearance of different emotions as well. In Section 6.3.2 we will see that reusing the emotion models trained on isolated actions is still able to distinguish emotions in sequential actions to some extent. However, adjusting for the change in appearance can improve emotion recognition markedly.

## 6.2 Segmentation framework

In this section I describe in detail the methods I have used for segmenting sequences of connected actions into isolated actions. The major requirement for a segmentation framework is its reuse of individual action models as trained on isolated actions. This is important for three reasons:

- Within the scope of this dissertation I would like to exploit the fact that I have already developed a set of action models in Chapter 3. The segmentation method should make use of this information as much as possible.
- Starting out with isolated models allows me to evaluate their performance for connected actions directly. I will discuss how isolated models can be adapted to perform better on connected actions. The gained insights are very illuminating for understanding the differences between emotions expressed through isolated and connected actions.
- Being able to build a connected action recogniser from isolated action models would make the construction of a real-world system easier. In particular, if a new action category needs to be supported, it suffices to collect training data and train an according model in isolation.

### 6.2.1 Problem statement

In Chapter 3 I developed a set of action models $\lambda_c$. I will call this full set of individual models $\Lambda = \{\lambda_c\}$. The goal of segmentation can be formulated as follows. Given a sequence of observations $\boldsymbol{v}(1) \dots \boldsymbol{v}(T)$, find the most likely sequence of actions $\lambda(1) \dots \lambda(L)$ which best explains the observations. Here $L$ denotes the total number of actions in the sequence which is unknown in general. I furthermore need to record the action boundaries $b(0) \dots b(L)$ such that $\lambda(i)$ explains frames $b(i-1) \dots b(i)$. Typically $b(0) = 1$ and $b(L) = T$.

At this point it is important to remember that the human body is modelled as two distinct regions and that actions from the sets $\mathcal{C}_U$ and $\mathcal{C}_L$ can happen in parallel (see Section 4.2.2).

I will therefore derive two independent sequences of actions $\lambda_U(l)$ and $\lambda_L(l)$ as well as their respective boundaries.

### 6.2.2 Solution using Level Building

The segmentation problem can be solved efficiently using Level Building (LB) [MR81c]. It treats a sequence of $L$ actions as a succession of $L$ levels $l = 1 \dots L$. At each individual level $l$ the LB algorithm performs a Viterbi decoding in much the same way I used it to find the most likely isolated action in Section 4.2.3. The Viterbi parse is calculated for each model $\lambda_c$. Crucially, however, the Viterbi parse at level $l + 1$ can make use of the probabilities calculated at level $l$ in order to start $\lambda(l)$ at the most likely frame. It also stores the maximising Viterbi probability for every time frame $t$. In particular, there are three outputs computed at every level $l$:

1. $\hat{\lambda}(l, t)$: the model which is most probable to have given rise to the observation $\boldsymbol{v}(t)$
2. $\hat{P}(l, t)$: the Viterbi probability with which $\hat{\lambda}(l, t)$ produced $\boldsymbol{v}(t)$
3. $\hat{B}(l, t)$: the time frame at which $\hat{\lambda}(l, t)$ picked up from level $l - 1$.

$\hat{B}(l, t)$ represents the crux of the LB algorithm. It allows me to connect the independent Viterbi parses at each level. Once $\hat{\lambda}(L, t)$ and $\hat{B}(L, t)$ have been computed, the most likely sequence of actions can be found recursively as

$$b(l) = \begin{cases} T & \text{for } l = L \\ \hat{B}(l+1, b(l+1)) - 1 & \text{for } 1 \leqslant l < L \\ 1 & \text{for } l = 0 \end{cases} \tag{6.3}$$

$$\lambda(l) = \hat{\lambda}(l, b(l)) \tag{6.4}$$

In order to allow the connection from level $l - 1$ to level $l$ a simple change is necessary to the standard Viterbi algorithm. In its normal form the Viterbi algorithm computes a Dynamic Programming matrix $V$ one time frame at a time by making use of results from the previous time frame as shown in Figure 6.2 (a). In particular, matrix element $V(i, t)$ denotes

$$V(i, t) = \max_{\omega(1)\dots\omega(t-1)} P(\boldsymbol{v}(1) \dots v(t-1), \omega(1) \dots \omega(t-1), \boldsymbol{v}(t), \omega(t) = \omega_i) \tag{6.5}$$

That is, $V(i, t)$ stores the probability of the most likely sequence of states up to time frame $t$ and ending in state $\omega(t) = \omega_i$. Note that in order to account for models which have

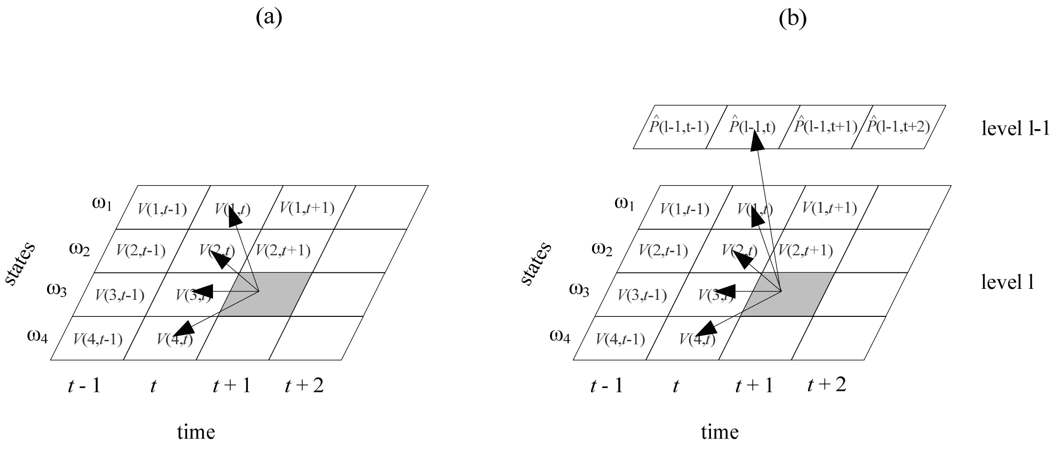

**Figure 6.2:** Dynamic programming computations for (a) standard Viterbi algorithm (b) Level Building.

outputs of different dimensionality, my implementation makes use of normalised Viterbi likelihoods as described in Section 4.2.3. The matrix is usually calculated recursively through dynamic programming as

$$V(i,t+1) = \max_j a_{j,i} \times P(\boldsymbol{v}(t)|i) \times V(j,t) \tag{6.6}$$

This calculation assumes that transitions can only occur between the hidden states of a single HMM and not between multiple models themselves. However, the probabilities $\hat{P}(l,t)$ can be used in the standard Viterbi optimisation in order to include transitions from the previous level. I accomplish this by taking $\hat{P}(l,t)$ into consideration in the recursive update of $V(i,t)$. In particular, for levels $l > 1$ I calculate $V(i,t)$ as

$$V(i,t+1) = \max \begin{cases} \max_j a_{j,i} \times P(v(t)|i) \times V(j,t) \\ t(\hat{\lambda}(l-1,t),c,l) \times \hat{P}(l-1,t)) \end{cases} \tag{6.7}$$

The crucial recursive steps are illustrated in Figure 6.2. We see that in the latter case the algorithm can decide to maximise the Viterbi probability by either transitioning from one of the internal states of the model *or* by transitioning from the most likely end state of the previous level.

The computation makes use of an additional transition function $t(i,j,l)$ which captures the probability of transitioning from $\lambda_i$ to $\lambda_j$ at level $l$. $t$ can be used to enforce the notion of an *action grammar*. More importantly for us, the function can be used in order to only allow a certain action sequence. Assume that, as in the case of the Glasgow corpus, we have a weak labeling $\lambda(1)\dots\lambda(L)$ for each sequence where each $\lambda(i) \in \Lambda$. Then a transition function $\hat{t}$ can be defined to enforce this order as follows:

$$\hat{t}(i,j,l) = \begin{cases} 1 & \text{if } \lambda(l-1) = \lambda_i \text{ and } \lambda(l) = \lambda_j \\ 0 & \text{otherwise} \end{cases} \tag{6.8}$$

I will use this method to compute the *best possible* segmentation achievable with LB. Without any prior information, however, these probabilities will be uniform and independent of $l$: $t(i,j,l) = \frac{1}{|\mathcal{E}|}$. This will be the definition of $t$ for the majority of conditions during my evaluation.

### Determining the optimal number of levels

Given the above formalism it is possible to find the optimal segmentation given a set of candidate models $\Lambda$, an observation sequence $\boldsymbol{v}$ and the number of actions in the sequence $L$. In order to segment the action sequences in the Glasgow corpus all this information is available as I know which actions can appear in each sequence and that each sequence has a length of $L = 8$. However, in a more general scenario, the number of levels is not known and needs to be determined automatically. The simplest approach to solve this problem is to pick the segmentation which has the largest likelihood. That is, I pick $\hat{l}$ such that

$$\hat{l} = \arg\max_{l} \hat{P}(l,T) \tag{6.9}$$

There is, however, a problem with this approach. HMMs are notoriously bad at modelling the duration of actions [Rab02]. In particular, as the number of levels is increased, shorter and shorter action snippets are inserted into the segmentation. Those snippets might only be a few frames long and therefore far shorter than any of the observed training examples. Because HMMs do not explicitly model duration, the Viterbi likelihood can still increase if these short snippets are inserted, making it very difficult to pick the correct $\hat{l}$.

In order to prevent this, I am introducing a regularisation term into the Viterbi optimisation which explicitly captures the duration of actions [RL85]. I am modelling the duration of an action category $c$ as a normally distributed random variable. The mean $\mu_c$ and standard deviation $\sigma_c$ are estimated from the isolated cases. When calculating $\hat{P}(l,t)$ and $\hat{\lambda}(l,t)$ I take the length of the action into consideration. It enters the calculation as a multiplicative term $P_c(d_t)^\gamma$ where $d_t$ denotes the length of the action in frames and $\gamma$ is a constant weight which controls the importance assigned to the regularisation term. $P_c$ is defined as the normal density

$$P_c(d_t) = \frac{1}{\sqrt{2\pi}\sigma_c} e^{-\frac{(d_t-\mu_c)^2}{2\sigma_c^2}} \tag{6.10}$$

Figure 6.3 illustrates the effect of picking different values of $\gamma$ on the values of $\hat{P}(l,T)$ for the Glasgow sequences. Remember that each sequence consists $l = 8$ component

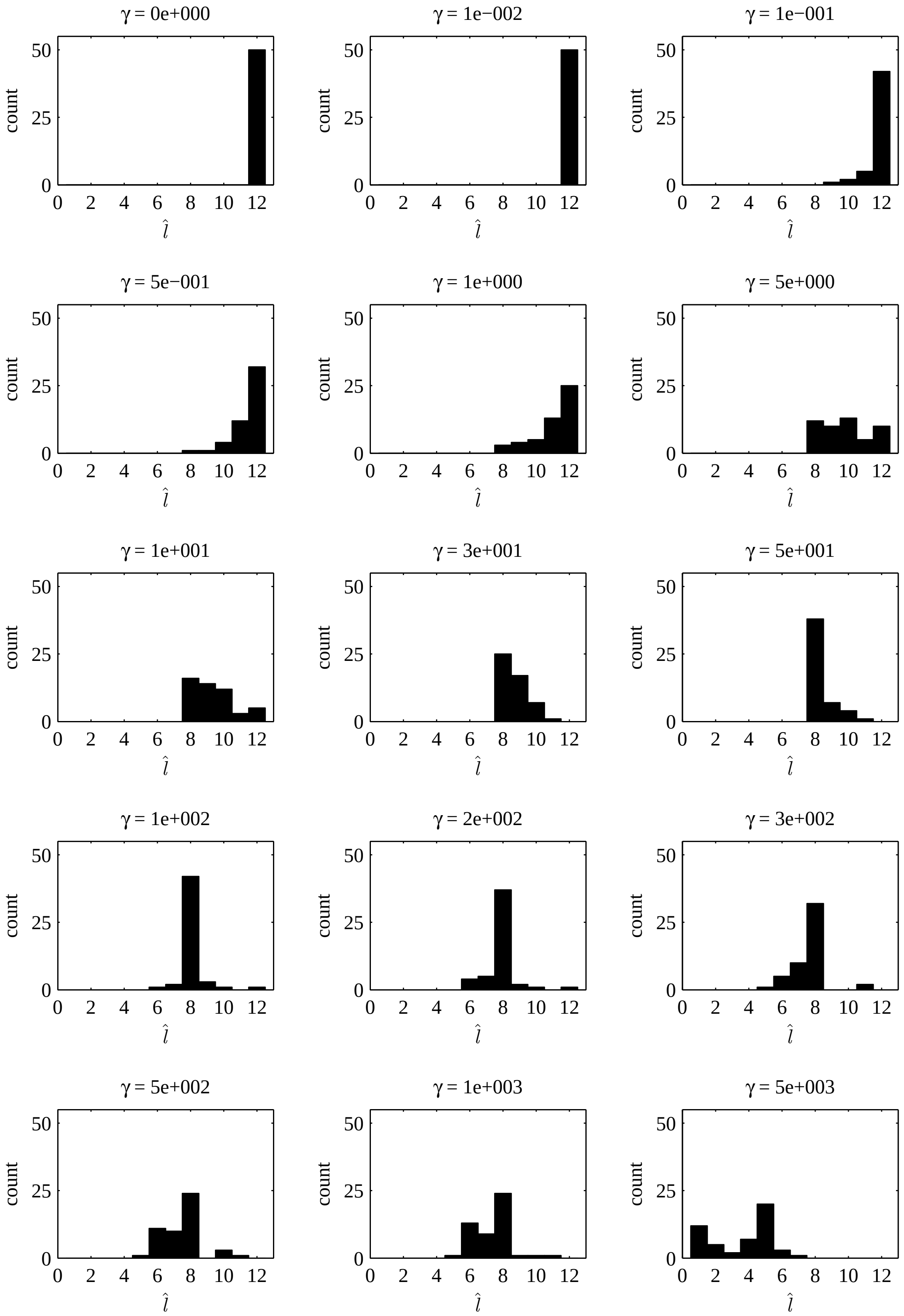

**Figure 6.3:** Distributions of $\hat{l}$ after introducing a duration-based regularisation term. $\gamma$ controls the importance assigned to the duration constraint.

actions. For this experiment I chose 50 sequences from the corpus at random and ran the LB algorithm for levels $1 \leqslant l \leqslant 12$. In particular, the figure shows a distribution of $\hat{l}$ calculated from all of the sequences. We see that for very small $\gamma$, adding new levels simply increases the overall $\hat{P}(l, T)$ and $\hat{l} = 12$. As $\gamma$ is increased, the distribution slowly moves away from the maximum. For $\gamma = 100$, 42 out of the 50 were judged correctly to have $\hat{l} = 8$. As I increase $\gamma$ further the weight given to the duration heuristic dominates the calculations and in the limit any evidence arising from the observations themselves is ignored. For large $\gamma$ the distribution diverges and favours segmentations which best agree with the action durations observed for the *isolated* cases. For the given scenario a value of $\gamma \approx 100$ strikes an optimal balance for taking into consideration the appearance and duration constraints.

In order to illustrate the effectiveness of the approach for sequences containing different numbers of actions I hand-segmented one of the sequences into its eight component actions. I then created eight new sequences where the $i$th sequence contains actions 1 to $i$. On each I ran the LB algorithm with $\gamma = 100$ and $L = 12$. The resulting curves for $\hat{P}(l, T)$ are shown in Figure 6.4. In all cases $\hat{l}$ coincides with the correct number of actions in the sequence.

### 6.2.3 Retargeting HMMs using bootstrapping

Reusing action models trained on isolated samples brings many advantages in practice. I have shown how to build a connected action recogniser from the isolated models using Level Building. However, models trained on isolated samples are not guaranteed to represent the same action in a sequential setting well. In Section 6.1 I discussed changes in appearance between isolated and sequential motions. Another issue is the tendency of HMMs to overfit to insufficiently varied training data. I demonstrated this effect in Section 4.2.4. As I will show in my evaluation in Section 6.2.4 using the resulting models in a Level Building algorithm can give suboptimal results.

In this section I describe and demonstrate an approach to training optimal models for connected action recognition. It works by iteratively refining the models constructed from isolated examples. One of the major observations is that actions are relatively long in duration, on the order of hundreds of frames. Only a small fraction of those tend to be affected by the surrounding actions and show coarticulation effects. For example, only the initial arm lift during a *Throwing* action is prolonged by an anticipatory effect when comparing the isolated to sequential cases. This leads me to the assumption that large parts of the isolated models are in fact still representative of the actions as they appear in a sequence. The models trained on isolated actions can therefore be seen as the first in a series of approximations to the optimal sequential models.

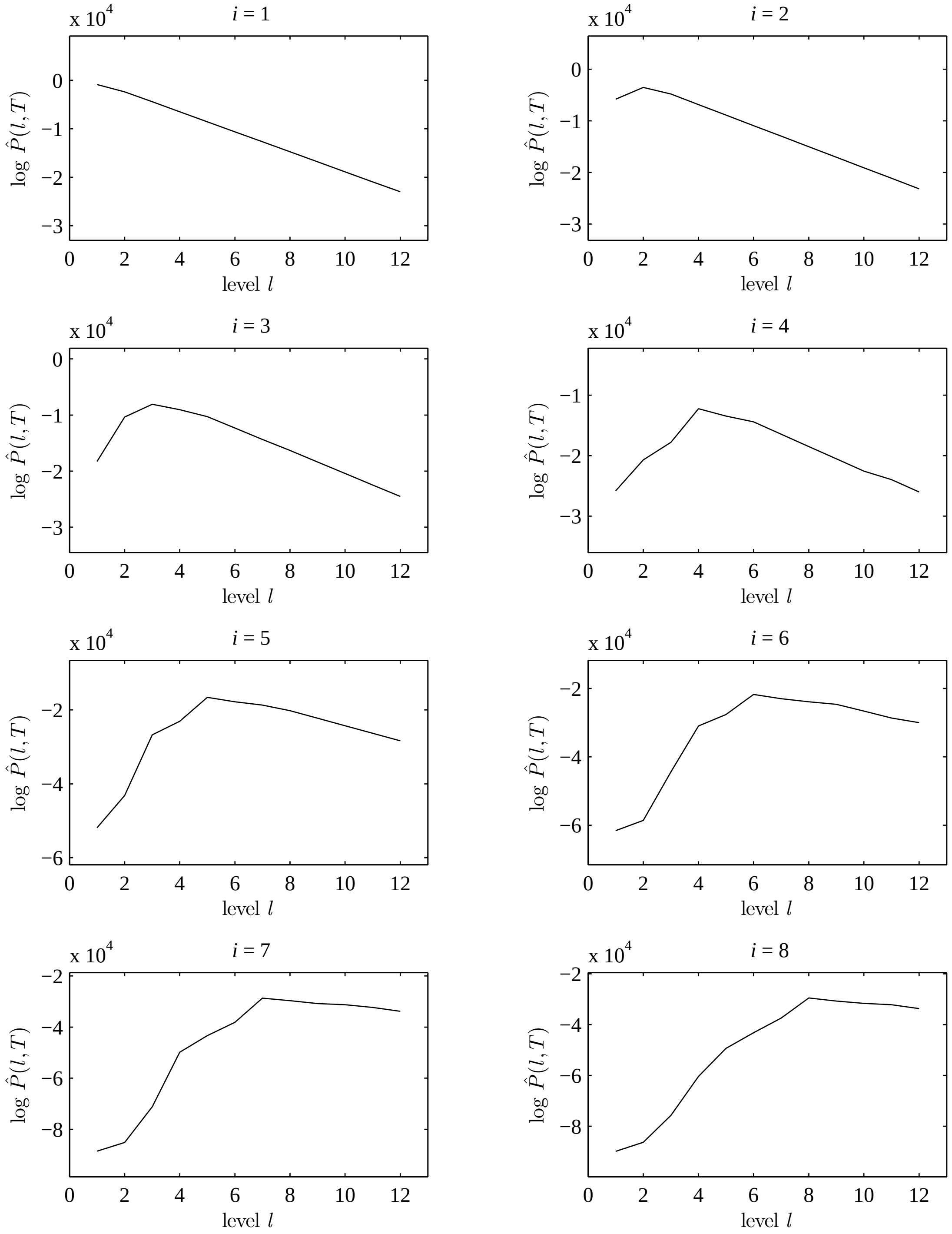

**Figure 6.4:** $\hat{P}(l,T)$ as computed for a number of sequences composed of $1 \leqslant i \leqslant 8$ actions. In each case the LB algorithm was run up to a maximum level of $L = 12$.

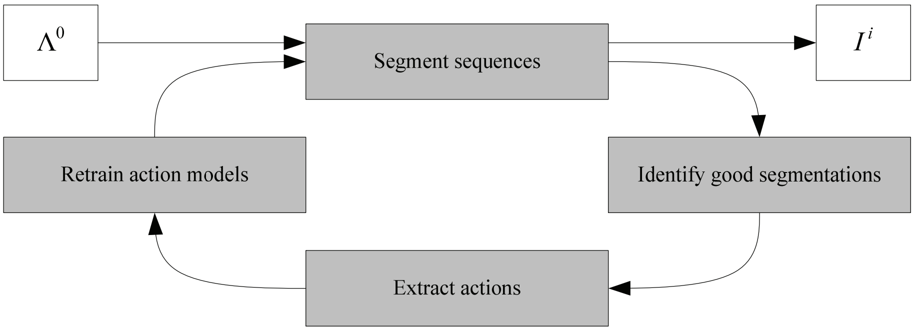

**Figure 6.5:** Bootstrapping loop used to improve action models.

### Bootstrapping

My algorithm uses a bootstrapping approach to iteratively refine models. Bootstrapping algorithms can be applied to an optimisation problem which aims to find an optimal model $\hat{\Lambda}$ based on a collection of data $\mathcal{V}$. Ultimately, the goal is to apply $\hat{\Lambda}$ to $\mathcal{V}$ in order to extract some higher level information $\hat{\mathcal{B}}$. In order to apply bootstrapping one usually supplies some initial information $\mathcal{B}^0$ believed to be extractable from $\mathcal{V}$. Bootstrapping then proceeds by searching $\mathcal{V}$ for evidence of $\mathcal{B}^0$. Based on the found patterns of evidence, a bootstrapping algorithm then builds a first approximation model $\Lambda^1$ which is capable of extracting $\mathcal{B}^0$. In an iterative procedure $\Lambda^1$ is then applied to $\mathcal{V}$ to find new information $\mathcal{B}^1$. Figure 6.5 illustrates the iterative process of finding new models $\Lambda^{i+1}$ based on the information $\mathcal{B}^i$ which then give rise to new information $\mathcal{B}^{i+1}$.

In my case, I start with an initial model approximation $\Lambda^0$ which is the set of HMMs derived from isolated action samples. I also make use of some initial information which states that all sequences respect a certain order of actions which was prescribed during data recording. I call these sequences *weakly labelled* as the boundaries for the actions are unknown. Instead, the exact segmentation boundaries comprise the information $\hat{\mathcal{B}}$ I am seeking. In this case, the set of data $\mathcal{V}$ is simply the set of action sequences to be segmented. I can find a first approximation $\mathcal{B}^0$ by applying $\Lambda^0$ using Level Building to find a first segmentation $\mathcal{B}^0$. I then need to update the action models accordingly. I do this by retraining the HMMs based on the sequential actions which were segmented into the correct order of actions (see Figure 6.5). This data is picked as training data to retrain the HMMs in $\Lambda^0$ to yield a new set of models $\Lambda^1$. The procedure can be repeated to yield new iterations of segmentations $\mathcal{B}^i$ based on the action models $\Lambda^i$.

## 6.2.4 Segmentation results

### Upper and lower body actions

For the first part of the evaluation of my segmentation approach I will focus on the ability of my framework to model actions of the upper and lower body happening in

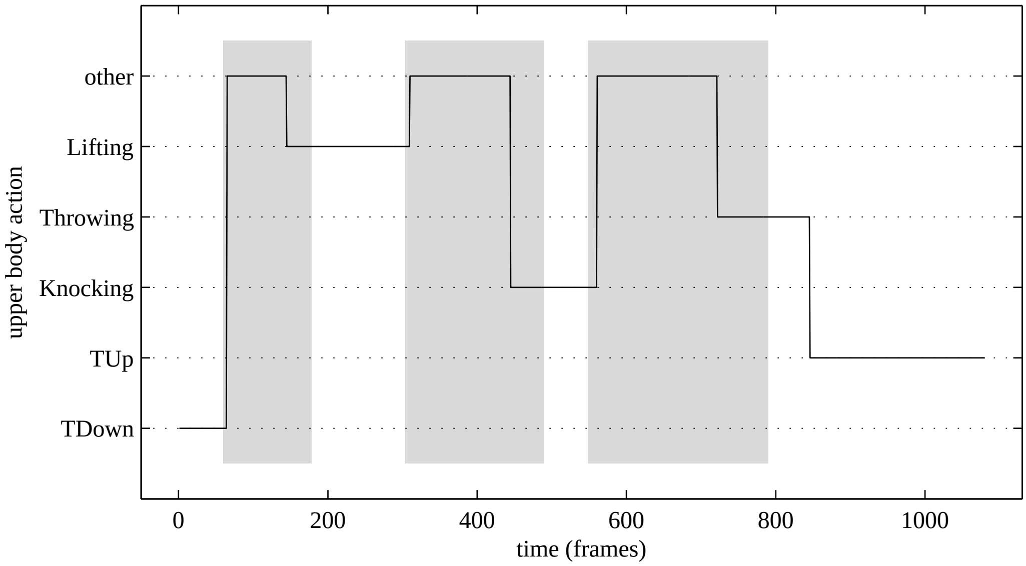

**Figure 6.6:** Segmentation result of an action sequence from the Glasgow corpus. Upper body actions are shown in the line plot. For lower body actions the grey areas denote *Walking* actions while white areas denote *other*.

parallel. This is facilitated by defining actions to be either an upper or lower body action (see Section 4.2.2) and parsing a sequence of observations into two independent action sequences. This is especially important when analysing connected actions as anticipatory movements and other coarticulation effects mean that subjects start a new action while still carrying out parts of the previous action.

Figure 6.6 shows the parse of a representative sequence picked from the Glasgow corpus. The segmentation results for the upper and lower body were taken after two bootstrapping iterations. The primary line shows the parse of the sequence into a number of upper body actions. We see that the sequence has been parsed into the correct order. The background of the figure represents the lower body actions. In particular, the grey areas denote frames where a *Walking* action was detected. As expected, they largely correspond to the *other* action of the upper body. However, in the transition phases we can see clear overlaps between *Walking* and upper body actions such as *Knocking*, *Lifting* and *Throwing*. The HMM-based segmentations therefore reflect my informal observations that subjects start many upper body actions before having stopped walking. The framework is able to successfully deal with upper and lower body actions appearing in parallel.

### Segmentation quality

Even though the segmentation in Figure 6.6 reflects our expectations, it does not allow me to evaluate the accuracy of the segmentations rigorously. A good segmentation accuracy is, however, important for my emotion recognition framework. In order to evaluate the segmentation quality achievable through Level Building, I first segmented the sequences

**Table 6.1:** Segmentation quality achieved at several iterations of bootstrapping for (a) unbiased initialisation (30 subjects) (b) biased initialisation (3 subjects).

(a) Experiment 1

| iteration | $\zeta$ | $\bar{\alpha}$ | $\sigma_\alpha$ |
|---|---|---|---|
| $\mathcal{B}_1^0$ | 0.65 | 0.90 | 0.11 |
| $\mathcal{B}_1^1$ | 0.86 | 0.92 | 0.07 |
| $\mathcal{B}_1^2$ | 0.85 | 0.91 | 0.07 |
| $\mathcal{B}_1^3$ | 0.85 | 0.90 | 0.07 |
| $\mathcal{B}_1^4$ | 0.78 | 0.89 | 0.08 |
| $\mathcal{B}_1^5$ | 0.72 | 0.88 | 0.08 |
| $\mathcal{B}_1^*$ | 1.00 | 0.92 | 0.07 |

(b) Experiment 2

| iteration | $\zeta$ | $\bar{\alpha}$ | $\sigma_\alpha$ |
|---|---|---|---|
| $\mathcal{B}_2^0$ | 0.45 | 0.87 | 0.09 |
| $\mathcal{B}_2^1$ | 0.84 | 0.91 | 0.08 |
| $\mathcal{B}_2^2$ | 0.87 | 0.91 | 0.07 |
| $\mathcal{B}_2^3$ | 0.83 | 0.91 | 0.08 |
| $\mathcal{B}_2^*$ | 1.00 | 0.92 | 0.07 |

in the Glasgow corpus by hand. Figure 6.6 illustrates that defining unique segmentation points in an action sequence by hand is difficult — especially because upper and lower body actions can overlap. In order to make manual segmentation tractable, I decided not to segment upper and lower body actions separately. I therefore only defined one sequence of 7 action boundaries and used them for both the upper and lower body actions. This inevitably leads to some inaccuracies. However, one of the main objectives of this section is to evaluate the relative *improvements* in segmentation quality that are achievable by iteratively updating HMM models through bootstrapping. Therefore, while absolute segmentation quality is difficult to evaluate, relative improvements can be judged more easily in this way.

To evaluate the quality of segmentation which can be achieved through bootstrapping I conducted two series of experiments. Because the upper body actions play the predominant role in the Glasgow corpus, the segmentation results quoted in this section were obtained solely from the upper body action models. The actions under consideration are hence $\mathcal{C}_U = \{$*Knocking*, *Throwing*, *Lifting*, *TUp*, *TDown*, *other*$\}$. The two experiments differ in how $\Lambda^0$ is initialised. In the first experiment I am training motion models using all available isolated actions from all 30 subjects. This forms the initial model $\Lambda_1^0$ for the bootstrapping framework. In Section 4.2.4 we saw that a limited amount of training data can lead to action models which do not generalise as well to samples from new subjects. In order to test how the bootstrapping approach copes with this situation, experiment 2 trains $\Lambda_2^0$ with data from only three of the subjects. A real-world scenario would probably lie somewhere in between these extremes with a moderate amount of initial training data and sequential data which could stem from previously unseen subjects.

Starting with the according $\Lambda^0$ in each case I perform a number of bootstrapping iterations. All results have been obtained with a fixed number of $L = 8$ levels and therefore without the necessity to estimate the optimal number of levels $\hat{l}$. Note, however, that I am making no assumptions about the *order* in which actions occur in the sequence. The algorithm therefore knows that there are eight actions in every sequence, but it does not know which actions they are and where the boundaries lie. Table 6.1 gives a number of segmentation quality metrics computed after each iteration. In particular, I calculate $\zeta$, the fraction of sequences which are segmented into the correct sequence of actions:

$$\zeta = \frac{\text{\# samples segmented into the correct sequence of actions}}{\text{\# samples in the corpus}} \tag{6.11}$$

A more detailed metric looks at the frame-wise agreement between a bootstrapped segmentation and the manual segmentation of the same sequence. The frame-wise agreement $\alpha$ achieved for a sequences is calculated as

$$\alpha = \frac{\text{\# frames in agreement between manual and bootstrapped segmentation}}{\text{\# frames in the sample}} \tag{6.12}$$

Table 6.1 lists mean and standard deviation for $\alpha$ as calculated from all sequences.

We can see that the initial segmentation quality in experiment 2 is slightly worse than in Experiment 1. This is especially apparent when we consider $\zeta$. That is to be expected as $\Lambda_2^0$ is biased towards the few subjects which were used for training. The subsequent bootstrapping iterations improve the segmentation quality. A peak is reached for $\Lambda_1^1$ and $\Lambda_2^2$ in Experiments 1 and 2 respectively. At that point the segmentation quality is virtually the same in both experiments. Therefore bootstrapping is able to correct for a biased initial training after only two iterations.

Experiment 1 also shows that the iterations do not converge to a maximum. They increase at first and then decrease. This is because noisy samples are picked up during the bootstrapping iterations in $\mathcal{B}^i$ which subsequently biases the HMM training. The bootstrapping process therefore does not *converge* towards $\hat{\Lambda}$, the optimal model which gives rise to the optimal segmentation $\hat{\mathcal{B}}$. It is, however, clear that a local maximum is reached within a few iterations of bootstrapping. These local maxima $\mathcal{B}_1^1$ and $\mathcal{B}_2^2$ are the optimal segmentations obtained from experiments 1 and 2 respectively.

A natural question to ask at this point is: How close are these local maxima to the global maximum $\hat{\mathcal{B}}$? In this sense I regard $\mathcal{B}^i$ as the optimal segmentation achievable through LB. We notice that $\zeta$ tends to increase as $\alpha$ increases. In other words, as the number of samples segmented into the correct sequence of actions increases, the total frame-wise agreement also improves. An approximation for the best possible segmentation $\hat{\mathcal{B}}$ might therefore be thought of as the one segments the sequence into the correct order of component actions. This would give rise to $\zeta = 1$. With LB this can be achieved by enforcing a known order of actions as defined through the transition function $\hat{t}(i, j, l)$ (see Equation 6.8). The algorithm's optimisation then reduces to finding the optimal frames for segmentation. Using the models optimised through bootstrapping, I ran this restricted version of Level Building to obtain segmentations $\hat{\mathcal{B}}_1$ and $\hat{\mathcal{B}}_2$ in experiments

1 and 2 respectively. Table 6.1 includes these results. We see that the segmentation quality reaches an average frame-wise error rate of $\bar{\alpha} = 0.08$ at a standard deviation of $\sigma_\alpha = 0.07$. These figures are virtually the same for Experiments 1 and 2, confirming the earlier observation that bootstrapping is relatively stable with respect to different initial models $\Lambda^0$. More importantly, the frame-wise segmentation quality achieved for $\hat{\mathcal{B}}_1$ and $\hat{\mathcal{B}}_2$ is also not much higher than those of the optimal segmentations I found without enforcing a particular sequence of actions. This boosts my confidence that the locally maximal segmentations are not very different from and hence good approximations for the optimal segmentation $\hat{\mathcal{B}}$.

After discussing related work in the next section I will demonstrate how important good action segmentation is for successful emotion recognition.

### 6.2.5 Related work

Segmenting and recognising connected actions faces many problems encountered in speech recognition. For example, the speech community has developed a host of techniques to deal with the problem of coarticulation. One of the most common techniques is to train models on connected units of speech, called n-grams [MS99]. Those allow the models to capture the interaction between neighbouring speech units. It is believed that training models that take these coarticulation effects into account is key to producing good speech recognition results [Rab02]. The complexity resulting from an explosion in the number of models that need to be trained is often reduced by tying the majority of model parameters together [MS99]. I decided to take a different approach and to adapt single action models. The main intuition was that, as opposed to speech units, actions tend to be much longer in duration and only small temporal portions of the motion signal tend to be affected by coarticulation effects. Other authors previously found that coarticulation effects in the gestural domain are better tackled using adaptation rather than training context-dependent models [VM97].

#### Segmentation

The models are used in a Level Building algorithm to segment the samples into individual actions. The idea of Level Building was first proposed by Rabiner and Levinson as a solution to connected speech recognition [MR81c]. At the time they used it in connection with Dynamic Time Warping (DTW), a concept I introduced in Section 4.3.2. The LB algorithm had been designed to support language grammars. I used this ability to enforce a known action sequence when computing the optimal segmentation $\hat{\mathcal{B}}$. Rabiner and Levinson also extended the algorithm to HMMs and introduced the idea of modelling word durations [RL85]. I used their ideas in order to find the optimal number of levels $\hat{l}$.

Level Building is not the only method which has been used in the past to segment continuous signals such as speech into component units. One of the most popular methods today is Token Passing [YRT89]. It was originally proposed by S.J. Young *et al.* as a conceptual abstraction of some more specialised algorithms, including Level Building. Token Passing can be implemented very efficiently and a very popular implementation exists in the HTK Toolkit developed and maintained by the Cambridge University Engineering Department [You93]. For my purposes Level Building proved sufficient and a use of Token Passing would have lead to very similar results. For a real-world implementation, however, a Token Passing algorithm would probably outperform a Level Building approach in terms of computation time.

Action recognition and segmentation has also received attention from the computer vision community. Gilbert *et al.* use corner-based spatio-temporal features and a sliding window to localise actions both in space and time [GIB09, GIB09]. Their approach is less suitable for segmenting complex actions of varying duration as the sliding window is fixed and needs to roughly agree with the action duration. Oikonomopoulos *et al.* employ a similar methodology based on a sliding window. They, however, use sliding windows of various durations and pick the one which results in the strongest correlation between the observed and learnt spatio-temporal features [OPP09]. This allows them to locate the beginning and end frames of actions more accurately.

Other approaches to segmenting connected symbols can be found in the sketch and handwriting recognition communities. Hierarchical HMMs (HHMMs) were proposed by Fine *et al.* [FST98]. They model complex multi-scale structures which often appear in natural time series such as cursive handwriting, language or speech. HHMMs were demonstrated to be especially effective at capturing large distance relationships between hidden states, such as grammatical constraints between words in a sentence [FST98]. Two-level HHMMs were also used for sketch interpretation [SD04]. Akin to a scene grammar, Simhon and Dudek model the temporal and spatial relationships between scene objects such as water, grass and clouds as a high level state graph. At the second level, each scene element is in turn modeled by an HMM which produces visible outputs. This idea can be developed into a more general graphical model. Sezgin *et al.* enhance the HHMMs to also model the interactions between neighbouring objects [Sez06], not unlike the effects of coarticulation.

### Model refinement

In order to optimise model parameters for connected actions from isolated models I used a bootstrapping-based approach. Bootstrapping has been used as an optimisation technique in many diverse domains. Recently, it has been very popular for data mining on large collections of text [AG00, Bri99, LYG03]. The most relevant recent work is by Oats *et al.* who use bootstrapping to detect different statistical sources for a set of time series [OFC01]. They are assuming that the nature and number of sources are unknown, and

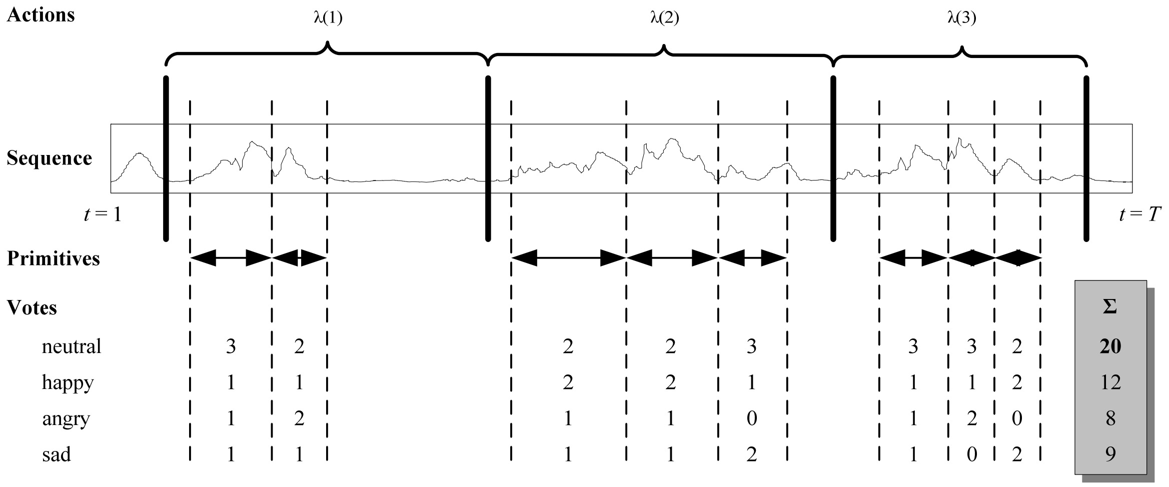

**Figure 6.7:** Emotion recognition process for a sequence of actions. For every primitive a number of votes are accumulated from binary SVMs (see Section 5.5.1).

therefore propose an unsupervised clustering solution. In contrast, I know which actions to expect in the sequences. They noted that in order to achieve good clustering performance it is important to start with good initial models $\Lambda^0$ or structure information $\mathcal{B}^0$. They obtain a good $\mathcal{B}^0$ by applying DTW to the sequences whereas I used the isolated action models to get a good initial estimate for $\Lambda^0$. Although Oats *et al.* used simple synthetic data, they found that the clustering is not perfect and that a number of misclassifications occur. Because they do not reclassify all samples at every iteration their algorithms still converged but they were largely facing the same stability problems which I report.

Bootstrapping algorithms are usually susceptible to these kinds of instability if noisy classifications enter the loop. In the field of information extraction, these problems are sometimes overcome with a semi-supervised approaches where information is regularly checked for spurious samples [Bri99]. Another promising approach comes from Lin *et al.* who propose to use negative as well as positive examples in the model training step during the bootstrap iterations [LYG03]. This creates competition between multiple categories which tends to constrain divergence [TR02]. The divergence problem arises in my case because the Baum-Welch optimisation trains HMMs purely generatively in the sense that parameters are learned such that they best match the observations. In the past authors have however proposed discriminative training methods for HMMs which could take into consideration negative training examples [Col02] as well. This might enable a more stable bootstrapping procedure as proposed by Lin *et al.*

## 6.3 Emotion recognition

Having found segmentations of various quality levels I now turn to the problem of recognising emotions from the segmented sequences. Much of the emotion classification approach will follow the procedure described for isolated actions in Section 5.5. In particular, I will reuse the recognition pipeline as depicted in Figure 5.2 for each individual segmented action. This will allow me to directly compare the results obtained for isolated and sequential actions.

One of the main goals of this section is to establish the extent to which emotion classifiers developed for isolated motions can be applied to sequential actions. In the simplest form, I could apply the classifiers as trained in Section 5.5 to the segmented actions directly. I pointed out previously, however, that the appearance of actions changes from isolated to sequential cases. This strongly suggests that emotion classifiers trained on the appearance of isolated actions will give suboptimal results when applied to sequential actions. I will therefore discuss and evaluate various methods for adapting the emotion classifiers to sequential actions. Those adaptation methods have different levels of complexity and cost associated with them. This kind of adaptation is analogous to adapting the isolated action models to action sequences through bootstrapping which I have already showed to lead to higher quality action segmentations.

As with the isolated cases, emotion classification is based on SVMs. I already described how a simple voting scheme can use binary SVMs to distinguish between multiple emotion classes in cases where exactly one feature vector per action is derived (action-global features). I also showed how the scheme generalises to cases where I compute multiple primitives for a single action (segment-local features). In this section I extend the voting scheme once more and combine votes from multiple actions to arrive at a single emotion label for a whole action sequence. As before, ties are resolved by assigning one of the candidate classes randomly. Figure 6.7 illustrates the entire voting procedure in more detail.

### 6.3.1 Experimental procedure

The primary goal of my evaluation is to establish how well the emotion recognition framework developed for isolated motions performs on sequential motions. As opposed to my evaluation in Chapter 5 my focus will therefore *not* be on the generalisation to new subjects and cross-validation. Instead, I will regard the Glasgow corpus as split into two large partitions: isolated training data and sequential testing data.

As for the isolated cases, my experimental evaluation focuses on a number of different factors which can affect emotion recognition from action sequences. In order to get the most comprehensive picture possible, I decided to adopt a factorial design. By investigating the factors in all possible combinations I get a rich set of classification results which provide the basis for a thorough discussion at the end of this chapter. Below I am giving a detailed description of the factors and according conditions analysed.

All experiments were carried out with custom Matlab implementations of LB and DTW

**Table 6.2:** Segmentations and their associated quality metrics used for the evaluation of my emotion recognition framework for actions sequences.

| iteration | $\zeta$ | $\bar{\alpha}$ | $\sigma_\alpha$ |
|---|---|---|---|
| $\mathcal{B}^{-2}$ | 0.00 | 0.13 | 0.11 |
| $\mathcal{B}^{-1}$ | 1.00 | 0.33 | 0.14 |
| $\mathcal{B}_2^0$ | 0.45 | 0.87 | 0.09 |
| $\mathcal{B}_2^1$ | 0.84 | 0.91 | 0.08 |
| $\mathcal{B}_2^2$ | 0.87 | 0.91 | 0.07 |
| $\mathcal{B}_2^3$ | 0.83 | 0.91 | 0.08 |
| $\hat{\mathcal{B}}_2$ | 1.00 | 0.92 | 0.07 |
| $\mathcal{B}^*$ | 1.00 | 1.00 | 0.00 |

for primitive extraction, as well as Joachim's SVMlight implementation of binary SVMs [Joa99]. Feature subsets were selected using the implementation in the WEKA package [HFH$^+$09, Wek].

### Factor 1: Action categories

In Section 5.5.3 we saw that SVM-based emotion recognition performance can vary between different action categories. As the action sequences all contain examples of *Knocking*, *Lifting*, *Throwing* and *Walking* motions, I can perform a similar post-hoc analysis to see which action categories provide the best grounds for distinguishing emotions in a sequential setup. In particular, I can choose to combine individual SVM votes for a whole sequence, but for each individual vote I will also record which action category it came from. This allows me to draw up results similar to those for isolated actions.

### Factor 2: Quality level of segmentation

One of the major problems I have focused on in this chapter is how to improve the segmentation of connected actions. Good action recognition and segmentation clearly have important applications in themselves. Within the scope of this dissertation, however, the main interest lies in the effect that segmentation quality has on emotion classification. I will use six segmentations produced in Section 6.2.3. Those are $\mathcal{B}_2^0$, $\mathcal{B}_2^1$, $\mathcal{B}_2^2$ and $\mathcal{B}_1^3$ computed through bootstrapping where $\mathcal{B}_2^0$ comes from the *biased* isolated motion models. This allows me to study a dense sample of quality levels $0.87 \leqslant \bar{\alpha} \leqslant 0.91$ achievable through bootstrapping.

At the top end of segmentation quality, I also use $\hat{\mathcal{B}}_2$ as obtained through Level Building when I enforce the correct action order. Finally, the arguably best segmentation is the segmentation obtained by hand, $\mathcal{B}^*$.

In order to compare these results to more degenerate segmentations, I also introduce two segmentations worse than $\mathcal{B}_2^0$. $\mathcal{B}^{-1}$ is a segmentation obtained by splitting each sequence into 8 regions of equal duration and assigning to them the emotion labels in the correct order as they appear in the corpus. This produces a frame-wise segmentation quality of $\bar{\alpha} = 0.33$ (see Table 6.2). Finally, as a worst-case I define $\mathcal{B}^{-2}$ similarly by splitting each sequence into 8 regions and assigning each an action category *at random*. This produces $\bar{\alpha} = 0.13$. Table 6.2 summarises all segmentations used in this experiment.

### Factor 3: Level of adaptation to sequential actions

There are multiple ways in which the SVM-based classification pipeline can be adapted from the isolated models. As a baseline, I can use the classifiers as trained on the isolated actions. In particular, this classifier assumes that the global normalisation constants and the person-specific bias are the same in both the isolated and sequential cases. It furthermore uses the SVMs as trained on the isolated data. As argued before, this model is not very likely to perform well because of the observed differences in appearance. I would expect these models to still perform better than random guessing, however, as appearances in sequences should still not change so extremely as to render the isolated models entirely useless.

As a next step, I could hypothesise that the observed differences are due to a *global* shift which changes the appearance of all actions equally when they appear in sequences. In other words, I could recompute the global normalisation constants (mean and standard deviation of features) based on the sequential data. This would mean replacing the global constant inputs to the recognition pipeline introduced back in Figure 5.2. This condition then assumes that all *personal* deviations are not affected and hence the personal bias and emotion classifiers remain unchanged.

The third hypothesis postulates that in fact personal biases change as well when moving from isolated to sequential motions. In other words, I propose that not only do actions appear differently for isolated and sequential cases, but for different individuals this change happens independently. This means that I need to recompute global constants *and* personal bias from the sequential actions in the recognition pipeline.

The final hypothesis postulates that isolated and sequential motions are entirely different. In order to get good recognition performance I need to fully retrain the emotion classifier SVMs on *sequential* training data. This level therefore re-estimates all the inputs to the recognition pipeline and totally ignores the classifiers trained for isolated actions. In order to estimate the recognition performance for this condition I use 10-fold cross validation (10F-CV) as introduced for the classification of isolated motions. Otherwise I would test on exactly the same samples that I trained the classifiers on which would positively bias the results (see Section 5.5.2).

I am calling the derived classifiers $M^0$, $M^1$, $M^2$ and $M^3$ according to how many inputs to the recognition pipeline have to be recomputed. Note that each recomputation of

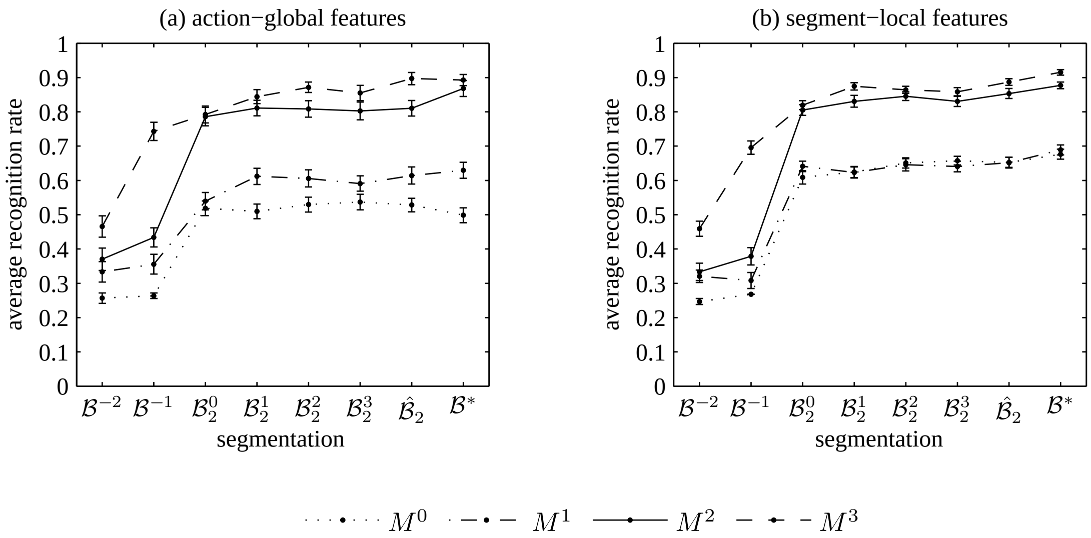

**Figure 6.8:** Emotion recognition rates for various levels of segmentation quality and levels of adaptation. The results were obtained using (a) the action-global feature set; (b) the primitive-local feature set.

pipeline inputs inflicts an increased cost as I require data labeled with increasingly rich information from component actions to individuals and emotions.

### Factor 4: Level of action modelling

In Section 4.3 I discussed the possibility of modelling actions at a finer scale of primitives in order to capture subtleties of emotional expressions which might otherwise be missed. We then saw evidence that using these primitive-local features can indeed improve the recognition performance for isolated actions. I will return to this question in this section and perform recognition using both action-global and primitive-local features. For the latter case, I use the master descriptions as introduced in Section 4.3.2 to first align each of the segmented actions and then propagate the previously defined primitive boundaries to the new action sample.

## 6.3.2 Results

The experiment provided a wealth of data which I will present in various forms in this section. Figure 6.8 provides a broad overview of the obtained recognition results. It plots the combined recognition rates for whole sequences. Each line represents the change in recognition rate as I improve the segmentation quality and keep the level of emotion classifier adaptation constant. Because the classification algorithm resolves tied votes by making a random choice I am also indicating the spread as obtained by multiple runs of the experiment with error bars. A general trend we are already observing at this point

**Table 6.3:** Emotion recognition results by action and for different experimental conditions. Individual confusion matrices list emotions in the order *neutral*, *happy*, *angry*, *sad*.

| | $(\mathcal{B}_2^0, M^0)$ | | | | $(\mathcal{B}_2^2, M^2)$ | | | | $(\mathcal{B}^*, M^3)$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Knocking* | **0.41** | 0.20 | 0.02 | 0.37 | **0.68** | 0.18 | 0.02 | 0.12 | **0.71** | 0.09 | 0.02 | 0.18 |
| | 0.23 | **0.46** | 0.20 | 0.11 | 0.20 | **0.69** | 0.09 | 0.02 | 0.14 | **0.75** | 0.07 | 0.04 |
| | 0.00 | 0.12 | **0.85** | 0.03 | 0.00 | 0.12 | **0.88** | 0.00 | 0.00 | 0.09 | **0.91** | 0.00 |
| | 0.15 | 0.07 | 0.00 | **0.78** | 0.27 | 0.02 | 0.00 | **0.71** | 0.14 | 0.06 | 0.00 | **0.80** |
| | **average rate: 0.62** | | | | **average rate: 0.74** | | | | **average rate: 0.79** | | | |
| *Throwing* | **0.32** | 0.25 | 0.18 | 0.25 | **0.61** | 0.09 | 0.04 | 0.26 | **0.66** | 0.13 | 0.02 | 0.19 |
| | 0.18 | **0.40** | 0.40 | 0.02 | 0.32 | **0.59** | 0.02 | 0.06 | 0.22 | **0.67** | 0.04 | 0.07 |
| | 0.00 | 0.02 | **0.98** | 0.00 | 0.00 | 0.09 | **0.91** | 0.00 | 0.04 | 0.05 | **0.91** | 0.00 |
| | 0.35 | 0.16 | 0.05 | **0.44** | 0.32 | 0.05 | 0.02 | **0.61** | 0.30 | 0.06 | 0.00 | **0.64** |
| | **average rate: 0.54** | | | | **average rate: 0.68** | | | | **average rate: 0.72** | | | |
| *Lifting* | **0.30** | 0.04 | 0.04 | 0.62 | **0.73** | 0.11 | 0.04 | 0.12 | **0.68** | 0.09 | 0.02 | 0.21 |
| | 0.11 | **0.47** | 0.15 | 0.27 | 0.08 | **0.66** | 0.20 | 0.06 | 0.16 | **0.72** | 0.05 | 0.07 |
| | 0.06 | 0.30 | **0.59** | 0.06 | 0.00 | 0.09 | **0.91** | 0.00 | 0.00 | 0.11 | **0.89** | 0.00 |
| | 0.02 | 0.00 | 0.02 | **0.96** | 0.28 | 0.00 | 0.02 | **0.70** | 0.22 | 0.04 | 0.00 | **0.74** |
| | **average rate: 0.58** | | | | **average rate: 0.75** | | | | **average rate: 0.76** | | | |
| *Walking* | **0.02** | 0.09 | 0.00 | 0.89 | **0.84** | 0.09 | 0.05 | 0.02 | **0.91** | 0.04 | 0.00 | 0.05 |
| | 0.13 | **0.35** | 0.00 | 0.52 | 0.29 | **0.51** | 0.20 | 0.00 | 0.23 | **0.56** | 0.21 | 0.00 |
| | 0.11 | 0.61 | **0.00** | 0.28 | 0.02 | 0.38 | **0.60** | 0.00 | 0.11 | 0.22 | **0.67** | 0.00 |
| | 0.02 | 0.04 | 0.00 | **0.94** | 0.36 | 0.00 | 0.00 | **0.64** | 0.19 | 0.00 | 0.00 | **0.81** |
| | **average rate: 0.33** | | | | **average rate: 0.65** | | | | **average rate: 0.74** | | | |
| *Combined* | **0.16** | 0.06 | 0.00 | 0.78 | **0.88** | 0.03 | 0.03 | 0.06 | **0.93** | 0.02 | 0.02 | 0.03 |
| | 0.07 | **0.47** | 0.06 | 0.40 | 0.12 | **0.74** | 0.11 | 0.03 | 0.13 | **0.82** | 0.05 | 0.00 |
| | 0.00 | 0.40 | **0.47** | 0.13 | 0.00 | 0.14 | **0.86** | 0.00 | 0.00 | 0.07 | **0.93** | 0.00 |
| | 0.00 | 0.00 | 0.02 | **0.98** | 0.23 | 0.01 | 0.00 | **0.76** | 0.08 | 0.02 | 0.00 | **0.90** |
| | **average rate: 0.51** | | | | **average rate: 0.81** | | | | **average rate: 0.89** | | | |

is that with increasing adaptation and segmentation quality better recognition rates are achieved. I will now consider the results with respect to the four factors outlined above.

### Factor 1: Action categories

Table 6.3 gives the detailed recognition rates as achieved on the connected actions after segmentation and split by action category. I also show the confusion matrices for combined recognition of a whole sequence. In order to make the results comparable to the isolated cases from Section 5.5.3 I give results for three conditions:

1. Largely unadapted models: segmentation $\mathcal{B}_2^0$ and $M^0$
2. Well-adapted models: segmentation $\mathcal{B}_2^2$ and $M^2$
3. Best possible models: segmentation $\mathcal{B}^*$ and $M^3$

In all cases results are based on action-global features. Across the different conditions none of the actions stands out as particularly easy or hard to analyse. One exception is *Walking* in the first condition. With a recognition rate of 33% the isolated emotion models for *Walking* are clearly unsuitable for analysing the connected cases. In fact, the combined classifier for the whole sequence would be better off not to include the *Walking* classifications as the combined result is considerably lower than the results for the other three action categories. In the other two conditions, however, the combined classification performance is consistently higher than the recognition rates for individual actions.

#### Factor 2: Quality level of segmentation

Figure 6.9 visualises the connection between the quality of segmentation $\bar{\alpha}$ and resulting emotion recognition rate. I am showing the results for both action-global and primitive-local features. The four adaptation levels are plotted as separate graphs. In order to evaluate the significance of the correlation I form the null-hypothesis that there is no correlation between the variable $\bar{\alpha}$ and the recognition rate. For each condition I am then transforming the values using a Student's t-distribution and calculate the probability $p$ that the null-hypothesis is true. The corresponding $p$-values are given in the plots. In all cases there is a significant correlation between the recognition rates and $\bar{\alpha}$. The plots confirm that the best recognition results are achieved with accurate action models which allow me to segment the sequences well. Studying the $p$-values in more detail, we see that the correlations tend to be more significant if using primitive-local features rather than only using features derived from the actions globally.

#### Factor 3: Level of adaptation to sequential actions

The influence of increasing levels of adaptation on the recognition rate can be best seen by looking back at Figure 6.8. Here different levels of adaptation are plotted as separate lines. We see that with increasing levels of adaptation the recognition rates tend to increase. Improvements can be very substantial. For example, the average gain in recognition rate obtained over all segmentation levels by going from $M^0$ to $M^2$ is 26% for action-global features and 18% if we work with primitive-local features. Note that I managed to achieve this improvement without retraining the actual emotion classifiers. The adaptation stemmed solely from appropriate preprocessing of the feature vectors. This means that the models were only adapted to the changed appearance of actions, disregarding any changes that might occur for individual emotions. If I am also willing to retrain the emotion classifiers on connected actions labeled with emotions, I can achieve improvements of 33% and 26% for action-global and primitive-local feature sets respectively.

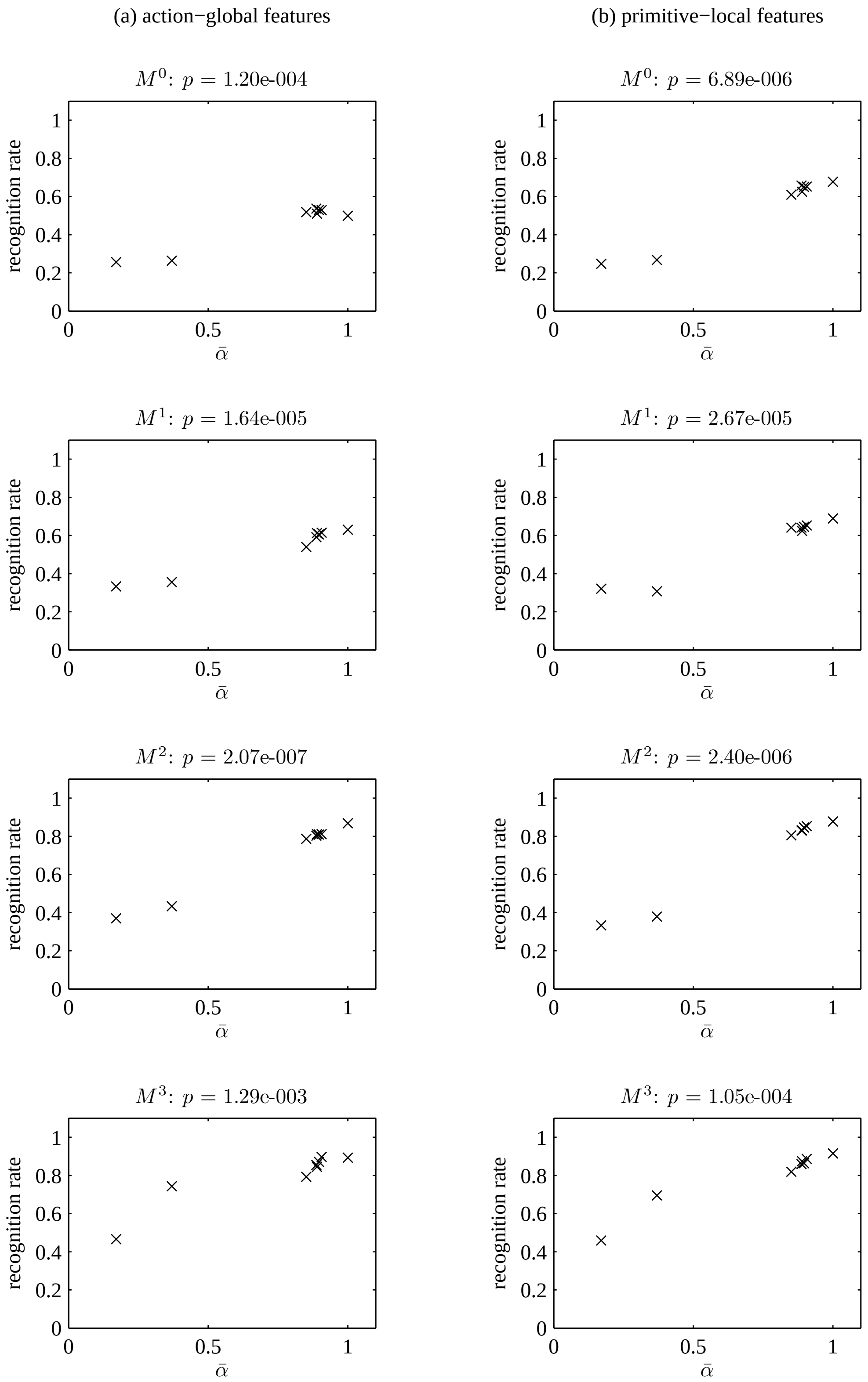

**Figure 6.9:** Emotion recognition rates plotted against $\bar{\alpha}$. The results were obtained using (a) the action-global feature set; (b) the primitive-local feature set.

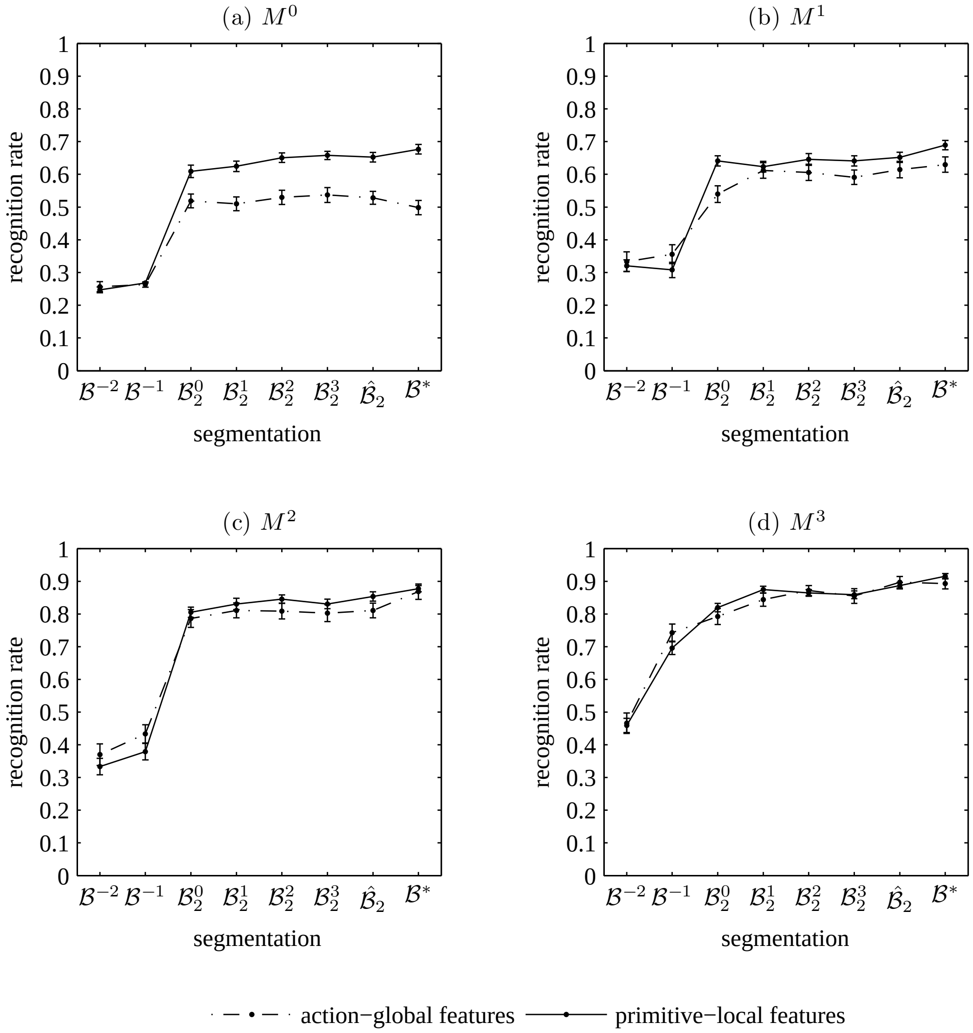

**Figure 6.10:** Comparison of emotion recognition rates using action-global and primitive-local features.

### Factor 4: Level of action modelling

Figure 6.10 shows a direct comparison of the recognition rates achieved for action-global and segment-local feature sets. In Section 5.5.3 we saw that there is a small but significant advantage in using segment-local features when analysing isolated motions. Figure 6.10 suggests that there is a notable difference between the two conditions only if the action models are not adapted at all. In this case segment-local features afford an improvement in recognition rate of 13%. In the other conditions, the segment-local features tend to perform slightly better than the action-global features. The improvements, however, are comparable to the margin of error introduced by the voting scheme.

## 6.4 Discussion

The experiments give many insights into the problem at hand. In this section I am going to highlight and discuss some of the more significant results. I will also relate the results from this chapter back to the recognition results achieved on isolated actions (Chapter 5) where appropriate.

The first thing to note is that at the low end of the scale the recognition rates can approach chance level of 25%. In my experiment this is the case if I use the classifiers without adaptation and use bad or random segmentations ($\mathcal{B}^{-1}$ and $\mathcal{B}^{-2}$). At the top end of the scale recognition rate of 92% using 10F-CV are achievable if sequences are segmented manually ($\mathcal{B}^{*}$) and the emotion classifiers are trained on connected data.

One interesting condition to pick out at the lower end of the scale is what is achieved with $\mathcal{B}^{-2}$ and $M^3$. For action-global features this condition yields an average recognition rate of 47%. In fact, this number gives a good idea of how well emotion recognition would perform if I assumed no knowledge about the influence of action patterns on the data. By splitting the sequences into a number of sections and assigning an action category at random, I am approximating a sliding window approach which ignores any knowledge about actions. Looking back at Chapter 3, this was the approach that seemed to work well for continuous expressive motions and for everyday activities it performs better than chance. However, by adding more knowledge about the structure of actions I can nearly double this rate. It is still important to take away from this experiment, however, that there seem to be fundamental patterns to the expression of emotions in action-driven body movements. These are strong enough to enable the detection of emotions significantly above chance level without any knowledge about the underlying action patterns.

At the other extreme, we see that building an understanding of action patterns down to a primitive level brings only marginal benefits to emotion recognition from connected actions. A real benefit for motion primitives arises, however, when there is no data to adapt the emotion classifiers ($M^0$). In that condition the recognition rates are considerably higher for primitive-local features. A further benefit of using primitive-local features is that the resulting recognition results show less variation. In fact, for each of the conditions the error bars indicating the variation across multiple iterations are smaller if we primitive-local features are used. This is not surprising as more primitives result in more votes to combine into a final classification. With more votes it is less likely to encounter a tie between two or more emotion classes. Another tendency is that for primitive-local features the correlation between segmentation quality and recognition rate is stronger as exhibited by smaller $p$-values in Figure 6.9. Therefore, although the average recognition rates for both cases are very similar, primitive-local features tend to produce more consistent results with less associated uncertainty. Of course, this marginal benefit comes at a relatively large cost of having to extract the primitives first.

We can furthermore compare the results obtained in this section with the results from the isolated cases (Section 5.5.3). In particular, the last column in Table 6.3 and the first column in Table 5.3 give results for exactly the same experiment carried out on isolated and hand-segmented connected actions respectively. In Section 5.5.3 we found that emotional patterns were particularly pronounced and hence easier to discriminate for isolated *Knocking* and *Walking* actions. When looking at the connected cases for these actions, however, we find that their recognition rates decrease by around 5%. This does not mean, however, that emotional patterns are generally less pronounced in connected actions. Some action categories actually become easier to analyse. *Throwing* and *Lifting* actions both show slight improvements in recognition rates as we move from isolated to connected cases. Overall the performance across different actions tends to be more uniform for connected actions.

Of course, what makes emotion recognition harder for connected actions is the fact that we need a good segmentation and well-adapted emotion classifiers. In particular, this is necessary to achieve the kinds of recognition rates considered for the comparison with isolated actions above. There I assumed that I have perfectly segmented action sequences labeled with emotions to train the emotion classifiers. In real-world situations this kind of data may be hard to come by. Therefore large parts of this chapter have focused on ways to improve action models and emotion classifiers obtained from training on isolated samples. The experiments demonstrated that action models derived through bootstrapping on weakly labeled sequences are able to improve segmentation quality which in turn leads to better emotion recognition.

Starting from action and emotion models trained on isolated data the arguably larger benefit, however, comes from adapting the emotion classifier to connected data. In particular, by far the single most significant improvement was achieved by recomputing the individual motion bias in $M^2$. While recomputing the global normalisation constants ($M^1$) can improve the results somewhat for global features, I really need to reconsider the difference between individuals in order to see a real benefit. This suggests that the change in appearance that we observe between isolated and connected actions cannot be explained through a global adjustment alone. Instead the results suggest that individuals' motion patterns change largely independently of each other. Once I have recomputed the personal bias, retraining the classifiers ($M^3$) provides a noticeable but much smaller improvement.

In conclusion, it is interesting to remember that, unsurprisingly, the best results are achieved by fully retraining the emotion classifiers using connected actions with the corresponding emotion labels. However, it is possible to achieve a large part of the improvement without the need for explicit emotion labels and by focussing on the individual differences of expression between subjects. Together with the previous results in Section 5.5.3 this confirms once more the importance of paying attention to differences between individuals for emotion recognition from human body motions.

# Chapter 7

# Conclusions

Within the scope of the research described in this dissertation I implemented and evaluated a number of approaches to analyse body movements and extract from them emotional cues. In this chapter I will summarise the principle contributions I made. I will also suggest how the results outlined in previous chapters can be applied and developed in future work.

## 7.1 Contributions

There are a number of core contributions arising from this work. Most importantly, this work has been the first thorough analysis of everyday actions for automatic emotion recognition. I have developed a framework for analysing complex actions and subsequently extract and classify emotion-communicating features. Secondly, I have focused on and discussed in detail the role of personal differences in the expression of emotions through the body. Equally importantly, I have developed methods for analysing naturally occurring connected action sequences and analysed the difference to more commonly studied isolated data. Finally, my framework makes use of algorithms which are in principle able to classify action sequences into emotions in real time, thus making it suitable for real-time applications.

I am going to discuss each of these contributions in more detail below.

### 7.1.1 Emotion recognition from everyday body motions

In this dissertation I have described the development and evaluation of an end-to-end system for the analysis and emotion classification of everyday actions. Despite the increased interest from psychologists there had not been any previous treatment of this problem by the computing community. Within the scope of affective computing many researchers had turned to the less ecologically valid archetypal emotion displays.

By creating the Cambridge corpus of expressive motions, I could show that features capturing the posture and motion dynamics of body joints are effective discriminators for emotion recognition in both archetypal and more realistic non-archetypal scenarios. In fact, many of the features I used were inspired by those used for the analysis of archetypal emotion portrayals.

In contrast to many previous attempts to recognise emotions from expressive body motions, I advocated a more holistic approach to emotion recognition from the human body. I described three factors which influence the appearance of human body motions in a systematic way: action, personal motion bias and emotion. The effect of each of these factors on the appearance of body movements had been studied in isolation in the past. My work represents an attempt to reconcile their effects. I showed that there are complex interactions between these factors and subsequently demonstrated the benefit of modelling each of them explicitly in an emotion recognition pipeline. Based on the data I concluded that a good understanding of each of the factors leads to significantly better emotion recognition. For example, we saw that having a model for different action categories is what makes emotion recognition from everyday motions possible. An even more detailed model for actions at the level of primitives lead to more robust and for isolated actions significantly better emotion recognition results.

I evaluated and validated my framework using the extensive Glasgow corpus of everyday actions. This work has been the first to report extensive findings from the computational analysis of this database. A subject count of 30 meant that the data showed significant amounts of variation representative of a large population. I conducted comprehensive experiments which allowed me to evaluate how well my algorithms would generalise to motions from unseen subjects. I found that the performance of my framework approaches the recognition rates of humans when confronted with videos of emotional knocking motions from the database.

### 7.1.2 Significance of personal differences

Guided by a holistic motion analysis framework I investigated the influence of individual motion idiosyncrasies on the appearance of emotional body movements. Building on ideas such as personality detection from gait patterns I demonstrated that in fact individual motion bias is present in all of the four action categories recorded for the Glasgow corpus. I described how to model this bias using a motion signature computed for each person and action. Removing this motion bias before attempting to learn and classify motion patterns into emotion classes makes a very significant difference — for all action categories and for both isolated and connected actions.

From a practical standpoint, I described how this concept can be developed into a scheme akin to speaker adaptation in connected speech recognition. I showed that the personal motion signatures for different action categories are in fact strongly correlated. I described a regression-based adaptation scheme which exploits this correlation to predict the motion bias for various actions from a single type of calibration motion. This scheme achieves results significantly better than when not using any adaptation at all.

### 7.1.3 Sequential *vs.* isolated data

Many emotion recognition results from the body and other modalities have in the past been reported on data samples recorded in isolation. In this dissertation I have shown, however, that the appearance of actions changes as we move from isolated to connected action sequences. This has been the first study to investigate these differences in appearance between isolated and connected actions at a large scale. I found that the differences in appearance are large enough to cause a significant degradation in the system's ability to discriminate between different emotions. These findings suggest that data collected under very constrained laboratory conditions is not necessarily representative of data occurring in more natural scenarios. I showed that systems trained on isolated data are therefore unlikely to perform well on real-world connected action data without further adaptation.

My experiments also revealed, however, that the observed changes in appearance do not mean that emotions in sequential actions are necessarily harder to detect — they only look different. Retraining the emotion classifiers on connected data resulted in recognition results comparable to isolated cases. I also demonstrated the benefit of combining classification results from multiple actions in a sequence. Because training emotion classifiers on labelled sequential data may be hard in practice, I also presented a number of adaptation strategies which do not rely on fully retraining the emotion classifiers. I found that most of the changes in appearance we see between isolated and connected actions are in fact independent of emotions and only affect the statistics of different actions and individuals. Those changes can hence be accounted for by statistical analysis across actions and individuals and without the need to consider emotions. This lead to the important result that emotion classification can be adapted very well from isolated to connected samples without a need for emotion-labelled sequence data.

As action sequences are one of the most common form of body motions in everyday scenarios my results bring the community one step closer to achieving emotion recognition in real-world scenarios.

### 7.1.4 Real-time applications

The emotion recognition pipeline as pictured in Figure 5.2 relies on five major processing steps:

1. Action recognition (HMM-based Level Building)

2. Primitive extraction (Dynamic Time Warping)
3. Feature extraction (vector arithmetic, differences, sums and square roots)
4. Global and bias normalisation (subtraction and division)
5. Emotion classification (SVMs)

The most computation-intensive steps are steps 1, 2 and 5. All of these steps rely on well-established and understood algorithms whose performance has been optimised by a generation of scientists. In particular, Level Building and Dynamic Time Warping were used successfully in the past for the real-time analysis of time series including speech [GR88, IN90, Vuc01], gesture and action recognition [CKKR08, DP93, ZRXL00]. In a real-time scenario the LB and DTW algorithms as well as feature extraction need to analyse the data at a rate of up to 100Hz whereas the final two steps can work on features at a much lower data rate. Steps 1 and 2 would be the most likely bottlenecks of the system. However, having been applied successfully in numerous real-time systems makes them an ideal candidate for handling the high volumes of body motion data.

It would seem that emotion classification might be a complex and expensive operation as I am trying to classify high-dimensional feature vectors using decision boundaries of potentially high complexity. However, SVMs are extremely efficient in solving these kinds of pattern classification problems. This is because complex boundaries are defined through the kernel functions $K(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$ and classifications depend on the data only through the computation of the kernel. In many cases this computation is very fast, including the polynomial and RBF kernels used in this dissertation. Furthermore, the kernel function only needs to be evaluated for the sample to be classified together with the *support vectors*. A typical SVM can have as few as 30 support vectors (binary *Knocking* SVM distinguishing *angry* and *sad*). This sparseness in representation leads to very few evaluations and fast classification.

For a moderate number of emotion classes I would only require a small number of SVM classifications per second. SVM-based emotion classification is therefore well-suited for the real-time classification of my emotion feature vectors. SVMs have been used for real-time application with good success in the past including for problems such as phoneme classification [KKK02], facial expression recognition [MK03, KP05] and the detection of cognitive distraction based on eye movements and other performance data [LRL07]. I also presented a prototype for a system based on the real-time evaluation of the emotion recognition pipeline described in this dissertation at the 26th CHI Conference in Florence [BR08].

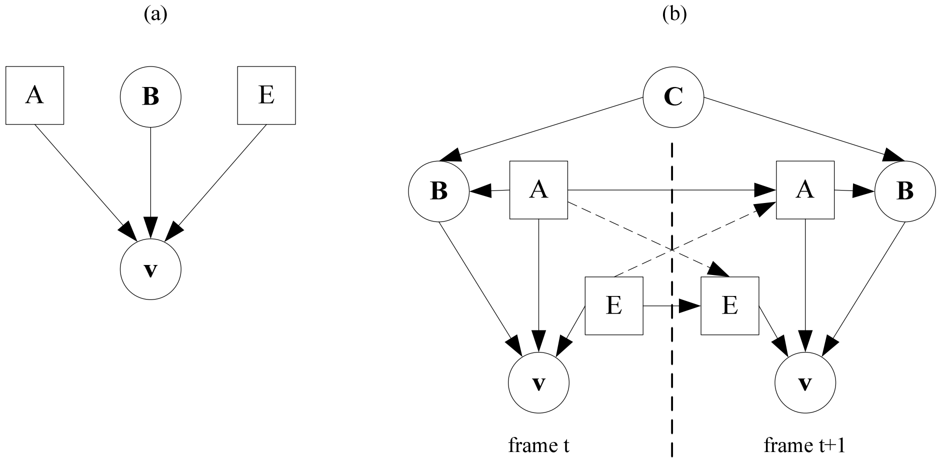

**Figure 7.1:** Graphical model visualising the statistical dependencies as arrows between action type (A), motion bias (**B**), emotion (E) and the observation vector (**v**); (a) models observations as feature vectors in a static Bayesian Network, (b) introduces a calibration variable (**C**) and models observations arising from a Dynamic Bayesian Network.

## 7.2 Future Work

### 7.2.1 Alternative models

This work has shown that the three factors of personal motion bias, action and emotion have a significant impact on an observed motion signal. Although these factors are notionally independent of each other, I decided to design my pipeline with the ultimate goal of detecting *emotions* and trained discriminative classifiers to distinguish patterns arising from different emotion classes. One compelling alternative is to model the problem generatively, as a *Bayesian Network* [Jen01]. In such a graphical model the independence between the three factors would be made explicit (see Figure 7.1(a)). The three factors are modeled as three independent random variables $A$ (action), $B$ (bias) and $E$ (emotion). This Bayesian view would emphasise the *holistic* nature of my approach as emotion is simply one of a number of random variables affecting the state of the observation variable $\boldsymbol{v}$. Among other things, this would enable the recognition of emotions given knowledge about the person ($B$ and $\boldsymbol{v}$ are known) or, conversely, to the identification of individuals if we know the emotion ($E$ and $\boldsymbol{v}$ are known). Standard training algorithms developed for these graph representations would enable the estimation of the underlying probability distributions according to observed data, in much the same way I trained the action models and emotion SVMs.

One of the powers of graphical models is that they can be extended easily to incorporate additional dependencies. For example, user adaptation as presented in Section 5.3 could be achieved by adding a new variable $C$ representing a person's calibration statistics with the appropriate dependencies. Figure 7.1(b) shows such a model based on all the dependencies modelled in my pipeline. Figure 7.1(b) is also a *Dynamic Bayesian Network* which allows action and emotion variables to change over time. A dynamic action variable would allow the construction of state-based dynamical action models, just as I did with the HMM action models described in Section 4.2. A dynamic emotion variable would allow a more sophisticated model of emotions developing over time. This relaxes the current implicit assumption that there is a single emotion giving rise to an observation sequence. In fact, the dotted lines in the figure indicate that I would be able to model even more complex relationships between actions and emotions over time. Eventually, this could enable knowledge such as "Going for a walk usually makes people happier" to be modeled explicitly. This power comes from a holistic view of peoples' movements and may be particularly well represented using a Bayesian Network. Pantic *et al.* also suggested that an affect recognition framework built on Graphical Models would enable a more systematic modelling of complicating factors such as personality, context and cultural biases [PSCH05].

Even if the pipeline as pictured in Figure 5.6 is retained, a number of different modelling techniques could be explored in the future. Particularly interesting candidates are briefly discussed below.

### User adaptation based on calibration motions

User adaptation as formulated in Section 5.3.2 is very similar to, and has indeed been inspired by, the problem of speaker adaptation in speech recognition. The speech community has developed a wealth of techniques to adapt speaker models ranging from maximum likelihood linear regression [LW95] to speaker space methods such as eigenvoices [KJNN00]. My regression-based scheme is relatively simple considering that this problem has sparked several PhD theses [Leg95, Ngu02]. For an emotion recognition system based on my approach to be usable in practice user adaptation is clearly crucial. In Section 5.5.3 we saw that calculating the true motion bias still results in a significantly better recognition performance than using the regression-based adaptation scheme. Improving user adaptation for human actions could therefore be a very fruitful goal for the future. The related research in speech recognition means that there is no shortage of inspiration.

### Parsing action sequences

In Section 6.2.3 I presented a bootstrapping-based technique for retraining isolated action models on connected samples. We saw that the retrained models afforded a better emotion recognition performance. The best performance was still achieved, however, for hand-segmented sequences. Revisiting the problem of adapting the action models to the appearance of connected actions could lead to even better sequence parses. As we have seen, this is likely to give rise to better emotion recognition.

Thinking about a real-world implementation of the technology presented in this dissertation, it would also be interesting to adapt HTK's token passing algorithm to the problem of connected action recognition. I would be able to make use of a robustly developed and maintained toolkit used and tested by a large number of the speech recognition researchers.

### 7.2.2 Naturalistic data

Based on my experimental results I concluded that data collected under very constrained laboratory conditions is not necessarily representative of data occurring in more natural scenarios. Of course, while I argued that the connected action data used for my experiments is more ecologically valid than data used in many others studies to date, it was only recorded under laboratory conditions as well. Returning to Picard's criteria introduced back in Section 2.4.1, the recording procedure could be made more ecologically valid in several ways, including

- acquiring *spontaneous* rather than *posed* data: Although the scenario approach helps to elicit an emotion naturally, the subjects still explicitly act the emotion they are asked to portray.
- the use of *hidden* recording equipment: Optical motion capture technology currently requires subjects to wear light tracking devices and possibly special clothing. This may have an effect on the expressiveness and appearance of actions. Marker-less motion capture is promising to enable hidden recordings of body motions in the near future (see below).
- presenting the recording motivation as *other-purpose*: In order to produce natural displays, subjects would not know that the focus of the experiment is their emotional body expressions.

Many researchers have in the past examined the differences between posed and spontaneous emotional expressions in the face. There is strong evidence that posed expressions have a different appearance from spontaneous ones [SML86, SKA$^+$06, VPAC06]. A number of studies report similar results for vocal expressions [ZLH$^+$79, SKB06]. Although no reports exist to date, it is very likely that similar results will arise for bodily expressions of emotions. It is furthermore likely that subjects' expressiveness will reduce if the recording setup is hidden and other-purpose. A significant reduction in the expressiveness of subjects would not just alter the patterns of emotional expressions, but possibly make them less distinct and therefore more challenging to distinguish using statistical pattern recognition techniques. All of these are interesting challenges for future work.

Many emotion researchers ultimately aim for detecting emotions in fully naturalistic settings in which the subject is unaware of the machine's analysis. An alternative view which is beginning to be discussed by the community is that of *intentional affect* [AMR09]. The idea is to make the user explicitly aware that a machine is analysing his or her body cues and hence create a kind of dialogue between the user's emotional behaviour and the machine's appropriate responses. This kind of situation is much more similar to person-person communication. The kinds of emotional signals that get sent during these kinds of interactions may be more intentional, very different and potentially easier to analyse by statistical algorithms. The concept of *Levels of Intentionality* may be useful in this context [Lew90]: Is the subject or user trying to actively communicate an emotion or not? While the more traditional view might argue for detecting first level intentionality ("I feel sad") with reflex-like motor actions, this view would make use of higher level intentionality ("I want you to believe that I feel sad") which is necessarily interactive in nature.

### 7.2.3 Marker-less data acquisition

One of the reasons why the analysis of body motions has seen a surge in the recent past is the increasing availability of recording equipment which allows accurate and straightforward tracking of 3D body joints. While this technology makes data acquisition at the scale of the Glasgow corpus feasible, the recording procedures are still relatively disruptive for subjects. In most cases subjects need to wear special markers or light emitters. Sometimes additional, heavier modules need to be worn which control active emitters. In other cases subjects need to wear special clothing to facilitate the placement of sensors such as accelerometers and gyroscopes. An additional constraint is that many setups only allow tracking bodies in relatively small spaces. All of these constraints mean that ecologically valid recordings of body motions are very hard to come by. At the same time, there is much active research trying to relax these constraints. Incorporating these new tracking technologies will be an interesting challenge.

In particular, there has been an increasing interest around the challenge of marker-less body tracking. The first commercial systems which facilitate the tracking of 3D body data solely based on a number of standard DV cameras are now becoming available. Recent progress in computer vision research suggests that real-time body pose estimation is becoming a real possibility [MOB06, BSB$^+$07]. Of course, marker-based systems will always be likely to give more accurate tracking results. However, early results on the recognition of emotions from relatively impoverished 2D body silhouettes have been encouraging [CVC07, CBCV07]. It will therefore be interesting to see in the future to what extent emotion recognition from more natural body motions will suffer from the inaccuracies introduced by marker-less tracking technologies.

### 7.2.4 Multimodal analysis

Human behaviour is extremely complex and body motions are only one way in which people communicate emotions and interpersonal stance. Indeed, for a long time the body was considered an inferior source of information for human emotions. This lead to early work on detecting emotions from facial movements, speech and physiological signals. Many approaches to analyse these channels independently now exist. Fusing the different sources of information is one of the most interesting challenges the field of affective computing faces [PR03, SCGH05]. Multimodal systems will need to deal with noisy, inconsistent and missing information from various channels. At the same time, in many situations cues from the body will be the easiest, most reliable and least intrusive to capture. This means that accurate emotion inference from human body motion is likely to play an important role in real-world emotion-aware systems in the future.

# Appendix A

# Rating results for emotional music pieces

In order to confirm that the 9 music pieces selected for the recording of expressive body motions convey the intended emotions I carried out two experiments. Each experiment was completed by with five independent subjects.

### Experiment 1

In the first experiment, I played each music piece once in randomised order and asked subjects to pick the one emotion which they most strongly associated with the piece. The choice was forced out of the following five emotions: *neutral*, *happy*, *sad*, *afraid*, *worried*. Some subjects were unsure about the difference between *afraid* and *worried*. I illustrated the difference with eliciting situations. A person is more likely to feel *afraid* of a physical threat, the response is likely to be relatively short-lived but more intense. A *worried* response is often elicited by a more abstract problem such as a troubling situation of a family member. The emotion can be longer-lasting.

Table A.1 shows the results for this experiment. For most pieces the judges agree with the label. The most problematic piece from this experiment is *(*Afraid 1). In this experiment it was predominantly classified as *worried*. According to the judges it conveys a predominantly worried emotion. It would therefore probably have been a better choice for the class *worried*. It is important to note, however, that the two emotions which contributed most confusion in this experiment were *afraid* and *worried*. These two emotions are very similar to each other. Confusions are hence expected for two reasons:

1. Subjects' personal interpretations and associations are more likely to affect the agreement for emotions which are very similar.

**Table A.1:** 5-way forced choice rating results of the 9 emotional music pieces. Ratings were done by 5 subjects. The second column lists the rates at which subjects agreed with the pieces' labels. The remaining columns give detailed rating results. Disagreements are highlighted in bold.

| piece | agreement | neutral | happy | sad | afraid | worried |
|---|---|---|---|---|---|---|
| Afraid 1 | 0.20 | 0.00 | 0.00 | **0.20** | 0.20 | **0.60** |
| Happy 1 | 0.80 | 0.00 | 0.80 | **0.20** | 0.00 | 0.00 |
| Sad 2 | 0.80 | 0.00 | 0.00 | 0.80 | 0.00 | **0.20** |
| Afraid 2 | 1.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 |
| Worried 2 | 0.60 | **0.20** | 0.00 | 0.00 | **0.20** | 0.60 |
| Neutral | 1.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Happy 2 | 1.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 |
| Worried 1 | 0.80 | 0.00 | 0.00 | **0.20** | 0.00 | 0.80 |
| Sad 1 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 |

2. Music pieces are very likely to evoke mixtures of emotions. If the emotions in question are very similar, music pieces which evoke one are likely to carry elements which also evoke the other, similar emotion.

With this in mind the rates of agreement for *Afraid 2* and *Worried 1* are very good indeed.

### Experiment 2

During my data collection the music was always used in connection with a suggested emotion label. Experiment 1 is therefore arguably not as representative for the use of the music. In a second experiment I chose a set up which is much closer to the actual use of the music.

In this experiment I emulated the use of explicit emotion labels. The music is used to elicit emotional body motions and it is therefore most important that the suggested emotion label is in agreement with the music. In particular, music pieces might express a mix of emotions such as *afraid* and *worried* at the same time. For every piece I therefore asked subjects explicitly whether they thought it evoked the intended emotion. For example, for piece *Afraid 1* I asked I gave them a yes-no choice of whether they thought the piece evoked *afraid*. In order to avoid a bias towards marking *yes* for every sample, I also included 9 random entries, for example asking whether they thought that *Sad 1* evoked *afraid*.

The results are presented in Table A.2. As expected, the agreement is a lot higher than for Experiment 1. The only disagreements occur for *Worried 2* and *Afraid 1*. Both disagree-

**Table A.2:** 2-way forced choice rating results of the 9 emotional music pieces. 5 subjects were asked whether they agreed with the labels given to each music piece.

| piece | agreement |
|---|---|
| Happy 1 | 1.00 |
| Sad 1 | 1.00 |
| Happy 2 | 1.00 |
| Worried 2 | 0.80 |
| Worried 1 | 1.00 |
| Afraid 1 | 0.80 |
| Neutral | 1.00 |
| Sad 2 | 1.00 |
| Afraid 2 | 1.00 |

ments come from the same subject and the same subject exhibited a lot of disagreements in Experiment 1 (5/9). I did not follow up on the reasons for this, but it might be justified to treat this subject as an outlier.

In conclusion, all music pieces showed a high degree of agreement in Experiment 2, which is the most representative for the use of the music. As the only exception, it might be worth reconsidering the label or use of *Afraid 1* in future work as it drew by far the most disagreement among the raters.

# Appendix B

# Subject instructions

The following instructions were given to all subjects for the recording of the Cambridge corpus of expressive body movements.

> This experiment is aimed at recording emotional body motions. During the session you are going to wear a motion capture jacket and a hat, so we can track your body as you are moving in the capture area. Before the capture starts, please ensure that you feel comfortable wearing the special clothing. Once you have been calibrated to the system, please do not move the hat or markers on the jacket as this will severely impair the data recorded. While moving during the capture session, please try to stay inside the capture area outlined on the floor.
>
> ## The session protocol
>
> The goal of the session is to invoke as strongly emotional body motions from you as possible. In order to help you feel comfortable expressing emotions during the session, we are providing the following environment:
>
> - Clear labels will be given on the screen as to which emotion you should express at any one time.
> - You will hear music with the corresponding emotional qualities. This will assist you to put yourself into the particular mood.
> - The room will not be accessed by other people (including the experimenter) during the session.
>
> Please try to exhibit body motions throughout the whole session, as your motions will be recorded throughout, not just when the emotional stimuli change. You are free to express the emotion in any way you want. It might

help if you try and put yourself into an emotional situation in accordance to the music (for example a situation in your life in which you felt very sad or worried). Try and be as expressive as possible.

The recording session will last approximately 20 minutes. The sequence of emotions will be as follows:

Neutral (30s)
Happiness (30s)
Worry (30s)
Sadness (30s)
Fear (30s)

Neutral (50s)
Worry (50s)
Happiness (50s)
Fear (50s)
Sadness (50s)

Neutral (40s)
Fear (40s)
Happiness (40s)
Worry (40s)
Sadness (40s)

Neutral (40s)
Sadness (40s)
Worry (40s)
Fear (40s)
Happiness (40s)

Neutral (30s)

You will be asked to fill in a short questionnaire at the end of the session.

Thank you for participating.

# Appendix C

# Evaluation of clustering-based primitives

In addition to the evaluation carried out in Section 4.3.1, I carried out the following statistical evaluation based on the parses of all 1200 *Knocking* motions. I recorded all transitions between primitives, the durations of primitives and the number of occurrences of each primitive type per action sample. In order to capture transitions to and from the start and end states, I added Primitive 0 (start) and 5 (end).

Given this set of six primitives, and their assumed semantic meanings I am considering the following 8 transitions as potentially correct:

0-1, 1-2, 1-3, 2-3, 2-4, 3-2, 3-4, 4-5.

All other transitions are unexpected and problematic. Table C.1 shows the transition counts between primitives. Problematic transitions are highlighted in bold. The fraction of problematic transitions is $611/8700 = 0.07$ . The biggest problem seems to come from the fact that Primitive 3 gets inserted just before the end (large transition count 3-5), most probably after Primitive 4 (large transition count 4-3). These kinds of errors can be avoided with the alignment-based method. This problem is also an indication for the existence of another kind of primitive which sometimes exists at the end of a *Knocking* action. This smaller-scale movement could be distinguished from the current Primitive 3 with the cluster-based method if primitives were defined in terms of absolute as well as relative joint trajectories.

Figures C.1 (left) visualises the distributions of primitive duration relative to the duration of the whole action. In this figure I am also including the durations of periods which were not assigned to any primitive. We can see that the vast majority of these unassigned periods are very short. This indicates that for *Knocking* motions the threshold-based segmentation method is relatively effective. Nevertheless, the long tail of the distribution is one motivation for a more top-down approach such as my alignment-based method.

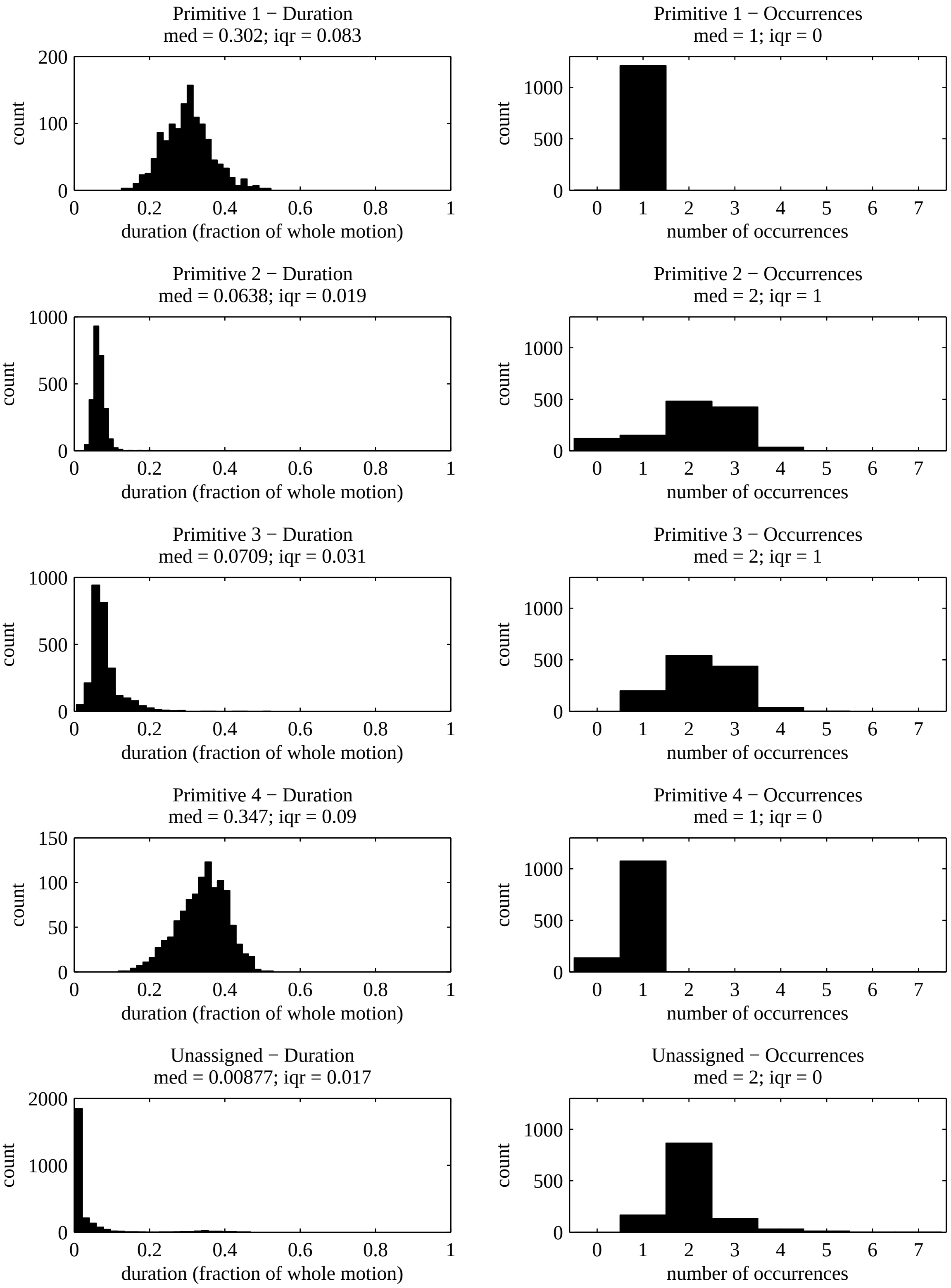

**Figure C.1:** Evaluation statistics for clustering-based primitives: Left: distributions of primitive duration relative to action duration; Right: distributions of the number of occurrences of each primitive type per action sample. The figure includes unassigned regions as a special kind of primitive present in the parses. Each histogram also lists the distribution's average (median) and spread (interquartile range).

**Table C.1:** Primitive transition statistics, including two artificial primitives 0 (start) and 5 (end). Problematic transitions with counts other than zero are highlighted in bold.

| | | to Primitive | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 |
| from Primitive | 0 | 0 | 1120 | **5** | **87** | 0 | 0 |
| | 1 | 0 | 0 | 507 | 701 | **2** | 0 |
| | 2 | 0 | **4** | 0 | 1792 | 618 | **99** |
| | 3 | 0 | **86** | 1968 | 0 | 455 | **185** |
| | 4 | 0 | 0 | **33** | **114** | 0 | 928 |
| | 5 | 0 | 0 | 0 | 0 | 0 | 0 |

Figure C.1 (right) shows the distributions of the number of occurrences of each primitive type per action sample. Again, I am including unassigned periods in the figure. As expected, Primitives 1 and 4 usually occur once, while Primitives 3 and 4 occur multiple times – on average twice but often also three times. The spread in Primitives 2 and 3 can be partly attributed to the fact that they are difficult to isolate using energy minima only. There is an average of two unassigned regions per sample. This is not unexpected as the way my extraction algorithm works tends to leave the first and last frames of each sequence unassigned (see Figure 4.5).

# Appendix D

# Evaluation of alignment-based primitives

The top-down approach followed for the alignment-based definition of motion primitives means that the transitions between primitives necessarily follow the expected order. One way in which I can statistically analyse whether the primitive extraction is robust is by considering the durations of the extracted segments. Because each primitive should represent one well-defined part of the action, each primitive's duration distribution should have a very clear average with a reasonably low spread. Furthermore, a large number of outliers would indicate that the extraction algorithm is not very robust.

Figure D.1 visualises the distributions of primitive duration relative to the duration of the whole action. All distributions are well-defined with small spread. Some primitives such as *Knocking*- Primitive 2 and *Throwing*- Primitive 2 have a relatively long tail which could indicate some problems. However, because the number of affected samples is very small and this dissertation is primarily concerned with the detection of emotions, a further analysis is left as future work.

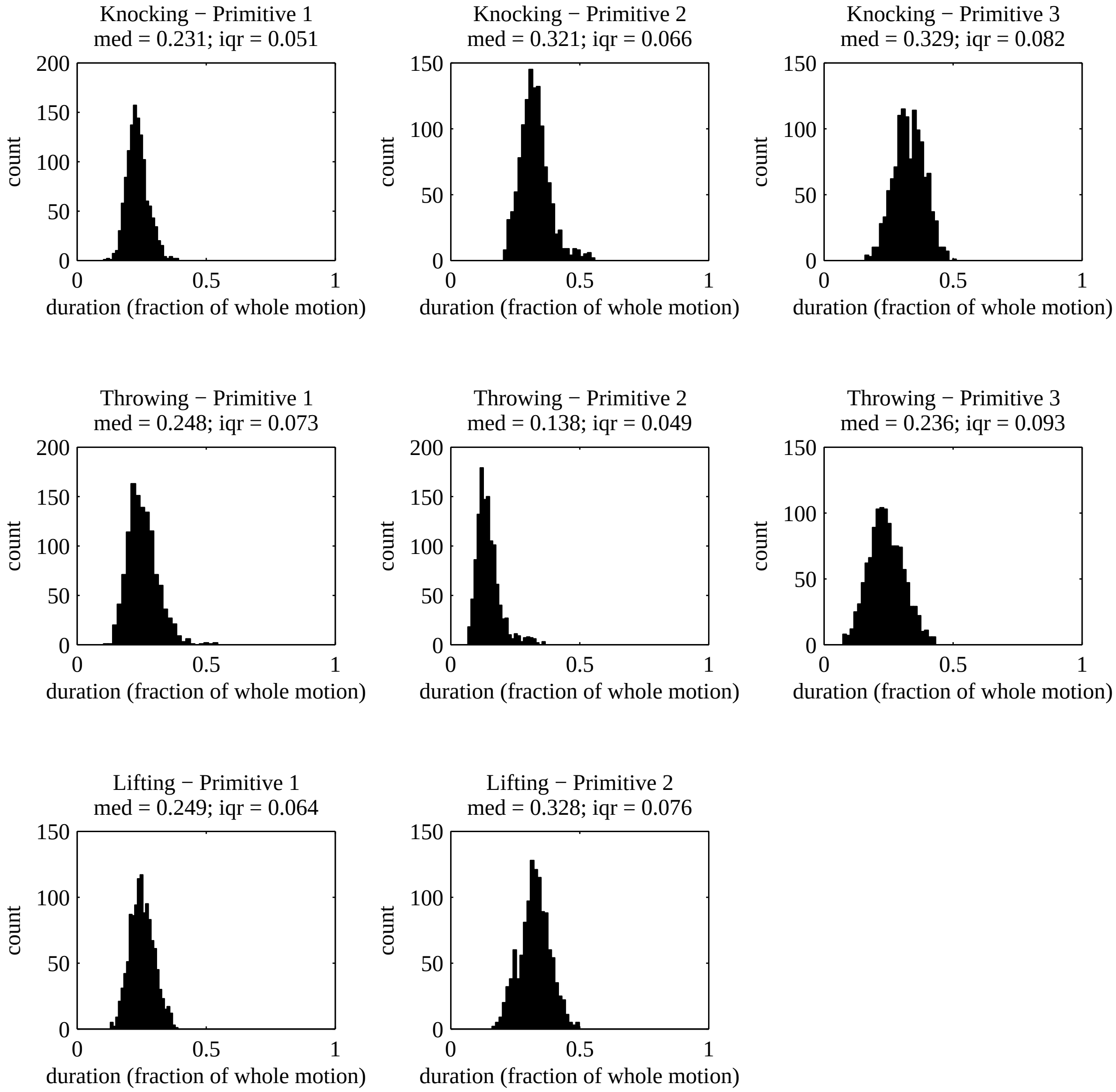

**Figure D.1:** Duration distributions for alignment-based primitives relative to action durations. Each histogram also lists the distribution's average (median) and spread (interquartile range).

# Appendix E

# Detailed feature analysis results

## E.1 Feature lists

Below is the list of features selected as informative in Section 5.4.2, Figure 5.8. These features were picked from vectors containing both primitive-local and action-global features.

### Knocking

*Primitive 1*:

median wrist height (primitive-local)
maximum elbow speed (primitive-local)
median wrist accel (primitive-local)

std dev of wrist height (action-global)
median elbow speed (action-global)
median wrist speed (action-global)
minimum elbow speed (action-global)
std dev of wrist speed (action-global)
median elbow accel (action-global)
median wrist accel (action-global)
minimum elbow accel (action-global)
std dev of wrist accel (action-global)
median wrist jerk (action-global)
std dev of wrist jerk (action-global)

median head roll (action-global)
median head pitch (action-global)
std dev of head pitch (action-global)
objective function median
objective function mean
optimal segmentation threshold

*Primitive 2*:

median wrist speed (primitive-local)
maximum wrist speed (primitive-local)
minimum elbow speed (primitive-local)
elbow-wrist phase shift
median wrist accel (primitive-local)
segment length
std dev of wrist height (action-global)
median elbow speed (action-global)
median wrist speed (action-global)
maximum wrist speed (action-global)

maximum wrist accel (primitive-local)
minimum elbow accel (primitive-local)
minimum wrist accel (primitive-local)
median elbow jerk (primitive-local)
maximum wrist jerk (primitive-local)
minimum elbow jerk (primitive-local)
minimum wrist jerk (primitive-local)
std dev of elbow jerk (primitive-local)
std dev of wrist jerk (primitive-local)
std dev of head pitch (primitive-local)

std dev of elbow speed (action-global)
std dev of wrist speed (action-global)
median wrist accel (action-global)
std dev of wrist accel (action-global)
median head roll (action-global)
median head pitch (action-global)
objective function mean
optimal segmentation threshold

*Primitive 3*:

median wrist height (primitive-local)

std dev of wrist height (action-global)
median elbow speed (action-global)
median wrist speed (action-global)
maximum wrist speed (action-global)
minimum elbow speed (action-global)
std dev of elbow speed (action-global)
median elbow accel (action-global)
median wrist accel (action-global)
minimum elbow accel (action-global)
std dev of wrist accel (action-global)
median wrist jerk (action-global)
median head roll (action-global)
median head pitch (action-global)
std dev of head pitch (action-global)
objective function median
objective function mean
optimal segmentation threshold

### Throwing

*Primitive 1*:

minimum elbow height (primitive-local)
maximum wrist speed (primitive-local)
minimum elbow speed (primitive-local)
median wrist accel (primitive-local)
maximum wrist accel (primitive-local)
maximum head pitch (primitive-local)

median wrist height (action-global)
maximum wrist height (action-global)
maximum wrist speed (action-global)
std dev of wrist speed (action-global)
median wrist accel (action-global)
maximum wrist accel (action-global)
std dev of wrist accel (action-global)
median wrist jerk (action-global)
maximum wrist jerk (action-global)
median head roll (action-global)
objective function mean

*Primitive 2*:

minimum wrist height (primitive-local)

maximum wrist speed (action-global)

std dev of wrist height (primitive-local)
median wrist speed (primitive-local)
maximum wrist speed (primitive-local)
median wrist accel (primitive-local)
maximum wrist jerk (primitive-local)

median wrist height (action-global)
maximum wrist height (action-global)
minimum elbow height (action-global)
std dev of wrist speed (action-global)
median wrist accel (action-global)
std dev of wrist accel (action-global)
median wrist jerk (action-global)
median head roll (action-global)
median head pitch (action-global)
objective function mean
optimal segmentation threshold

*Primitive 3*:

median elbow height (primitive-local)
maximum wrist height (primitive-local)
std dev of wrist height (primitive-local)
elbow-wrist phase shift
maximum head roll (primitive-local)

median wrist height (action-global)
maximum wrist height (action-global)
median wrist speed (action-global)
maximum wrist speed (action-global)
std dev of wrist speed (action-global)
median wrist accel (action-global)
maximum wrist accel (action-global)
std dev of wrist accel (action-global)
median wrist jerk (action-global)
maximum wrist jerk (action-global)
median head pitch (action-global)
objective function mean

**Lifting**

*Primitive 1*:

median wrist height (primitive-local)
median wrist speed (primitive-local)
maximum wrist speed (primitive-local)
std dev of elbow accel (primitive-local)
std dev of wrist accel (primitive-local)
median head pitch (primitive-local)
std dev of head roll (primitive-local)
segment length

median wrist height (action-global)
minimum wrist height (action-global)
std dev of wrist height (action-global)
median wrist speed (action-global)
maximum wrist speed (action-global)
minimum elbow speed (action-global)
std dev of wrist speed (action-global)
elbow-wrist phase shift
median elbow accel (action-global)
median wrist accel (action-global)
maximum wrist accel (action-global)
std dev of elbow accel (action-global)
std dev of wrist accel (action-global)
median wrist jerk (action-global)
std dev of wrist jerk (action-global)
std dev of head pitch (action-global)
objective function mean
optimal segmentation threshold

*Primitive 2*:

minimum wrist height (primitive-local)
std dev of wrist height (primitive-local)
median wrist speed (primitive-local)
median wrist accel (primitive-local)
maximum head pitch (primitive-local)
segment length

median wrist height (action-global)
maximum wrist height (action-global)
median wrist speed (action-global)
maximum wrist speed (action-global)
minimum elbow speed (action-global)
std dev of wrist speed (action-global)
elbow-wrist phase shift
median elbow accel (action-global)
median wrist accel (action-global)
maximum elbow accel (action-global)
maximum wrist accel (action-global)
std dev of elbow accel (action-global)
std dev of wrist accel (action-global)
std dev of wrist jerk (action-global)
std dev of head pitch (action-global)
objective function mean
optimal segmentation threshold

## E.2 Global-only features

The same feature selection as in Section 5.4 was run on the action-global feature vectors to isolate the single-most informative features for each action without confounding the list by the influence of motion primitives. The results are listed below.

*Knocking*:

median wrist height
std dev of wrist height
median elbow speed
median wrist speed
maximum wrist speed
minimum elbow speed
std dev of elbow speed
median elbow accel
median wrist accel
minimum elbow accel
std dev of wrist accel
median wrist jerk
median head roll
median head pitch
std dev of head pitch
objective function median
objective function mean
optimal segmentation threshold

*Throwing*:

median wrist height
maximum wrist height
minimum elbow height
median wrist speed
maximum wrist speed
std dev of wrist speed
median wrist accel
maximum wrist accel
std dev of wrist accel
median wrist jerk
maximum wrist jerk
median head roll
median head pitch
objective function mean

*Lifting*:

median wrist height
minimum wrist height
std dev of wrist height
median wrist speed
maximum wrist speed
minimum elbow speed
std dev of wrist speed
elbow-wrist phase shift
median elbow accel
median wrist accel
maximum elbow accel
maximum wrist accel
std dev of wrist accel
median wrist jerk
maximum head pitch
std dev of head roll
objective function mean
optimal segmentation threshold

## E.3 Informative features by emotion

In order to find out which features are particularly informative for individual emotions, I employed a slightly different methodology. The earlier results were all derived by considering all four emotions simultaneously. The previously listed features are therefore informative for distinguishing between any of the four emotions. Below are the results obtained for rerunning the feature selection procedure after singling out a particular emotion to investigate.

All samples of the particular emotion in question were re-labelled with a new emotion label 1 while *all other* samples were re-labelled with emotion label 2. The latter set of samples was sub-sampled to ensure an even distribution of samples for emotions 1 *vs*. 2 during the feature selection procedure. The features listed for a particular emotion below are thus informative for distinguishing between the particular emotion in question and the remaining three emotions in the set. As expected, the number of features needed to separate one single emotion is in every case smaller than the number of features informative for distinguishing between four emotions as found in the previous section.

### Knocking

*Neutral* vs. *all*:

std dev of wrist height
median elbow speed
median wrist speed
maximum wrist speed
median wrist accel
maximum wrist accel
minimum wrist accel
median head pitch
std dev of head pitch
objective function mean
optimal segmentation threshold

*Happy* vs. *all*:

median elbow speed
maximum wrist speed
std dev of wrist speed
median wrist accel
std dev of wrist accel
median wrist jerk
maximum head roll
std dev of head pitch
objective function mean

*Angry* vs. *all*:

maximum wrist height
minimum wrist height
median elbow speed
minimum elbow speed
minimum wrist speed
elbow-wrist phase shift
median wrist accel
std dev of wrist accel
median wrist jerk
std dev of wrist jerk
std dev of head roll
std dev of head pitch
objective function median
objective function mean
optimal segmentation threshold

*Sad* vs. *all*:

maximum elbow height
std dev of wrist height
median elbow speed
median wrist speed
maximum elbow speed
maximum wrist speed
minimum elbow speed
std dev of wrist speed
elbow-wrist phase shift
median wrist accel
minimum elbow accel
median head roll
median head pitch
maximum head pitch
objective function mean
optimal segmentation threshold

**Throwing**

*Neutral* vs. *all*:

median elbow height
median wrist height
maximum elbow height
median wrist speed
maximum wrist speed
minimum elbow speed
std dev of wrist speed
median elbow accel
median wrist accel
maximum wrist accel
median wrist jerk
median head pitch
std dev of head roll
objective function mean

*Happy* vs. *all*:

median wrist height
minimum elbow height
std dev of elbow height
maximum wrist speed
median wrist accel
std dev of wrist accel

*Angry* vs. *all*:

median wrist height
maximum wrist height
minimum wrist speed
elbow-wrist phase shift
median wrist accel
maximum wrist accel
std dev of wrist accel
minimum head roll
objective function median

*Sad* vs. *all*:

median wrist height
maximum wrist height
minimum elbow height
std dev of wrist height
std dev of elbow speed
median elbow accel
median wrist accel
minimum elbow accel
minimum wrist accel
median wrist jerk
minimum wrist jerk
median head pitch
objective function mean
optimal segmentation threshold

**Lifting**

*Neutral* vs. *all*:

median elbow height
median wrist height
minimum elbow height
minimum wrist height
median wrist speed
maximum wrist speed
std dev of elbow speed
std dev of wrist speed
median wrist accel
maximum wrist accel
std dev of elbow accel
std dev of wrist accel
median head pitch
objective function mean

*Happy* vs. *all*:

median wrist height
std dev of elbow height
median wrist speed
maximum wrist speed
maximum wrist accel
std dev of wrist accel
maximum head pitch
std dev of head roll

std dev of wrist speed
median wrist accel
std dev of head pitch

*Angry* vs. *all*:

minimum wrist height
std dev of wrist speed
median wrist accel
maximum elbow accel
maximum wrist accel
minimum wrist accel
std dev of elbow accel
std dev of wrist accel
median wrist jerk
minimum wrist jerk
std dev of elbow jerk
objective function mean

*Sad* vs. *all*:

median wrist height
maximum wrist height
std dev of elbow height
std dev of wrist height
median wrist speed
maximum wrist speed
std dev of wrist speed
elbow-wrist phase shift
median elbow accel
median wrist accel
minimum wrist accel
std dev of elbow accel
median head pitch
std dev of head pitch
objective function median
objective function mean
optimal segmentation threshold

# Appendix F

# Error analysis for bias prediction

My primary means of evaluating the accuracy of motion bias prediction has been in terms of the resulting emotion recognition performance. Only in this way it is possible to really judge how accurate the bias prediction really has to be in order to build a more elaborate recognition system on top of it. In this appendix I am giving additional results for evaluating the bias prediction framework independently.

As ground truth I am using every person's motion bias ($\hat{\boldsymbol{\phi}}_{\text{Knock},p}$, $\hat{\boldsymbol{\phi}}_{\text{Throw},p}$, $\hat{\boldsymbol{\phi}}_{\text{Lift},p}$, $\hat{\boldsymbol{\phi}}_{\text{Walk},p}$) as computed from the action samples. In order to evaluate the accuracy of the regression-based prediction, I am using a 10-fold cross validation approach. For each fold $n$, a regression function $B_{A,n}$ is estimated during the training phase based on the 27 training subjects' *TDown* adaptation variables ($\hat{\boldsymbol{\phi}}_{\text{TDown},p}$) and bias variables ($\hat{\boldsymbol{\phi}}_{A,p}$). During the evaluation phase, the remaining 3 subjects' bias variables are estimated based on their observed *TDown* variables as $\hat{\boldsymbol{\phi}}_{A,p} = B_{A,n}(\hat{\boldsymbol{\phi}}_{\text{TDown},p})$. I then compute the root mean squared error (RMSE) between the true bias as computed from the action itself and the predicted bias. Because each dimension is normalised prior to this step, I do not need to use a more complex error measure like the Mahalanobis distance in this case.

Table F.1 lists the RMSE for each subject and action combination as well as aggregates over all subject. We can see that the estimates for the *Knocking* motion bias are most accurate. The accuracy of bias estimation as computed in this experiment also correlates well with the resulting emotion recognition accuracies (see Section 5.5.3). Finally, the significantly worse bias prediction performance for *Walking* is not unexpected, as bias variables for a lower body action have to be estimated from variables derived from the upper body action *TDown*. In Section 5.5.3 we saw that trying to estimate the motion bias for *Walking* actions leads to no improvements in emotion recognition over fully biased signals.

**Table F.1:** Root mean squared errors for bias estimation using linear regression.

| Subject | *Knocking* | *Throwing* | *Lifting* | *Walking* |
|---|---|---|---|---|
| subject 1 | 1.88 | 2.41 | 0.82 | 0.92 |
| subject 2 | 1.21 | 1.92 | 2.74 | 1.76 |
| subject 3 | 1.69 | 1.53 | 1.98 | 2.64 |
| subject 4 | 2.96 | 1.81 | 1.22 | 2.57 |
| subject 5 | 0.68 | 1.47 | 0.77 | 2.43 |
| subject 6 | 1.38 | 1.96 | 2.66 | 12.99 |
| subject 7 | 1.85 | 2.65 | 1.97 | 5.08 |
| subject 8 | 0.49 | 1.55 | 1.05 | 1.03 |
| subject 9 | 0.85 | 1.96 | 1.15 | 1.23 |
| subject 10 | 0.97 | 2.05 | 1.32 | 1.80 |
| subject 11 | 1.81 | 0.52 | 1.40 | 1.32 |
| subject 12 | 0.83 | 0.65 | 0.84 | 1.47 |
| subject 13 | 0.62 | 1.71 | 1.35 | 2.84 |
| subject 14 | 0.88 | 0.95 | 1.13 | 0.82 |
| subject 15 | 2.22 | 0.68 | 2.11 | 2.50 |
| subject 16 | 0.54 | 1.30 | 1.83 | 1.82 |
| subject 17 | 1.91 | 1.39 | 1.24 | 3.60 |
| subject 18 | 0.79 | 1.51 | 1.52 | 1.99 |
| subject 19 | 0.78 | 1.13 | 1.38 | 1.65 |
| subject 20 | 2.45 | 1.48 | 1.22 | 2.25 |
| subject 21 | 0.46 | 1.58 | 1.73 | 3.70 |
| subject 22 | 0.49 | 1.77 | 1.61 | 1.17 |
| subject 23 | 1.05 | 0.95 | 1.20 | 1.85 |
| subject 24 | 0.24 | 1.26 | 1.92 | 1.01 |
| subject 25 | 1.11 | 1.13 | 1.38 | 1.20 |
| subject 26 | 2.81 | 1.32 | 1.88 | 1.56 |
| subject 27 | 0.78 | 1.74 | 0.84 | 2.55 |
| subject 28 | 0.56 | 2.10 | 1.15 | 1.05 |
| subject 29 | 1.12 | 2.01 | 1.67 | 3.49 |
| subject 30 | 1.14 | 1.23 | 1.24 | 3.46 |
| mean | 1.22 | 1.52 | 1.48 | 2.46 |
| std deviation | 0.72 | 0.50 | 0.50 | 2.23 |

# Appendix G

# Symbols

| | |
|---|---|
| $\boldsymbol{v}(t)$ | Observation vector at time frame $t$ |
| $\boldsymbol{V}^T$ | Sequence of observations |
| $\mathcal{V}$ | Set of observation sequences |
| $W(x)$ | Dynamic Time Warp function |
| $\tilde{\boldsymbol{V}}^{\tilde{T}}$ | Time-warped observation sequence |
| $\delta(\boldsymbol{v}_1, \boldsymbol{v}_2)$ | Distance function for two observations |
| $\delta_s(v_1, v_2)$ | Distance function for two sets of observations |
| $\Delta(W)$ | Distance function for two warped sequences |
| $\dot{\theta}$ | Joint speed |
| $E(t)$ | Motion energy at time frame $t$ |
| $c$ | Action category |
| $\mathcal{C}$ | Set of actions |
| $\mathcal{J}_c$ | Body joints used by action $c$ |
| $\lambda_c$ | Hidden Markov Model (HMM) for action $c$ |
| $d_c$ | Dimensionality of the observation vectors of $\lambda_c$ |
| $s$ | Number of distinct HMM states |
| $\omega_i$ | HMM state |
| $\boldsymbol{\omega}^T$ | HMM state sequence |
| $\omega(t)$ | HMM state at time frame $t$ |
| $\boldsymbol{A}$ | HMM transition matrix |
| $\boldsymbol{\pi}$ | HMM state prior |
| $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ | Normal density parameterised by mean and covariance |
| $\Lambda$ | Set of action HMMs |
| $\lambda(l)$ | Action at level $l$ in an action sequence |
| $b(0) \dots b(L)$ | Action boundaries in an action sequence |
| $\mathcal{B}$ | Segmentations of a set of action sequences |
| $V$ | Dynamic programming matrix |
| $t(i, j, l)$ | Transition function between levels in an action sequence |

| | |
|---|---|
| $P_c(d_t)$ | Regularisation term for segments of duration $d_t$ |
| $\gamma$ | Regularisation weight |
| $\phi$ | Feature used for emotion recognition |
| $\Phi$ | Set of features |
| $\boldsymbol{\phi}_{m,p}$ | Feature vector for motion primitive $m$ and person $p$ |
| $\bar{\boldsymbol{\phi}}_{m,p}$ | Motion bias |
| $\hat{\boldsymbol{\phi}}_{m,p}$ | Unbiased feature vector |
| $B(\boldsymbol{\phi})$ | Bias adaptation function |
| $e$ | Emotion class |
| $\mathcal{E}$ | Set of emotions |
| $M^m_{e_1,e_2}$ | Binary SVM for motion primitive $m$ and emotion classes $e_1$ and $e_2$ |
| $K(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$ | SVM kernel function |
| $\rho_{\phi_1,\phi_2}$ | Correlation coefficient between two features |
| $\tau$ | Energy/correlation threshold |
| $numseg_E(\tau)$ | Function computing the number of segments by thresholding the motion energy signal |
| $Merit(\Phi)$ | Quality measure of a feature subset for predicting emotion |
| $\zeta$ | Fraction of correctly segmented action sequences |
| $\alpha$ | Fraction of correctly labeled frames in an action sequence |
| $\eta$ | Emotion recognition efficiency |

# Appendix H

# Glossary

**Action-global features** Features derived from the statistics of a whole action rather than individual motion primitives.

**Adaptation** Process of adjusting a number of predefined parameters in a machine recognition engine to better fit a particular user's idiosyncrasies.

**Biased features** Original motion features before subtracting the person-specific motion signature.

**Bias-adapted features** Person-normalised motion features generated by subtracting an *approximated* motion signature from the original, biased features.

**Bootstrapping** Machine learning technique which improves a classifier by iteratively training and evaluating it on a set of data with partially known structure.

**Cambridge Corpus** Database of motion-captured expressive body motions collected for this research.

**Coarticulation** Effects of neighbouring units of speech or motion on each others' appearance.

**Cross Validation** Evaluation technique which estimates the generalisation performance of a classifier. Data is repeatedly split into training and validation sets, each of which are disjoint and come from different subjects.

**Discriminant function** In pattern classification, a scalar function defined over a feature space and representing a particular class. A classifier evaluates all discriminant functions at a certain point in feature space to assign it to the class with the highest discriminant function output.

**Discriminative classifier** In machine learning, a model which has been trained to best separate samples belonging to a number of distinct classes.

**Dynamic Time Warping** Method of aligning two time series allowing for non-linear variations in correspondences.

**Emotional response triad** Three major components involved in an emotional response: physiological arousal, motor expression, subjective feeling.

**Feature subset selection** Process of identifying statistical features which are maximally correlated with a certain class but minimally correlated with each other.

**Feature vector** Multidimensional representation of motion statistics indicative of emotional state.

**Generalisation** Measure of a classifier's ability to classify samples from previously unseen subjects.

**Generative classifier** In machine learning, a model which has been trained to best represent the data associated with a certain class independent of other classes.

**Glasgow Corpus** Database of motion-captured actions performed in different emotional styles collected at the Psychology Department, University of Glasgow.

**Guide tree** Rooted acyclic graph determining the order of individual alignments for multiple sequence alignment.

**Hidden Markov Model** Model for a temporally evolving system producing multi-dimensional observations. Observations are conditioned only on hidden states governed over time by a Markov Process.

**Level Building** Approach for segmenting connected gestures by combining individual HMMs through Dynamic Programming.

**Mood induction procedure** In experimental psychology, way of systematically manipulating a subject's felt mood/emotion.

**Motion bias** Posture and dynamic idiosyncrasies exhibited by a person while carrying out an action, see also biased and unbiased features.

**Motion capture** Process of recording body movement at high spatial and temporal resolution using special hardware.

**Motion descriptor** Channel capturing an aspect of continuously varying body posture or motion dynamics.

**Motion primitive** Basic section of movement, several of which are combined to produce complex actions.

**Motion quality** Parameters of body motion such as smoothness and speed.

**Multiple Discriminant Analysis** Statistical method to find the linear combinations of features which best separate two or more classes of samples.

**Multiple sequence alignment** The problem of temporally aligning multiple sequences of symbols or time series.

**Posture** Static configuration of body parts, including head pose, arm, leg and trunk configuration.

**Primitive-local features** Features derived from the statistics of individual motion primitives used to model a more complex action.

**Support Vector Machine** Feature-based linear classifier aimed at maximising the margin between samples of two classes during statistical training. Feature vectors are commonly represented in high-dimensional spaces using kernels

**Unbiased features** Person-normalised motion features generated by subtracting the person-specific motion signature from the original, biased features.

# References

[ABC96] Kenji Amaya, Armin Bruderlin, and Tom Calvert. Emotion from motion. In *GI '96: Proceedings of the conference on Graphics interface '96*, pages 222–229, Toronto, Ont., Canada, Canada, 1996. Canadian Information Processing Society.

[Abd02] Chiraz B. Abdelkader. Stride and cadence as a biometric in automatic person identification and verification. In *FGR '02: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 372–377, Washington, DC, USA, 2002. IEEE Computer Society.

[ADGY04] A. P. Atkinson, W. H. Dittrich, A. J. Gemmell, and A. W. Young. Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception*, 33:717–746, 2004.

[AG98] Marcell Assan and Kirsti Grobel. Video-based sign language recognition using hidden markov models. In *Gesture and Sign Language in Human-Computer Interaction: Proceedings of the International Gesture Workshop*, pages 97–109. Springer, 1998.

[AG00] Eugene Agichtein and Luis Gravano. Snowball: extracting relations from large plain-text collections. In *DL '00: Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94, New York, NY, USA, 2000. ACM.

[AMR09] Shazia Afzal, Cecily Morrison, and Peter Robinson. Intentional affect: an alternative notion of affective interaction with a machine. In *BCS HCI '09: Proceedings of the 2009 British Computer Society Conference on Human-Computer Interaction*, pages 370–374, Swinton, UK, UK, 2009. British Computer Society.

[ATD07] A. Atkinson, M. Tunstall, and W. Dittrich. Evidence for distinct contributions of form and motion information to the recognition of emotions from body gestures. *Cognition*, 104(1):59–72, July 2007.

[AZ07] Marilyn Augustyn and Barry Zuckerman. From mother's mouth to infant's brain. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 92(2):F82+, March 2007.

[BC01] Thomas R. Boone and Joseph G. Cunningham. Children's expression of emotional meaning in music through expressive body movement. *Journal of Nonverbal Behavior*, 25(1):21–41, March 2001.

[BFH⁺03] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. How to find trouble in communication. *Speech Commun.*, 40(1-2):117–143, April 2003.

[BH00] Matthew Brand and Aaron Hertzmann. Style machines. In *SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 183–192, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co.

[Bir70] Ray L. Birdwhistell. *Kinesics and Context: Essays on Body Motion Communication*. University of Pennsylvania Press, 1970.

[BLF⁺06] Marian S. Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek1, Ian Fasel, and Javier Movellan. Fully automatic facial action recognition in spontaneous behavior. In *FGR '06: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, pages 223–230, Washington, DC, USA, 2006. IEEE Computer Society.

[BLFM03] Marian S. Bartlett, Gwen Littlewort, Ian Fasel, and Javier R. Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *CVPRW '03: Proceedings of the Conference on Computer Vision and Pattern Recognition, Workshop, 2003*, volume 5, page 53, 2003.

[BLV⁺08] Marian Bartlett, Gwen Littlewort, Esra Vural, Kang Lee, Mujdat Cetin, Aytul Ercil, and Javier Movellan. Data mining spontaneous facial behavior with automatic expression coding. In *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*, pages 1–20, Berlin, Heidelberg, 2008. Springer-Verlag.

[Bob97] Aaron F. Bobick. Movement, activity and action: The role of knowledge in the perception of motion. *Philosophical Transactions: Biological Sciences*, 352(1358):1257–1265, 1997.

[BR07] Daniel Bernhardt and Peter Robinson. Detecting affect from non-stylised body motions. In *ACII'07: Proceedings of the Seconnd International Conference on Affective Computing and Intelligent Interaction*, pages 59–70, Berlin, Heidelberg, 2007. Springer-Verlag.

[BR08] Daniel Bernhardt and Peter Robinson. Interactive control of music using emotional body expressions. In *CHI '08: CHI '08 extended abstracts on Human factors in computing systems*, pages 3117–3122, New York, NY, USA, 2008. ACM.

[Bre04] C. Breazeal. Function meets style: insights from emotion theory applied to HRI. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 34(2):187–194, 2004.

[Bri99] Sergey Brin. Extracting patterns and relations from the world wide web. In *The World Wide Web and Databases: Proceedings of the International Workshop WebDB'98*, pages 172–183. Springer, 1999.

[BRI$^+$05] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias. Emotion analysis in man-machine interaction systems. In *Procceedings of the First International Workshop on Machine Learning for Multimodal Interaction*, pages 318–328. Springer, 2005.

[BS07] Tanja Bänziger and Klaus Scherer. Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus. In *ACII '07: Proceedings of the Second International Conference on Affective Computing and Intelligent Interaction*, pages 476–487. Springer, 2007.

[BSB$^+$07] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker. Detailed human shape and pose from images. In *CVPR '07: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007*, pages 1–8, 2007.

[Bul87] Peter E. Bull. *Posture and Gesture*, volume 16. Pergamon Press, 1987.

[Bur98] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, June 1998.

[Cam97] J. P. Campbell. Speaker recognition: a tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, 1997.

[Cas08] Ginevra Castellano. *Movement expressivity analysis in affetive computers: from recognition to expression of emotion*. PhD thesis, Department of Communication, Faculty of Engineering, University of Genova, 2008.

[CBCV07] Ginevra Castellano, Roberto Bresin, Antonio Camurri, and Gualtiero Volpe. User-centered control of audio and visual expressive feedback by full-body movements. In *ACII'07: Proceedings of the Second International Conference on Affective Computing and Intelligent Interaction*, pages 501–510, 2007.

[CBF+05] T. J. Clarke, M. F. Bradshaw, D. T. Field, S. E. Hampson, and D. Rose. The perception of emotion from body movement in point-light displays of interpersonal dialogue. *Perception*, 34:1171–1180, 2005.

[CCS+07] Ellen D. Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean C. Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, Noam Amir, and Kostas Karpouzis. The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. In *ACII'07: Proceedings of the Second International Conference on Affective Computing and Intelligent Interaction*, pages 488–500, Berlin, Heidelberg, 2007. Springer-Verlag.

[CCZB00] Diane Chi, Monica Costa, Liwei Zhao, and Norman Badler. The EMOTE model for effort and shape. In *SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 173–182, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co.

[CD97] Scott S. Chen and Peter Desouza. Speaker adaptation by correlation (ABC). In *EUROSPEECH-1997*, pages 2111–2114, 1997.

[CDC96] R. Cowie and E. Douglas-Cowie. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In *ICSLP 96: Proceedings of the Fourth International Conference on Spoken Language, 1996*, volume 3, pages 1989–1992 vol.3, 1996.

[CDCT+02] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, August 2002.

[Ced95] C. Cedras. Motion-based recognition a survey. *Image and Vision Computing*, 13(2):129–155, March 1995.

[CG07] Elizabeth Crane and Melissa Gross. Motion capture and emotion: Affect detection in whole body movement. In *ACII'07: Proceedings of the Second International Conference on Affective Computing and Intelligent Interaction*, pages 95–101. Springer, 2007.

[CKC08] Ginevra Castellano, Loic Kessous, and George Caridakis. Emotion recognition through multiple modalities: Face, body gesture, speech. In *Affect and Emotion in Human-Computer Interaction*, pages 92–103. Springer, 2008.

[CKK08] George Caridakis, Kostas Karpouzis, and Stefanos Kollias. User and context adaptive neural networks for emotion recognition. *Neurocomputing*, 71(13-15):2553–2562, 2008.

[CKKR08] S. Cherla, K. Kulkarni, A. Kale, and V. Ramasubramanian. Towards fast, view-invariant human action recognition. In *CVPRW '08: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Workshops, 2008*, pages 1–8, 2008.

[CLV03] Antonio Camurri, Ingrid Lagerlöf, and Gualtiero Volpe. Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies*, 59(1-2):213–225, 2003.

[CMC$^+$08] Ginevra Castellano, Marcello Mortillaro, Antonio Camurri, Gualtiero Volpe, and Klaus Scherer. Automated analysis of body movement in emotionally expressive piano performances. *Music Perception*, 26(2):103–119, December 2008.

[CMV04] Antonio Camurri, Barbara Mazzarino, and Gualtiero Volpe. Analysis of expressive gesture: The EyesWeb expressive gesture processing library. In *5th International Gesture Workshop: Gesture-Based Communication in Human-Computer Interaction*, pages 469–470. Springer, 2004.

[Coh06] Jeffrey F. Cohn. Foundations of human computing: facial expression and emotion. In *ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces*, pages 233–238, New York, NY, USA, 2006. ACM.

[Col02] Michael Collins. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 1–8, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[Cou04] Mark Coulson. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of Nonverbal Behavior*, 28(2):117–139, June 2004.

[CS02] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2002.

[CS04] Jeffrey Cohn and Karen Schmidt. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, 2, March 2004.

[CS05] Roddy Cowie and Marc Schröder. Piecing together the emotion jigsaw. In Samy Bengio and Hervé Bourlard, editors, *MLMI 2004: Revised Selected*

*Papers of the First International Workshop on Machine Learning for Multimodal Interaction*, volume 3361 of *Lecture Notes in Computer Science*, pages 305–317, Berlin, 2005. Springer Verlag.

[CSG+03] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S. Chen, and Thomas S. Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding*, 91(1-2):160–187, August 2003.

[CSM93] Linda A. Camras, Jean Sullivan, and George Michel. Do infants express discrete emotions? adult judgments of facial, vocal, and body actions. *Journal of Nonverbal Behavior*, 17(3):171–186, September 1993.

[CVC07] Ginevra Castellano, Santiago Villalba, and Antonio Camurri. Recognising human emotions from body movement and gesture dynamics. In *ACII'07: Proceedings of the Second International Conference on Affective Computing and Intelligent Interaction*, pages 71–82, Berlin, Heidelberg, 2007. Springer-Verlag.

[Dar72] Charles Darwin. *The Expression of the Emotions in Man and Animals.* University of Chicago Press, Chicago, 1872.

[Das01] Sanmay Das. Filters, wrappers and a boosting-based hybrid for feature selection. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 74–81, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[DCDM+05] Ellen Douglas-Cowie, Laurence Devillers, Jean-Claude Martin, Roddy Cowie, Suzie Savvidou, Sarkis Abrilian, and Cate Cox. Multimodal databases of everyday emotion: Facing up to complexity. In *INTERSPEECH 2005*, pages 813–816, 2005.

[DCW+08] Sidney D'Mello, Scotty Craig, Amy Witherspoon, Bethany McDaniel, and Arthur Graesser. Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction*, 18(1):45–80, February 2008.

[DCXL06] Youtian Du, Feng Chen, Wenli Xu, and Yongbin Li. Recognizing interaction activities using dynamic bayesian network. In *Proceedins of the International Conference on Pattern Recognition*, volume 1, pages 618–621, Los Alamitos, CA, USA, 2006. IEEE Computer Society.

[DHS00] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Calssification.* Wiley, 2nd edition, 2000.

[DP93] T. Darrell and A. Pentland. Space-time gestures. In *CVPR '93: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1993*, pages 335–340, 1993.

[DSBB04] P. Ravindra De Silva and Nadia Bianchi-Berthouze. Modeling human affective postures: an information theoretic characterization of posture features. *Computer Animation and Virtual Worlds*, 15(3-4):269–276, 2004.

[DSOMM06] P. R. De Silva, M. Osano, A. Marasinghe, and A. P. Madurapperuma. Towards recognizing emotion with affective dimensions through body gestures. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 269–274, April 2006.

[DTLM96] W. H. Dittrich, T. Troscianko, S. E. G. Lea, and D. Morgan. Perception of emotion from dynamic point-light displays represented in dance. *Perception*, 25(6):727–738, 1996.

[EF74] Paul Ekman and Wallace V. Friesen. Detecting deception from body or face. *Journal of Personality and Social Psychology*, 29:288–298, 1974.

[EF76] Paul Ekman and Wallace V. Friesen. *Pictures of Facial Affect.* Consulting Psychologists Press, 1976.

[EF78] Paul Ekman and Wallace V. Friesen. *Facial action coding system.* Consulting Psychologists Press, 1978.

[Ekm64] Paul Ekman. Differential communication of affect by head and body cues. *Journal of Personality and Social Psychology*, pages 726–735, 1964.

[Ekm92] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200, 1992.

[Ekm94] Paul Ekman. Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique. *Psychology Bulletin*, 115(2):268–287, 1994.

[FA05] Jean-Marc Fellous and Michael A. Arbib, editors. *Who needs emotions? - The brain meets the robot.* Oxford University Press US, 2005.

[FH85] T. Flash and N. Hogan. The coordination of arm movements: an experimentally confirmed mathematical model. *Journal of Neuroscience*, 5(7):1688–1703, July 1985.

[FMJ02] Ajo Fod, Maja J. Matarić, and Odest C. Jenkins. Automated derivation of primitives for movement classification. *Autononmous Robots*, 12(1):39–54, 2002.

[FRL04] Katherine Forbes-Riley and Diane J. Litman. Predicting emotion in spoken dialogue from multiple knowledge sources. In *HLT-NAACL*, pages 201–208, 2004.

[FSH04] Petra Fagerberg, Anna Ståhl, and Kristina Höök. eMoto: emotionally engaging interaction. *Personal Ubiquitous Computing*, 8(5):377–381, 2004.

[FST98] Shai Fine, Yoram Singer, and Naftali Tishby. The Hierarchical Hidden Markov Model: Analysis and applications. *Machine Learning*, 32(1):41–62, 1998.

[FT05] N. Fragopanagos and J. G. Taylor. Emotion recognition in human-computer interaction. *Neural Networks*, 18(4):389–405, May 2005.

[Fur86] S. Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(1):52–59, 1986.

[Gav99] D. M. Gavrila. The visual analysis of human movement: a survey. *Comput. Vis. Image Underst.*, 73(1):82–98, January 1999.

[GCP07] H. M. Genc, Z. Cataltepe, and T. Pearson. A new pca/ica based feature selection method. In *Signal Processing and Communications Applications, 2007. SIU 2007. IEEE 15th*, pages 1–4, June 2007.

[GG04] R. D. Green and Ling Guan. Quantifying and recognizing human movement patterns from monocular video images-part I: a new framework for modeling human motion. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(2):179–190, 2004.

[GGB06] D. Giuliani, M. Gerosa, and F. Brugnara. Improved automatic speech recognition through speaker normalization. *Computer Speech & Language*, 20(1):107–123, January 2006.

[GIB09] A. Gilbert, J. Illingworth, and R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *Proc. Int. Conference Computer Vision (ICCV09)*, 2009.

[Gie92] Heinz J. Giergerich. *English phonology: An introduction.* Cambridge University Press, 1992.

[Gil08] A. L. Gilet. Mood induction procedures: a critical review. *L'Encéphale*, 34(3):233–239, June 2008.

[GP05] H. Gunes and M. Piccardi. Affect recognition from face and body: early fusion vs. late fusion. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pages 3437–3443 Vol. 4, 2005.

[GP06] H. Gunes and M. Piccardi. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *ICPR: Proceedings of the 18th International Conference on Pattern Recognition*, volume 1, pages 1148–1153, 2006.

[GP07] Hatice Gunes and Massimo Piccardi. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30(4):1334–1345, 2007.

[GP08] H. Gunes and M. Piccardi. Automatic temporal segment detection and affect recognition from face and body display. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 39(1):64–84, August 2008.

[GPP08] H. Gunes, M. Piccardi, and M. Pantic. *From the Lab to the real world: affect recognition using multiple cues and modalities*, pages 185–218. InTech Education and Publishing, 2008.

[GR88] A. L. Gorin and D. B. Roe. Parallel level-building on a tree machine. In *ICASSP-88: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 1988*, pages 295–298, 1988.

[Gri02] W. E. L. Grimson. Gait analysis for recognition and classification. In *FGR '02: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, page 155, Washington, DC, USA, 2002. IEEE Computer Society.

[HCGM06] Alexis Heloir, Nicolas Courty, Sylvie Gibet, and Franck Multon. Temporal alignment of communicative gesture sequences: Research articles. *Computer Animation and Virtual Worlds*, 17(3-4):347–357, 2006.

[HDR08] Robert Horlings, Dragos Datcu, and Leon J. M. Rothkrantz. Emotion recognition using brain activity. In *CompSysTech '08: Proceedings of the 9th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing*, New York, NY, USA, 2008. ACM.

[HFH$^+$09] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009.

[HG04] Corey B. Hart and Simon F. Giszter. Modular premotor drives and unit bursts as primitives for frog motor behaviors. *Journal of Neuroscience*, 24(22):5269–5282, June 2004.

[HH99] William Hardcastle and Nigel Hewlett. *Coarticulation: Theory, Data and Techniques*. Cambridge University Press, 1999.

[HL02] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, August 2002.

[HS88] D. G. Higgins and P. M. Sharp. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73(1):237–244, December 1988.

[HS97] Mark A. Hall and Lloyd A. Smith. Feature subset selection: a correlation based filter approach. In *1997 International Conference on Neural Information Processing and Intelligent Information Systems*, pages 855–858. Springer, 1997.

[HS01] D. P. Huijsmans and N. Sebe. Extended performance graphs for cluster retrieval. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:26+, 2001.

[Hut77] A. Hutchinson. *Labanotation*. Theatre Art Books, 1977.

[HW77] P. W. Holland and R. E. Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics: Theory and Methods*, A6:813–827, 1977.

[IN90] Y. Ishikawa and K. Nakajima. A real time connected word recognition system. In *Pattern Recognition, 1990. Proceedings., 10th International Conference on*, volume ii, pages 215–217 vol.2, 1990.

[Iza94] C. E. Izard. Innate and universal facial expressions: evidence from developmental and cross-cultural research. *Psychological bulletin*, 115(2):288–299, March 1994.

[JB01] Amos Johnson and Aaron Bobick. A multi-view method for gait recognition using static body parameters. In *AVBPA 2001: Proceedings of the Third International Conference on Audio- and Video-Based Biometric Person Authentication*, pages 301–311. Springer, 2001.

[Jen01] Finn V. Jensen. *Bayesian Networks and Decision Graphs*. Springer, 2001.

[JKP94] George H. John, Ron Kohavi, and Karl Pfleger. Irrelevant features and the subset selection problem. In *Proceedings of the International Conference on Machine Learning*, pages 121–129, 1994.

[Joa99] Thorsten Joachims. *Making large-scale support vector machine learning practical*, pages 169–184. MIT Press, Cambridge, MA, USA, 1999.

[Joh67] Stephen Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, September 1967.

[JSL+08] Daniel Janssen, Wolfgang Schöllhorn, Jessica Lubienetzki, Karina Fölling, Henrike Kokenge, and Keith Davids. Recognition of emotions in gait patterns by means of artificial neural nets. *Journal of Nonverbal Behavior*, 32(2):79–92, June 2008.

[JT95] Ollie Johnston and Frank Thomas. *The Illusion of Life: Disney Animation.* Hyperion, 1995.

[Kal05] Rana E. Kaliouby. *Mind-reading machines: automated inference of complex mental states.* PhD thesis, University of Cambridge, Computer Laboratory, 2005.

[KBB07] Andrea Kleinsmith and Nadia Bianchi-Berthouze. Recognizing affective dimensions from body posture. In *ACII'07: Proceedings of the Second International Conference on Affective Computing and Intelligent Interaction*, pages 48–58. Springer, 2007.

[KBP07] Ashish Kapoor, Winslow Burleson, and Rosalind W. Picard. Automatic prediction of frustration. *Int. J. Hum.-Comput. Stud.*, 65(8):724–736, August 2007.

[KCHL03] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee. Emotion recognition by speech signals. In *EUROSPEECH-2003*, pages 125–128, 2003.

[Ken04] Adam Kendon. *Gesture: Visible Action as Utterance.* Cambridge University Press, 2004.

[KFN07] F. Kawsar, K. Fujinami, and T. Nakajima. Approaching intelligent environment through sentient artefacts. In *iccit 2007: Proceedings of the 10th international conference on Computer and information technology, 2008*, pages 1–6, 2007.

[KJ97] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Journal of Artificial Intelligence*, 97(1-2):273–324, 1997.

[KJNN00] R. Kuhn, J. C. Junqua, P. Nguyen, and N. Niedzielski. Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, 8(6):695–707, 2000.

[KKK02] Taeyoon Kim, Yongsung Kang, and Hanseok Ko. Achieving real-time lip synch via SVM-based phoneme classification and lip shape refinement. In *Proceedings of the IEEE International Conference on Multimodal Interfaces*, volume 0, pages 299+, Los Alamitos, CA, USA, 2002. IEEE Computer Society.

[KKVB$^+$05] Asha Kapur, Ajay Kapur, Naznin Virji-Babul, George Tzanetakis, and Peter Driessen. Gesture-based affective computing on motion capture data. In *ACII'05: Proceedings of the First International Conference on Affective Computing and Intelligent Interaction*, pages 1–7. Springer, 2005.

[KP05] I. Kotsia and I. Pitas. Real time facial expression recognition from image sequences using support vector machines. In *ICIP 2005: Proceedings of the IEEE International Conference on Image Processing, 2005*, volume 2, pages II–966–9, 2005.

[KR04] Rana E. Kaliouby and Peter Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *VPRW '04: Proceedings of the Conference on Computer Vision and Pattern Recognition, Workshop, 2004*, page 154, 2004.

[KR05] Rana Kaliouby and Peter Robinson. Generalization of a vision-based computational model of mind-reading. In *ACII'05: Proceedings of the First International Conference on Affective Computing and Intelligent Interaction*, pages 582–589. Springer, 2005.

[KS05] Taesoo Kwon and Sung Y. Shin. Motion modeling for on-line locomotion synthesis. In *SCA '05: Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 29–38, New York, NY, USA, 2005. ACM.

[KSBB05] Andrea Kleinsmith, P. Silva, and Nadia Bianchi-Berthouze. Grounding affective dimensions into posture features. In *ACII'05: Proceedings of the First International Conference on Affective Computing and Intelligent Interaction*, pages 263–270. Springer, 2005.

[KSR$^+$04] A. Kale, A. Sundaresan, A. N. Rajagopalan, N. P. Cuntoor, A. K. Roy-Chowdhury, V. Kruger, and R. Chellappa. Identification of humans using gait. *IEEE Transactions on Image Processing*, 13(9):1163–1173, 2004.

[KTP06] Rana E. Kaliouby, Alea Teeters, and Rosalind W. Picard. Invited talk: An exploratory social-emotional prosthetic for autism spectrum disorders. *International Workshop on Wearable and Implantable Body Sensor Network*, 0:3–4, 2006.

[Lab75] Rudolf Laban. *Principles of dance and movement notation.* McDonald & Evans, 2 edition, 1975.

[Las87] John Lasseter. Principles of traditional animation applied to 3D computer animation. In *SIGGRAPH '87: Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, pages 35–44, New York, NY, USA, 1987. ACM.

[LBF$^+$04] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. In *CVPRW '04: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop, 2004*, page 80, June 2004.

[LBL07] Gwen C. Littlewort, Marian S. Bartlett, and Kang Lee. Faces of pain: automated measurement of spontaneous facial expressions of genuine and posed pain. In *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces*, pages 15–21, New York, NY, USA, 2007. ACM.

[LE06] Chan-Su Lee and Ahmed Elgammal. Human motion synthesis by motion manifold learning and motion primitive segmentation. In *AMDO 2006: Proceedings of the 4th International Conference on Articulated Motion and Deformable Objects*, pages 464–473. Springer, 2006.

[LEAP05] Tian Lan, D. Erdogmus, A. Adami, and M. Pavel. Feature selection by independent component analysis and mutual information maximization in eeg signal classification. In *Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference on*, volume 5, August 2005.

[Leg95] Christopher J. Leggetter. *Improved acoustic modelling for HMMs using linear transformations.* PhD thesis, University of Cambridge, 1995.

[Lew90] Michael Lewis. The development of intentionality and the role of consciousness. *Psychological Inquiry*, 1(3):231–247, 1990.

[LG02] L. Lee and W. E. L. Grimson. Gait analysis for recognition and classification. In *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, 2002*, pages 148–155, 2002.

[LHP05] Karen C. Liu, Aaron Hertzmann, and Zoran Popović. Learning physics-based motion style with nonlinear inverse optimization. *ACM Transactions on Graphics*, 24(3):1071–1081, 2005.

[LN04] Christine L. Lisetti and Fatma Nasoz. Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP Journal on Applied Signal Processing*, 2004:1672–1687, 2004.

[LN06] Fengjun Lv and Ramakant Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *ECCV'06: Proceedings of the 9th European Conference on Computer Vision*, pages 359–372, 2006.

[LN07] Fengjun Lv and R. Nevatia. Single view human action recognition using key pose matching and viterbi path searching. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007.

[LR04] Diane J. Litman and Kate F. Riley. Predicting student emotions in computer-human tutoring dialogues. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 351+, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[LRL07] Yulan Liang, M. L. Reyes, and J. D. Lee. Real-time detection of driver cognitive distraction using support vector machines. *IEEE Transactions on Intelligent Transportation Systems*, 8(2):340–350, 2007.

[LW95] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, 9(2):171–185, April 1995.

[LYG03] Winston Lin, Roman Yangarber, and Ralph Grishman. Bootstrapped learning of semantic classes from positive and negative examples. In *Proceedings of the ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data*, 2003.

[Mat93] David Matsumoto. Ethnic differences in affect intensity, emotion judgments, display rule attitudes, and self-reported emotional expression in an american sample. *Motivation and Emotion*, 17(2):107–123, June 1993.

[mcn92] David mcneill. *Hand and Mind: What Gestures Reveal about Thought*. The University of Chicago Press, 1992.

[Mei89] Marco Meijer. The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal Behavior*, 13(4):247–268, December 1989.

[MK03] Philipp Michel and Rana E. Kaliouby. Real time facial expression recognition in video using support vector machines. In *ICMI '03: Proceedings of the 5th international conference on Multimodal interfaces*, pages 258–264, New York, NY, USA, 2003. ACM.

[MKZA99] Joann Montepare, Elissa Koff, Deborah Zaitchik, and Marilyn Albert. The use of body movements and gestures as cues to emotions in younger and older adults. *Journal of Nonverbal Behavior*, 23(2):133–152, June 1999.

[MOB06] Antonio Micilotta, Eng-Jon Ong, and Richard Bowden. Real-time upper body detection and 3D pose estimation in monoscopic images. In *ECCV 2006: Proceedings of the European Conference on Computer Vision*, pages 139–150, 2006.

[Moz02] S. Mozziconacci. Prosody and emotions. In *Proc. Speech Prosody*, pages 1–9, 2002.

[MP03] Selene Mota and Rosalind W. Picard. Automated posture analysis for detecting learner's interest level. *Computer Vision and Pattern Recognition Workshop*, 5:49+, 2003.

[MPP06] 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. *Behavior Research Methods*, 38(1):134–141, February 2006.

[MR81a] C. Myers and L. Rabiner. Connected digit recognition using a level-building DTW algorithm. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(3):351–363, 1981.

[MR81b] C. Myers and L. Rabiner. Connected word recognition using a level building dynamic time warping algorithm. In *ICASSP '81: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 951–955, 1981.

[MR81c] C. Myers and L. Rabiner. A level building dynamic time warping algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(2):284–297, 1981.

[MS99] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing.* MIT Press, 1999.

[MSSS04] T. Mori, Y. Segawa, M. Shimosaka, and T. Sato. Hierarchical recognition of daily human actions based on continuous hidden markov models. In *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004*, pages 779–784, 2004.

[Mur] Kevin Murphy. Hidden Markov Model (HMM) Toolbox for Matlab.
urlhttp://www.cs.ubc.ca/˜murphyk/Software/HMM/hmm.html.

[NALF04] Fatma Nasoz, Kaye Alvarez, Christinel Lisetti, and Neal Finkelstein. Emotion recognition from physiological signals using wireless sensors for presence technologies. *Cognition, Technology & Work*, 6(1):4–14, February 2004.

[NCN$^+$99] M. S. Nixon, J. N. Carter, J. M. Nash, P. S. Huang, D. Cunado, and S. V. Stevenage. Automatic gait recognition. In *IEE Colloquium on Motion Analysis and Tracking*, pages 3/1–3/6, 1999.

[Ngu02] Patrick Nguyen. *Speaker adaptation: Modelling variabilities.* PhD thesis, Ecole polytechnique fédérale de Lausanne, 2002.

[NSHU$^+$94] H. Ney, V. Steinbiss, R. Haeb-Umbach, B. H. Tran, and U. Essen. An overview of the philips research system for large-vocabulary continuous-speech recognition. *International Journal of Pattern Recognition and Artificial Intelligence, Special Issue on Speech Recognition for Different Languages*, 8(1):33–70, 1994.

[NW70] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, March 1970.

[OCC75] A. Ortony, G. L. Clore, and A. Collins. *The Cognitive Structure of Emotions.* Cambridge Univerity Press, 1975.

[OFC01] Tim Oates, Laura Firoiu, and Paul Cohen. Using dynamic time warping to bootstrap hmm-based clustering of time series. In *Sequence Learning: Paradigms, Algorithms and Applications*, pages 35–52. Springer, 2001.

[OHG02] N. Oliver, E. Horvitz, and A. Garg. Layered representations for human activity recognition. In *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces, 2002*, pages 3–8, 2002.

[OPP09] A. Oikonomopoulos, I. Patras, and M. Pantic. An implicit spatiotemporal shape model for human activity localization and recognition. *Computer Vision and Pattern Recognition Workshop*, 0:27–33, 2009.

[Ost02] Jörn Ostermann. *Face Animation in MPEG-4*, pages 17–56. Wiley Blackwell, 2002.

[Par07] Les Pardew. *Character Emotion in 2D and 3D Animation.* Delmar, 2007.

[Pat02] Helena M. Paterson. *The perception and cognition of emotion from motion.* PhD thesis, Department of Psychology, University of Glasgow, 2002.

[PBL02] M. Pardas, A. Bonafonte, and J. L. Landabaso. Emotion recognition based on MPEG-4 facial animation parameters. In *ICASSP '02: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002*, volume 4, pages IV–3624–IV–3627 vol.4, 2002.

[Pic97] Rosalind W. Picard. *Affective Computing.* The MIT Press, 1997.

[Pol01] Tony P. Polichroniadis. *High Level Control of Virtual Actors.* PhD thesis, Computer Laboratory, University of Cambridge, 2001.

[PP06] M. Pantic and I. Patras. Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 36(2):433–449, March 2006.

[PPBS01] Frank E. Pollick, Helena M. Paterson, Armin Bruderlin, and Anthony J. Sanford. Perceiving affect from arm movement. *Cognition*, 82(2):B51–B61, December 2001.

[PPS01] Helena M. Paterson, Frank E. Pollick, and Anthony J. Sanford. The role of velocity in affect discrimination. In *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, pages 756–761, 2001.

[PR00] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[PR02] M. Pantic and L. J. M. Rothkrantz. Facial gesture recognition in face image sequences: a study on facial gestures typical for speech articulation. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 2002*, volume 6, pages 6 pp. vol.6+, 2002.

[PR03] M. Pantic and L. J. M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

[PRP05] J. Posner, J. A. Russell, and B. S. Peterson. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734, 2005.

[PSCH05] Maja Pantic, Nicu Sebe, Jeffrey F. Cohn, and Thomas Huang. Affective multimodal human-computer interaction. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 669–676, New York, NY, USA, 2005. ACM.

[PVH01] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions in Pattern Analysis and Machine Intelligence*, 23(10):1175–1191, October 2001.

[PVW07] P. Peursum, S. Venkatesh, and G. West. Tracking-as-recognition for articulated full-body human motion analysis. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007.

[Rab02] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, August 2002.

[RCB98] Charles Rose, Michael F. Cohen, and Bobby Bodenheimer. Verbs and adverbs: Multidimensional motion interpolation. *IEEE Compututer Graphics and Applcations*, 18(5):32–40, September 1998.

[RD08] Bogdan Raducanu and Fadi Dornaika. Dynamic vs. static recognition of facial expressions. In *AmI '08: Proceedings of the European Conference on Ambient Intelligence*, pages 13–25, Berlin, Heidelberg, 2008. Springer-Verlag.

[RK90] Elaine Rich and Kevin Knight. *Artificial Intelligence*. McGraw-Hill, 1990.

[RL85] L. Rabiner and S. Levinson. A speaker-independent, syntax-directed, connected word recognition system based on hidden markov models and level building. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(3):561–573, 1985.

[RL05] D. B. Redpath and K. Lebart. Observations on boosting feature selection. *Multiple Classifier Systems*, pages 32–41, 2005.

[RQD00] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41, January 2000.

[SAE07] V. Sethu, E. Ambikairajah, and J. Epps. Speaker normalisation for speech-based emotion detection. In *Proceedings of the 15th International Conference on Digital Signal Processing, 2007*, pages 611–614, 2007.

[SC78] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49, 1978.

[SC01] K. L. Schmidt and J. F. Cohn. Dynamics of facial expression: normative characteristics and individual differences. In *ICME 2001: Proceedings of the IEEE International Conference on Multimedia and Expo, 2001*, pages 547–550, 2001.

[SC07] Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.*, 11(5):561–580, 2007.

[SCGH05] Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S. Huang. Multimodal approaches for emotion recognition: a survey. In Simone Santini, Raimondo Schettini, and Theo Gevers, editors, *Internet Imaging VI: Proceedings of the Society of Photo-Optical Instrumentation Engineers*, volume 5670, pages 56–67. SPIE, 2005.

[Sch00] Klaus R. Scherer. *Psychological Models of Emotion*, chapter 6, pages 137–162. Oxford University Press, 2000.

[Sch05a] Klaus R. Scherer. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729, December 2005.

[Sch05b] Klaus Schittkowski. Optimal parameter selection in support vector machines. *Journal of Industrial and Management Optimization*, 1(4):465–476, 2005.

[SD04] S. Simhon and G. Dudek. Sketch interpretation and refinement using statistical models. In *Eurographics Symposium on Rendering*, 2004.

[Sez06] Tevfik M. Sezgin. *Sketch Interpretation Using Multiscale Stochastic Models of Temporal Patterns*. PhD thesis, Massachussetts Institute of Technology, 2006.

[SKA$^+$06] Schmidt, Karen, Ambadar, Zara, Cohn, Jeffrey, Reed, and L. Movement differences between deliberate and spontaneous facial expressions: Zygomaticus major action in smiling. *Journal of Nonverbal Behavior*, 30(1):37–52, March 2006.

[SKB06] Felix Schaeffler, Vera Kempe, and Sonja Biersack. Comparing vocal parameters in spontaneous and posed child-directed speech. In *Proceedings of the 3rd international conference on speech prosody*, 2006.

[SML86] Craig A. Smith, Gregory J. Mchugo, and John T. Lanzetta. The facial muscle patterning of posed and imagery-induced expressions of emotion by expressive and nonexpressive posers. *Motivation and Emotion*, 10(2):133–157, June 1986.

[SNH98] H. Sawada, T. Notsu, and S. Hashimoto. Japanese sign-language recognition based on gesture primitives using acceleration sensors and datagloves. In *Proccedings of the Second European Conference on Disability, Virtual Reality and Associated Technologies*, pages 149–157, 1998.

[SP95] T. Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. In *Proceedings of the International Symposium on Computer Vision, 1995*, pages 265–270, 1995.

[SP01] Klaus R. Scherer and Martin Peper. *Psychological theories of emotion and neuropsychological research*, volume 5, chapter 2, pages 17–48. Elsevier Science, 2 edition, 2001.

[SS08] Tal Sobol-Shikler. *Analysis of affective expression in speech.* PhD thesis, University of Cambridge, Computer Laboratory, 2008.

[SSH05] Petra Sundström, Anna Ståhl, and Kristina Höök. emoto: affectively involving both body and mind. In *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, pages 2005–2008, New York, NY, USA, 2005. ACM.

[Ste92] Gerhard Stemmler. *Differential Psychophysiology: Persons in Situations.* Springer, 1992.

[SW90] K. R. Scherer and H. G. Wallbott. *Ausdruck von Emotionen*, chapter 6, pages 345–422. Hogrefe, 1990.

[SWMK88] K. R. Scherer, H. G. Wallbott, D. Matsumoto, and T. Kudoh. *Emotional experience in cultural context: A comparison between Europe, Japan and the United States*, chapter 1, pages 5–30. Erlbaum, Hillsdale, New Jersey, 1988.

[TAD00] Laurel J. Trainor, Caren M. Austin, and Renée N. Desjardins. Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychological Science*, 11(3):188–195, 2000.

[TE01] Marion Trew and Tony Everett. *Human movement: An introductory text.* Elsevier, 4 edition, 2001.

[TF00] Joshua B. Tenenbaum and William T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000.

[TFS00] Helen Tager-Flusberg and Kate Sullivan. A componential view of theory of mind: evidence from williams syndrome. *Cognition*, 76(1):59–90, July 2000.

[THG94] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680, November 1994.

[TKC02] Y. I. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(2):97–115, August 2002.

[TLJ07] Yan Tong, Wenhui Liao, and Qiang Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(10):1683–1699, October 2007.

[TM08] Jessica L. Tracy and David Matsumoto. The spontaneous expression of pride and shame: Evidence for biologically innate nonverbal displays. *Proceedings of the National Academy of Sciences*, 105(33):11655–11660, 2008.

[TR02] Michael Thelen and Ellen Riloff. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 214–221, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[TS00] Kurt A. Thoroughman and Reza Shadmehr. Learning of action through adaptive combination of motor primitives. *Nature*, 407(6805):742–747, October 2000.

[TV04] Kinh Tieu and Paul Viola. Boosting image retrieval. *International Journal of Computer Vision*, 56(1):17–36, January 2004.

[TVC07] P. K. Turaga, A. Veeraraghavan, and R. Chellappa. From videos to verbs: Mining videos for activities using a cascade of dynamical systems. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007.

[UF04] R. Urtasun and P. Fua. 3D tracking for gait characterization and recognition. In *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004*, pages 17–22, 2004.

[VA06] Thurid Vogt and Elisabeth André. Improving automatic emotion recognition from speech via gender differentiation. In *LREC 2006: Proceedings of the Language Resources and Evaluation Conference*. ELRA, 2006.

[Vap00] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer, 2 edition, 2000.

[vD00] Stijn van Dongen. *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, 2000.

[VdSRdG07] J. Van den Stock, R. Righart, and B. de Gelder. Body expressions influence recognition of emotions in the face and voice. *Emotion*, 7(3):487–494, August 2007.

[Vec02] Del Vecchio. Primitives for human motion: a dynamical approach. In *Proceedings of the 15th IFAC World Congress on Automatic Control*, 2002.

[VGP07] Michel F. Valstar, Hatice Gunes, and Maja Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces*, pages 38–45, New York, NY, USA, 2007. ACM.

[VGS$^+$06] V. Vinayagamoorthy, M. Gillies, A. Steed, E. Tanguy, X. Pan, C. Loscos, and M. Slater. Building expression into virtual characters. In *Eurographics Conference State of the Art Reports 2006*, 2006.

[VKD05] Stefan Vacek, Steffen Knoop, and Rüdiger Dillmann. Classifying human activities in household environments. In *International Joint Conference on Artificial Intelligence*, 2005.

[VM97] C. Vogler and D. Metaxas. Adapting hidden markov models for asl recognition by using three-dimensional computer vision methods. In *Proceedings of the 1997 IEEE International Conference on Systems, Man, and Cybernetics, 1997*, volume 1, pages 156–161, 1997.

[VM99] C. Vogler and D. Metaxas. Parallel hidden markov models for american sign language recognition. In *Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999*, volume 1, pages 116–122, 1999.

[VM01] Christian Vogler and Dimitris Metaxas. A framework for recognizing the simultaneous aspects of american sign language. *Computer Vision and Image Understanding*, 81(3):358–384, 2001.

[VP06] Michel Valstar and Maja Pantic. Fully automatic facial action unit detection and temporal analysis. In *CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, pages 149+, Washington, DC, USA, 2006. IEEE Computer Society.

[VP07] Michel Valstar and Maja Pantic. Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. In *HCI'07: Proceedings of the IEEE Workshop on Human Computer Interaction*, pages 118–127, 2007.

[VPAC06] Michel F. Valstar, Maja Pantic, Zara Ambadar, and Jeffrey F. Cohn. Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In *ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces*, pages 162–170, New York, NY, USA, 2006. ACM.

[VSM00] C. Vogler, H. Sun, and D. Metaxas. A framework for motion recognition with applications to american sign language and gait recognition. In *Proceedings of the Workshop on Human Motion, 2000*, pages 33–38, 2000.

[VSWR07] Bogdan Vlasenko, Björn Schuller, Andreas Wendemuth, and Gerhard Rigoll. Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing. In *ACII '07: Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction*, pages 139–147, Berlin, Heidelberg, 2007. Springer-Verlag.

[Vuc01] V. Vuckovic. Dynamic time-warping method for isolated speech sequence recognition. In *TELSIKS 2001: Proceedings of the 5th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Service, 2001*, volume 1, pages 257–260, 2001.

[Wal98] Harald G. Wallbott. Bodily expression of emotion. *European Journal of Social Psychology*, 28(6):879–896, 1998.

[WB99] Andrew D. Wilson and Aaron F. Bobick. Parametric hidden markov models for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):884–900, 1999.

[WBR07] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3d exemplars. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–7, October 2007.

[Wek] Weka. Weka data mining software: online resources. `http://www.cs.waikato.ac.nz/ml/weka/`.

[WH99] Ying Wu and Thomas Huang. Vision-based gesture recognition: A review. In *Proceedings of the International GestureWorkshop, GW'99*, pages 103–115. Springer, 1999.

[Whi96] M. Whittle. Clinical gait analysis: A review. *Human Movement Science*, 15(3):369–387, June 1996.

[Whi02] Michael Whittle. *Gait analysis: an introduction.* Elsevier, 3 edition, 2002.

[WLF$^+$09] Jacob Whitehill, Gwen Littlewort, Ian Fasel, Marian Bartlett, and Javier Movellan. Toward practical smile detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2106–2111, 2009.

[WLH07] Tsu-Yu Wu, Chia-Chun Lian, and Jane Y. Hsu. Joint recognition of multiple concurrent activities using factorial conditional random fields. In *2007 AAAI Workshop on Plan, Activity, and Intent Recognition, Technical Report WS-07-09. The AAAI Press, Menlo Park*, 2007.

[WLZ03] Dong Wang, Lie Lu, and Hong J. Zhang. Speech segmentation without speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 468–471, 2003.

[WMH08] Ning Wang, Stacy Marsella, and Tim Hawkins. Individual differences in expressive response: a challenge for ECA design. In *AAMAS '08: Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems*, pages 1289–1292, Richland, SC, 2008. International Foundation for Autonomous Agents and Multiagent Systems.

[WSSH96] Rainer Westermann, Kordelia Spies, Günter Stahl, and Friedrich W. Hesse. Relative effectiveness and validity of mood induction procedures: a meta-analysis. *European Journal of Social Psychology*, 26(4):557–580, 1996.

[Wun05] W. Wundt. *Grundzüge der physiologischen Psychologie.* Engelmann, 5 edition, 1905.

[YKO+00] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book Version 3.0.* Cambridge University Press, 2000.

[You93] S. J. Young. The HTK hidden markov model toolkit: design and philosophy. Technical Report CUED/F-INFENG/TR.152, Department of Engineering, University of Cambridge, 1993.

[YRT89] S. J. Young, N. H. Russell, and J. H. S. Thornton. Token passing: a simple conceptual model for connected speech recognition systems. Technical Report CUED/F-INFENG/TR.38, Department of Engineering, University of Cambridge, Cambridge, UK, 1989.

[ZBC00] Liwei Zhao, Norm I. Badler, and Monica Costa. Interpreting movement manner. *Computer Animation*, 0:98+, 2000.

[ZDlTCZ09] Yunfeng Zhu, Fernando De la Torre, Jeffrey F. Cohn, and Yu-Jin Zhang. Dynamic cascades with bidirectional bootstrapping for spontaneous facial action unit detection. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–8, 2009.

[Zip49] George K. Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology.* Addison-Wesley Press, 1949.

[ZJ05] Yongmian Zhang and Qiang Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(5):699–714, 2005.

[ZLH+79] Miron Zuckerman, Deborah T. Larrance, Judith A. Hall, Richard S. Defrank, and Robert Rosenthal. Posed and spontaneous communication of emotion via facial and vocal cues. *Journal of Personality*, 47(4):712–733, 1979.

[ZPRH07] Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang. A survey of affect recognition methods: audio, visual and spontaneous expressions. In *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces*, pages 126–133, New York, NY, USA, 2007. ACM.

[ZRXL00] Yuanxin Zhu, Haibing Ren, Guangyou Xu, and Xueyin Lin. Toward real-time human-computer interaction with continuous dynamic hand gestures. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000*, pages 544–549, 2000.

[ZTPH08] Zhihong Zeng, Jilin Tu, B. M. Pianfetti, and T. S. Huang. Audio&#x2013;visual affective expression recognition through multistream fused hmm. *Multimedia, IEEE Transactions on*, 10(4):570–577, May 2008.

[ZWFW97] Puming Zhan, Martin Westphal, Michael Finke, and Alex Waibel. Speaker normalization and speaker adaptation - a combination for conversational speech recognition. In *EUROSPEECH-1997*, pages 2087–2090, 1997.