# Manual Interaction:
# Multimodality, Decomposition, Recognition

**Alexandra Barchunova**

PhD Thesis

Bielefeld University, Germany

# Manual Interaction:<br>Multimodality, Decomposition, Recognition

**Dissertation**

**zur Erlangung des akademischen Grades**
**Doktor der Naturwissenschaften**

**der technischen Fakultät der Universität Bielefeld**

**vorgelegt von**
**Alexandra Barchunova**

| | |
|---|---|
| **Erster Gutachter:** | **Prof. Dr. Helge Ritter (Universität Bielefeld)** |
| **Zweiter Gutachter:** | **Prof. Dr. Franz Kummert (Universität Bielefeld)** |

The official print version has been printed on age-resistant paper in accordance with DIN-ISO 9706.

# Acknowledgements

I would like to express my greatest appreciation and gratitude to my research supervisors Dr. Robert Haschke and Prof. Dr. Helge Ritter for their most patient guidance, encouragement and advice throughout this work. They played a decisive role in helping me approach the most puzzling questions.

A focused work on this thesis would not have been possible without the financial and technical support provided by Cor-Lab, Honda Research Institute Europe and the Neuroinformatics group. They provided limitless possibilities for work, cooperation as well as cutting-edge facilities.

I am also very grateful to my external supervisor Dr. Mathias Franzius from Honda Research Institute Europe for his assistance in keeping my progress on schedule. Parent-like advice and encouragement from Prof. Dr. Barbara Hammer and Prof. Dr. Friederike Eyssel have been a great help.

I would like to thank my friends and colleagues, Matthias Behnisch, Carsten Schürmann, Alex Schulz, Flo Schmidt, Slobodan Vukanovic, Jonathan Maycock, Christian Leichsenring and Sebastian Zehe for their encouragement, assistance and patience during the collection of my data. My special thanks are extended to my office colleagues Matthias Schöpfer and Erik Weitnauer, for their helpfulness, patience and flexibility. The latter I additionally thank for taking the bulk of the responsibility of caring for our office plants!

I wish to acknowledge my colleagues Jan Moringen and Ulf Grossekathöfer for fruitful discussions and help with both theoretical and practical issues during this crucial learning period. I am particularly grateful for assistance I received with the technical infrastructure, which importantly includes the coffee-machine, provided by Oliver Lieske and Martin Vorfeld. I thank Ruth Moradbakhti and Susanne Strunk for helping me with all organizational matters.

I would like to thank my family for their support and tolerance of my constant absence. I also wish to thank my extended family including Sigrid and Joachim Bartsch, Annelore Brammer, Bobbie and Sam Francis, Gloria and Peter Grothkopf, Otto and Christa Schmudlach, and Angela and Erhard Stölting, who were supportive and sensitive to my needs and always backed me up in difficult situations. A great thanks to Louise Vasvari who kindly gifted me boutique attire that I wore at my thesis defense and during my time at work.

Finally, a special thanks to the Science Museum, London for allowing me to use their picture for my title page.

To my grandmothers Elena and Nata, and to my great grandmother Natalia
for being role models and hard-working professionals.

# Contents

# Chapter 1

# Introduction

## 1.1 Motivation and Goals

For humans, manual interaction with the surrounding objects and its recognition is an essential cognitive ability. When we observe, how others interact with objects, we usually see continuous movements of the fingers accompanied in some cases by an acoustic noise. Nevertheless, we are capable of splitting these observations into chunks and assigning them to semantic categories, such as “grasping”, “pouring” or “shaking”.

The common challenge of the various scientific disciplines investigating recognition of manual interaction, is a deep understanding and modeling of this human ability. On the one hand, neuroscience and psychology are looking for insights into the embodied and cognitive representation of interaction [38]. On the other hand, humanoid robotics requires a method for replication of human manual interaction and its recognition for applications in human-robot interaction [24]. Vital for both research directions is the human self-relative perception for which the multimodality, encompassing vision, proprioception and hearing, plays a crucial role. Therefore, the aspiration of this work is to establish methods for recognition of manual interaction on a semantic level, incorporating a rich set of modalities comparable to human perception.

In the context of this work, the most general question “how to recognize manual interaction” can be split up in three more specific questions. The first question is: How to conceptualize the recognition of manual interaction? This question is of a great relevance for our work, however we do not propose a conceptual framework of our own. From the very beginning we seize a widely recognized conceptualization, characterized by interaction decomposition. Embracing this approach as a conceptual basis, the second question follows directly: How can an interaction be decomposed into smaller chunks, such as action primitives, to make the recognition of the rich variety of interactions feasible? Examination of this question leads to a search for a definition and a computational model of an action primitive. Due to the crucial role played by multimodality in human perception, the third question is: What is the role played by different perceptual modalities?

During the past decade the theories concerning the first question "how to conceptualize recognition of interaction" seem to converge towards a hierarchical approach, the so called Activity Theory. The roots of this approach can be traced back to Russian psychology in the beginning of the 20th century (a detailed discussion of this issue can be found in Chapter 2). Rather than specifying concrete action categories, this approach envisions abstract interaction decomposition and identification on hierarchically organized semantic levels called "action primitives", "actions" and "activities"[1]. Recent psychological experiments have derived manual action primitives in an action segmentation task [39]. Hemeren et al. argue that decomposition is guided by change in the low-level features, including the velocity-based features of the hand.

Motivated by these findings, our work mainly focusing on the second question, "how to decompose interaction into action primitives", is guided by the goal to detect change within a homogeneous observation flow of an interaction. To this end, we employ a Bayesian segmentation method, accommodating besides the homogeneity characteristics of the action primitives, also their length [27]. Our aspiration is that this decomposition method can serve as a building block for a higher-level modeling and representation of interaction. To support this claim, we present an approach towards unsupervised recognition of the decomposed interaction.

The third posed question is concerned with the role of multiple modalities, such as vision, proprioception, the sense of force, temperature, and texture, for recognition of manual interaction, in particular, the action primitives. Based on the neuroscientific findings, including those of B. Stein and A. Meredith, it can be assumed that multiple modalities are directly integrated for the recognition of manual interaction during self-relative perception or observation of others. Therefore, based on multimodal time series captured with a number of wearable and ambient sensors, we investigate the benefit of multiple modalities for both, decomposition of interaction and identification of action primitives.

The rest of the chapter is structured as follows: Section 1.2 introduces our approach to the multimodal interaction on the data level. Section 1.3 establishes the background of our approach to multimodal interaction recognition, consisting of decomposition into action primitives and their recognition. The final section presents the structure of the thesis.

## 1.2 Multimodal Manual Interaction Data

Historically, the research of interaction developed from speech recognition and analysis of video sequences. The development of unimodal video-based approaches focused on interaction recognition from observations of actions conducted by other individuals, and evolved from heavily constrained to more challenging realistic scenarios (described by e.g. Lopes et al. [58]). Despite the great advances in this field, experience with processing of mono, stereo or multiple view-point recordings has showed some difficulties connected with purely video-based approaches: computational complexity, occlusions, or ambiguity caused by e.g. 3D into 2D projection. Some of these problems can be solved by employing marker-based motion tracking setups capable of recording selected three-dimensional trajectories, e.g. VICON. Nevertheless, the problem of occlusion remains.

Recently developed interaction recognition applications take the occlusion problem into consideration and, therefore, much stronger rely on data captured by a variety of wearable or

[1] Note that "action primitive" corresponds to the lowest level in the hierarchy.

non-visual sensors [87, 65, 82]. Nevertheless, the employed methods further build upon the ones that have been successfully applied in video-based approaches. Importantly, in contrast to the video-based methods, the data acquired with the help of wearable sensors can be captured continuously throughout the interaction, and is therefore gap- and occlusion-free.

Similarly, in our approach to multimodal interaction capture, a combination of different non-visual sensors providing gap-free data for three modalities – audio, hand posture and the applied force – plays a central role. Firstly, human self-relative action perception strongly relies on multiple modalities, such as proprioception and hearing that undoubtedly provide an essential sensory feedback. Secondly, this choice of sensors helps to overcome the above-mentioned limitations of video-based methods.

Altogether, our multisensory data acquisition for the left and the right hands serves as a basis for a novel approach to interaction recognition encompassing both essential aspects of manual interaction, the multimodality and the bimanuality. To our knowledge, bimanual interaction recognition based on the combination of the three above-mentioned modalities has not yet been conducted before. Our experiments are based on a representative set of bimanual action primitives, including grasping, shaking, pouring, and screwing.

## 1.3 Recognition of Manual Interaction Through Action Primitives

Until now, there has been no consensus on how an action primitive should be defined. On a very general level, it is commonly described as the smallest unit of a semantic relevance and characterized by a homogeneity of a space-time trajectory in some configuration space (e.g. [14]). Identification of such homogeneous regions, entailing a semantically relevant decomposition of manual interaction into action primitives, is the focus of our work. The main challenges of this task include the high dimensionality and multimodality of the data, variability of action execution, unknown structure of the interaction. The purpose of the next paragraphs is to address these challenges, and, in the context of the previously employed methods, to derive the requirements on our approach.

The first challenge for the identification of action primitives is the unknown structure of an interaction, i.e. unknown action primitives as well as their number and locations. In previous work, two approaches have been commonly employed for identification of action primitives. The simplest method involves manual segmentation, or augmentation of data with pauses or special moves, that can be easily detected and deleted during the postprocessing. The other common approach involves domain-specific knowledge, such as action primitive templates or segmentation heuristics, that enable identification of the predefined action types. Although this method is well suited for the limited domains, it has a major drawback: it does not scale to situations with previously unknown actions, which are common in everyday scenarios.

In order to overcome the above-mentioned drawbacks, and provide a method that generalizes well for a wide range of actions, we employ an alternative approach based on homogeneity as a central characteristic of an action primitive. Mainly, our choice is motivated by the recent psychological finding, deriving action primitives based on change within the homogeneous interaction flow[2] [39]. Other than a change detection algorithm employed for e.g. fault detection and designed to generate arbitrarily small or large segments, estimation of change for detection of action primitives requires a Bayesian approach. In our work

[2] A similar motivation can be found in e.g. the work of Kohlmorgen et al. [49]

we demonstrate the first application of a Bayesian algorithm for multiple change detection introduced by P. Fearnhead [26, 27] to decomposition of interaction. The algorithm has been previously used for detection of multiple change points in one-dimensional time series, and in multivariate time series [89] (in the following paragraph we outline our extension of Fearnhead's algorithm for applications to multimodal data). Two further characteristics of this method are vital for our approach. Firstly, the generated segmentation is optimal in the sense that a combination of a prior distribution on segmentations and the segment-wise likelihoods is maximized. Secondly, Fearnhead's approach employs marginal likelihood, therefore avoiding a tuning of model parameter (a detailed description of the method can be found in Chapter 4).

The next challenge for identification of action primitives is the integration of multisensory data acquired for both hands from a variety of multimodal hardware devices. Traditionally, modalities are segmented independently from one another and later common borders over all modalities are calculated with the help of a heuristic procedure. With the growing dimensionality and multimodality of bimanual data it becomes more challenging to find such a procedure. In order to address the challenge of multimodal integration, we propose two novel approaches to bimanual multimodal segmentation: a hierarchical approach and a parallel approach. The main characteristic of the *hierarchical approach* is its sequential consideration of individual modalities in a series of segmentation and subsegmentation steps. The *parallel approach* processes all modalities in a single pass, thus finding action primitives with a coherent temporal structure w.r.t. all modalities.

The introduced multimodal segmentation methods are not primarily designed to provide semantically relevant labeling of the generated segments. Nevertheless, they can serve as a building block for higher-level interaction recognition methods, e.g. classification of action primitives. To support this claim, we present a robust procedure for identification and representation of action primitives based on the proposed segmentation methods. We show an example of clustering of action primitives based on ordered means models (OMMs) [36]. The resulting multimodal approach is modular with respect to multiple modalities, has a high generalization capability and can be employed in a wide range of applications, e.g. supporting robot learning from interactions performed by human demonstrators, for processing of financial and sociological time series, or in music, for producing automatized transcriptions.

## 1.4 Structure of the Thesis

The structure of the thesis is as follows: Chapter 2 presents an overview of the current approaches in segmentation and recognition of interaction. We present the relevant concepts and terminology, fields of application, major challenges and solution strategies, which provides the basis for our choice of relevant algorithms.

In Chapter 3 we describe the data acquisition and the experimental scenario. Within this chapter we give a detailed description of the hardware devices, the recorded data, and the acquisition of the ground truth used for evaluation of experimental results for both, segmentation and classification. Finally, we give an overview of the data characteristics for all modalities, motivating our approach to modeling and segmentation.

Chapter 4 is dedicated to the theoretical background of the segmentation framework. The first Section 4.1 introduces the Bayesian segmentation framework by P. Fearnhead.

Integrated into this framework is our choice of models, suitable for detection of structural changes in the recorded data (described in Section 4.2). In the last Section 4.3 we present our two approaches to multimodal segmentation: the hierarchical and the parallel approach. Chapter 5 describes and compares experimental results for unimodal, bimodal, and multimodal segmentation. Furthermore, it presents quantitative evaluations of the influence of central parameters on the segmentation.

Chapter 6 is dedicated to an approach towards higher-level representation. The first two sections of this chapter present the theoretical background of the unsupervised classification approach (Sections 6.1 and 6.2). Furthermore, Section 6.4 presents experimental results. The influence of different modalities on the results of clustering is investigated. Parts of the Chapters 3-6 are based on earlier publications (see [34, 5, 7, 6]).

We conclude this work and discuss its possible implications in Chapter 7.

# Chapter 2

# Conceptual Basis and Related Work

In this chapter we review the relevant findings addressing multimodal interaction recognition. We begin our discussion with the question: How to conceptualize the interaction? Building upon this discussion, we then describe the neuroscientific, the psychological and the computational aspects of the multimodal interaction recognition.

## 2.1 Action Primitive, Action, Activity

When discussing conceptualizations of interaction, most researchers in psychology and cognitive robotics refer to different levels of granularity or complexity (e.g. [85, 13, 51]). These range from recognition of detailed local trajectory chunks to action goals and styles [55]. Hence, in order to solve the complex task of recognizing a large number of interactions, the strategy consists in breaking up an interaction episode into chunks (that may e.g. correspond to subtasks) and then focus on the recognition of these simpler constituents.

This approach can be traced back to the Activity Theory, a conceptual framework that has been developed by a group of Russian psychologists led by S. Rubinstein, L. Vygotsky, A. Luria and A. Leontiev and starting from the beginning of the 20th century. The theory includes principles of *object-orientedness* and the *hierarchical structure* of activity [54]. The principle of object-orientedness proposes that humans' interaction with the world is organized around objects [45]. The second principle states a three-layered hierarchical structure of activity. Activities constitute the topmost level and are composed of actions. Actions are then composed of action primitives on the lowest level[1] (see Figure 2.1). According to this principle, an activity is directed by a *motive*, actions are oriented towards *goals* and action primitives are adjusted to *conditions*. We will return to the discussion of these terms shortly. Activity Theory has been recently described in [15, 83].

Most approaches to interaction recognition in computer science are inspired by the Activity Theory. Surprisingly, until now there is not yet a consensus on the use of terminology in the related literature. A wide variety of terms has been used to describe components of

[1] For consistency reasons, in this work, we use the term "action primitive" instead of "operation".

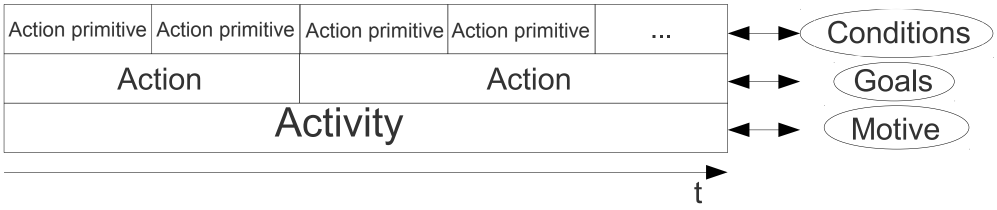


Figure 2.1: Sketch of the hierarchical structure of activity according to the Activity Theory (similar to [45]).

the above-mentioned hierarchy. In 1988 in his work on machine perception of motion Hans Nagel refers to "change, event, verb, episode, history" [63]. During the following twenty years the following terms have also been employed: movements, simple actions, actions, primitives, complex actions, behaviors, activities and many others.

In our work, following the original Activity Theory, we adopt the previously introduced terminology (see Figure 2.1): "action primitives", "actions" and "activities". **Action primitives** are the smallest chunks used in connection with recognition of human movements and involve an approximately homogeneous motion pattern. Bobick describes action primitives as follows[2]: "motion whose execution is consistent and easily characterized by a definite space time trajectory in some configuration space" [13]. Traditionally the meaning of the terms is illustrated on an example of playing tennis [51, 60]. Swinging the tennis racket back to the left, to the right, or up, hitting the ball from the left, right or up could serve as examples for action primitives. Following Bobick we describe an **action** as a "statistical sequence of movements"[3]. "Recognition of such a motion requires knowledge about both the appearance of each constituent movement and the statistical properties of the temporal sequence" [13]. Returning to the tennis example, "back-hand", "forehand" or a "volley" serve as examples for actions. In order to recognize activity "a system has to include a rich knowledge base about the domain and be able to hypothesize and evaluate possible semantic descriptions of the observed motion", writes Bobick in [13]. This notion lies at "a boundary of where perception meets cognition". "Playing tennis" traditionally serves as an example to illustrate **activity**.

Our work is dedicated solely to the lowest level of the activity hierarchy: decomposition of interaction into action primitives and their recognition. The rest of this chapter is organized as follows:

- Section 2.2 presents an overview of the domains employing uni- and multimodal methods for interaction recognition, motivating the direction taken in this work.
- Section 2.3 discusses relevant findings in neuroscience and psychology.
- Section 2.4 outlines the state of the art methods employed for time series segmentation, and motivates the choice of the segmentation method employed in this work.

[2] An action primitive is called "movement" in [13].
[3] An action is called "activity" in [13].

- Section 2.5 discusses the related work for representation and identification of interaction focusing on approaches to interaction recognition based on multiple modalities.
- Section 2.6 summarizes this chapter.

## 2.2 Unimodal and Multimodal Interaction Recognition

Until nowadays, domain-specific interaction recognition has been dominated by unimodal approaches. Hoey et al. describe automatic video-based guidance and prompting of patients to compensate for cognitive disability [40]. The method is illustrated on the example of a set of manual tasks. Analysis of manual interaction with the help of surface electromyography (sEMG) has been used for prosthesis control [18]. Applications based on acceleration are directed towards fitness control in e.g. [12]. Grosshauser et al. describe how the manual pressure recorded with tactile sensors mounted on a violin can be used to detect inaccurate playing, cramping, or malposition [35]. A framework for gesture and sign language recognition and generation is described in e.g. [90].

However, with the development of wearable on-body sensors, such as acceleration, orientation sensors and microphones employed in e.g. millions of mobile devices, multimodal integration has become one of the central issues in interaction analysis. Motivated by context aware collaboration and intelligent information presentation, multimodal integration is conducted for manual and whole-body activity recognition and monitoring in industrial settings (see e.g. [82]). Multimodal integration for activity recognition in absence of static assumptions about sensor availability is the aim of the European Union research project "Opportunity" [73], which is centered around a ten-modality synchronized benchmark data set. Furthermore, in human-computer interaction, Perceptual User Interfaces (PUIs) are being developed to enhance the traditional mouse-keyboard interaction with online multimodal action and activity recognition: "With flexible multimodal interfaces users can take advantage of more than one of their natural communication modes during human-computer interaction, selecting the best mode or combination of modes that suit their situation and task" [21]. For these purposes, a wide range of different multimodal devices are used: "Multimodal systems process two or more combined user input modes - such as speed, pen, touch, manual gestures, gaze and head and body movements - in a coordinated manner with multimedia system output" [43]. Oviatt et al. emphasize: "Our voice, hands, and entire body, once augmented by sensors such as microphones and cameras, are becoming the ultimate transparent and mobile multimodal input device" [67]. Common are speech- and gesture-based solutions, exemplified by e.g. text-input interface by Hoste et al. [41].

Because many of the above-mentioned domains require an interaction recognition method for multiple modalities, the aspiration of this work is towards a generic interaction recognition framework. Due to a big diversity of application domains, and a large number of employed modalities, our work focuses on two essential aspects: 1) integrating a wide range of modalities and 2) scaling to a large number of interactions. Building upon the Activity Theory, and targeting identification of a small representative set of action primitives, constituting a large number interactions, our approach envisions embedding in numerous applications in wearable and pervasive computing, human-computer interaction, etc. For example in humanoid robotics, biologically-inspired generation and recognition of actions aspires to empower a robot to conduct tasks in everyday scenarios and to facilitate human-robot interaction. Hence, in the context of e.g. imitation learning [16], our framework aspires to be employed for learning and identification of interaction chunks within the recorded multimodal interaction time series.

## 2.3 Neuroscientific and Psychological Experiments

For our work in modeling of human manual interaction, the findings of neuroscience and experimental psychology, addressing the questions of biological interaction perception, such as multisensory integration and identification of action primitives, are essential. In the context of our work, the central questions are: How do humans integrate multiple perceptual modalities? How do we decompose interaction and what is the embodied representation of action primitives?

The question of multimodal integration has not played a central role until the beginning of the 90s. Until then research had been mostly conducted with the unimodal approach. King [47] reports: "... studies in sensory physiology have concentrated on the primary neural pathways that encode sensory information in a modality-specific way". In [50] Krebs describes the "old" approach to multisensory integration:

> In this "old" view information is processed initially on a sense-by-sense basis, with each sense processed in a specific part of the cortex – sound in the auditory cortex; touch in the somato-sensory cortex and vision in the visual cortex, then and only then, are the individual fully formed sense perceptions integrated much later in sensory processing [50].

However, the unimodal approach has been challenged by the findings that a modality-specific perception can be directly influenced by other senses. Understanding of the mechanisms by which the brain combines and integrates different sensory sources has become a fundamental issue (e.g. [17, 50]).

For integration of multiple modalities, the superior colliculus and the multisensory neurons found in animal studies and explored at the level of the single cell, play an important role: "The superior colliculus is of a particular interest for the study of multisensory integration because it contains topographically aligned visual, auditory and somatosensory representations, and also because many of the neurons in its deeper layers receive inputs from more than one modality" [47]. Barry Stein, Alex Meredith and their colleagues have conducted a long-term study of superior colliculus and formulated the spatial, temporal and inverse effectiveness principles of neural multisensory integration[4]. In the latest experiments on rats, the multisensory neurons found at the borders between the neighboring modality-specific cortices have demonstrated the ability to integrate a cross-modal input [86]. These findings motivate the early integration of multiple modalities that we pursue in both, the segmentation and the recognition of action primitives (see Chapters 4 and 6 for a detailed discussion).

The current theories concerning the second question, the embodied representation of actions in humans, are commonly based on the mirror system. Briefly, the mirror neurons play a fundamental role in both, the visual recognition and motor execution of certain

[4] The spatial and temporal principles predict that the firing rate of multisensory neurons increases when two or more stimuli of different modalities arise approximately from the same location or at the same time. The inverse effectiveness principle states that the magnitude of multisensory integration inversely depends on the magnitude of the isolated unimodal stimuli [81].

actions [78]. Flanagan et al. suggest that the mirror neurons become activated when an object-oriented goal-directed action is observed, but not when its components are observed [29]. Several theories are concerned with internal modeling of actions and action primitives. Pazzo et al. hypothesize that mirror neurons are a basis for action representation and perception based on internal models [72]. The chain model described by Chersi et al. suggests that local pools of mirror neurons representing action primitives are connected in chains in order to create a complex action [19].

Based on the mirror system, the newly published work by Hemeren et al. proposes to derive the representation of action primitives based on their perception [39]. In their work Hemeren et al. investigate the results of a segmentation task. Participants of the experiment observed object-centered hand and arm actions represented by 12 point-light movies in two scenarios. The results of the experiment are as follows: in both scenarios, where the high-level recognition has been either impaired or not impaired, the participants reliably segmented the actions according to lower-level kinematic variables, such as change of direction, velocity, and acceleration of wrist (thumb and finger tips) [39]. Based on the segmentation of movement kinematics, and the mirror system, coupling perception with possible internal representation, the findings suggest an embodied representation of action primitives as parts of more complex actions. Hemeren et al. also suggest that the results obtained for the impaired and the not impaired scenarios, indicate that both, top-down and bottom-up action perception lead to similar results when performing an unconstrained segmentation task. Inspired by the findings of Hemeren et al. [39], our approach to decomposition of interaction is guided by detection of change, e.g. of low-level velocity-based features of the fingers (see Section 5.3).

## 2.4 Recognition of Interaction Through Decomposition

As described in the previous sections, our work focuses on the decomposition of interaction into action primitives as the first step towards interaction recognition. Inspired by the psychological findings presented in Section 2.3, we pursue this goal by detecting change in the low-level features of time series representing manual interaction. Thus, the following discussion of the state of the art segmentation methods, such as template- and heuristic-based segmentation, focuses on segmentation by change point detection. Before we resume the discussion on the state of the art segmentation approaches, Subsection 2.4.2 briefly introduces the basics of change point detection (CPD), and outlines the main characteristics, essential for usage of a CPD as a first step towards recognition of interaction.

### 2.4.1 Change Point Detection for Recognition

Historically, CPD has been motivated by monitoring of plants, quality control, industrial maintenance and automatic fault detection. Common applications can be found in geophysical signal processing, continuous speech recognition [3], econometric modeling [70], spam filtering and medical diagnostics.

Change detection in time series typically examines whether one or multiple changes have occurred and identifies the corresponding time points, which we will refer to as change points[5]. The goal of a CPD algorithm is, given a finite sequence of observations $y_1, \dots, y_N$,

[5] Accordingly, the data between a pair of adjacent change points is called a “segment”.

to detect an unknown set of change points $\Xi$:

$$\Xi := \{\tau_0, \dots, \tau_k\}, \quad 1 = \tau_0 < \dots < \tau_k = N. \tag{2.1}$$

In some cases the output includes a set of corresponding models $\{\theta_0, \dots, \theta_{k-1}\}$, which may represent the mean, variance, correlation, spectral properties, etc. of the data between the adjacent change points. CPD procedures can be categorized according to three main characteristics: 1) additive vs. nonadditive, 2) Bayesian vs. non-Bayesian, and 3) online vs. offline.

According to the first characteristic, the type of detected change, the CPD algorithms can be divided into two groups: additive changes and nonadditive or spectral changes. Additive changes are “changes in a signal or a linear system that result in changes only in the mean value of the sequence of observations” [10]. Nonadditive changes are “more general and difficult cases where changes occur in the variance, correlations, spectral characteristics, dynamics of the signal or system” [10]. According to the second characteristic, CPD methods are divided into Bayesian and the non-Bayesian approaches [10]. In a Bayesian approach typically a prior distribution of the segment length is assumed. The first Bayesian change detection problem was proposed in the year 1952 [30] to solve an online quality control problem. The first investigation of non-Bayesian change detection algorithms was made in [68]. According to the third characteristic, CPD methods are divided into online and offline algorithms. Depending on the application requirements one of these classes may be more suitable. The online procedures detect change based on sequential hypothesis testing as the data arrives. The offline change detection procedures receive the complete time series at once and are required to output all detected change points.

In order to employ a CPD method as a basis for a recognition application, specific characteristics are particularly important. In [10] Basseville proposes:

> “A possible approach to recognition-oriented signal processing consists of using an automatic segmentation of the signal as the first processing step. A segmentation algorithm splits the signal into homogeneous segments, the lengths of which are adapted to the local characteristics of the analyzed signal. The homogeneity of a segment can be in terms of the mean level or in terms of the spectral characteristics”.

First of all, incorporating prior length modeling in the CPD enables generation of segments whose lengths can be influenced by prior knowledge, therefore an arbitrary length is less probable. Secondly, the usage of an offline method can be beneficial, because an online algorithms that make decisions without the knowledge of the complete time series may be less precise. Finally, model uncertainty in each segment should be allowed to be able to model different types of homogeneities on the lowest level. All these requirements are satisfied by the Bayesian CPD method by Fearnhead [26] discussed in the following section.

### 2.4.2 State of the Art Segmentation Approaches

Major state of the art segmentation approaches are based on the change point detection, templates or data-related segmentation heuristics.

Bayesian multiple change point detection procedures for offline estimation of unknown location and number of change points based on Markov chain Monte Carlo (MCMC) methods have been proposed by Green [32] and used by Punskaya et al. [74]. In contrast to Punskaya, Fearnhead proposes a deterministic method to solve the problem of multiple change point detection in scalar time series with a finite set of models [27]. The advantage of this approach is that it avoids MCMC's problem of diagnosing convergence. Based on Fearnhead's work, Xuan and Murphy propose a procedure for automatic detection of change points in multivariate time series [89]. Employing singular spectrum analysis (SSA) for segmentation of scalar time series has been proposed by Moskvina [61]. All above-mentioned approaches have been previously used on scalar time series or, in the work of Xuan and Murphy, multivariate time series.

Approaches towards detection of change within action sequences with the goal of estimating the borders of action primitives without action-specific knowledge have been researched by many groups. The related work by Kohlmorgen et al. [49] and Kulic et al. [53] is based on the assumption that data belonging to the same action primitive has similar statistical properties. Koenig et al. [48] estimate boundaries of action primitives by means of variance analysis within a sliding window. Takano [84] learns probabilistic correlation and further uses the difference between the predicted and actual feature vectors for detection of boundaries of motion patterns. Ward et al. in [87] segment a sequence of continuous workshop activities analyzing the sound intensity recorded at two different locations and employing a threshold-based method.

A number of segmentation procedures involve action-specific prior knowledge: motion templates, action-specific heuristic description or a sequence-specific heuristic for generation of segment borders. In the first case a set of action primitives is specified a priori [52] and can be then recognized in action sequences. Individual heuristic characterization of action primitives can be found in [91]. In [46] Kawasaki et al. show a multimodal approach to heuristic segmentation on a pick-and-place task, based on several features: object velocity, fingertip position, etc. In order to segment and identify predefined components of the sequence, the authors heuristically specify the corresponding patterns. Ibarguren et al. describe a method for determining of the change points based on a heuristic rule [42]. Here the authors define the segment borders by examining the dynamics of the hand during execution of different gestures of the sign language. In this case low hand activity corresponds to intervals in which the sign is being showed.

The restriction of these methods is the usage of domain-specific knowledge that concerns either the primitives themselves or the heuristics for determining of the segment borders, the change points. Often in the recognition-oriented approaches the problem of automatic segmentation is completely ignored. The time series is either segmented manually or particular action primitives are executed and recorded separately.

Altogether, the interaction segmentation approach proposed in this work aims at high scalability w.r.t multiple modalities as well as a wide range of interaction scenarios. As a most suitable starting point for its development we use the offline Bayesian change point detection based on the method introduced by Fearnhead [27]. To realize the above goals, we extend Fearnhead's method for application in multimodal and bimanual segmentation (see Chapter 4). In contrast to the heuristic-based approaches described in e.g. [48, 53], the proposed approach allows multimodal and simultaneous consideration of various spectral and additive modality-specific properties. Extension to multimodal time series could not be conducted with the SSA-based methods in a similar fashion. In contrast to other template-based approaches e.g. [52, 91, 46], the proposed segmentation method does not employ any action-specific knowledge. The number and locations of action primitives do not need to be specified, although a prior distribution on segment lengths is required. Due to the marginal likelihood approach in the segmentation framework of Fearnhead [26] no model parameters have to be estimated or learned. In Chapter 6 we show that this approach can serve as a building block for a recognition approach.

## 2.5 Recognition of Interaction with Multiple Modalities

Historically, recognition of interaction has been conducted based on video sequences with mono- and multicamera systems, starting with the work by Nagel [62]. Until now the most common applications are e.g. assisting and monitoring systems, surveillance, HCI. Turaga et al. [85], Moeslund et al. [60], Aggarwal et al. [1] and Poppe et al. [71] describe applications in perception of the human actions and activity based on video. A meta survey can be found in a 46-page work by Lopes et al. “Action Recognition in Videos: from Motion Capture Labs to the Web” [58].

Table 2.1: Modeling approaches for action representation.

| Domain | Method | References |
|---|---|---|
| whole-body motion | temporal templates | Bobick et al. [14] |
| whole-body motion | spatio-temporal templates | Gorilick et al. [31] |
| manual manipulations | semantic graphs | Aksoy et al. [2] |
| whole-body motion | semantic attributes | Liu et al. [57] |
| simple whole-body interactions | dynamic Bayesian network | Park et al. [69] |
| simple whole-body movements | prob. context-free grammars | Ogale et al. [64] |
| activity detection | state space models | Cuntoor et al. [22] |

Processing of video sequences is commonly guided by the hierarchical framework described in the beginning of this chapter. On the lowest level feature extractors such as optical flow, point trajectories, blob detection are often used. On the level of action primitive description, different modeling methods are used to represent the captured data. Traditionally modeling is motivated by interpersonal and intrapersonal variance in recorded trajectories, differences in execution velocity, high data volume, etc. Most popular models for representation of action primitives and actions use different variants of dynamic Bayesian networks, manifolds, and linear dynamic systems (e.g. [53, 69, 80]). For activity recognition graphical models, context-free grammars, and logic-based systems have been employed. Modeling techniques associated to different approaches are listed in the Table 2.1.

Since huge advances have been made in the field of wearable sensors, action recognition from video sequences (e.g. [66]) coexists with action and activity spotting and recognition based on different types of markers, ambient and on-body sensors. The methods developed for modeling of interaction based on vision features are further applied to the features extracted from wearable sensor output.

Data recorded by a number of unimodal and multimodal wearable sensors has been used for recognition of manual interaction. An approach based on 24-dimensional joint-angle data has been proposed by Steffen et al. [80]. The authors employ an unsupervised kernel regression (UKR) method for representation of a series of gestures. Matsuo et al. [59] propose a method for learning of action primitives based on tactile feedback of a specially constructed tactile object. The primitives, formed with an EM-based algorithm, are mapped onto a robotic hand to impose appropriate contact forces. In [11] Bernardin et al. describe an approach towards multimodal classification of grasps based on CyberGlove and an array of tactile sensors. This approach uses an offline trained HMM-recogniser as a mechanism of integration of different input modalities. Ogris et al. introduce in [65] a method for multimodal activity spotting, based on motion and force sensors and ultra-wide band (UWB) tags for tracking user position. Their way of dealing with multiple modalities is *masking passes*. The processing by a sequence of masking passes applies a pass, defined by a modality- or feature-specific classifier. After the masking is finished a final merge of the classifier outputs is necessary. Li et al. describe their approach for both, mutlimodal segmentation (Vicon and CyberGlove) and recognition of hand gesture stream by classification [56]: "SVM classifiers with class probability estimates are explored for classifying the feature vectors in order to segment and recognize motion streams."

In our work we show an application of EM-clustering to chunks of multimodal data generated by the proposed segmentation procedures (see Chapter 6). As a method for representation we have chosen to employ a HMM-based model, an established method for representation of dynamic and multimodal data. A specific realization of the model are the ordered means models (OMMs) described in [36]. The absence of transition probabilities distinguishes this model from the HMM and makes it more efficient and robust in classification of incomplete segment data with different execution speeds and sampling rates [34]. OMMs have demonstrated a good generalization capacity in a large number of applications [88, 33, 35]. In Chapter 6 we describe our recognition approach in detail and discuss the influence of different modalities on the classification results.

## 2.6 Summary

Our approach to manual interaction recognition is conceptually based on the Activity Theory, and motivated by neuroscientific and psychological findings, suggesting i) multisensory neurons for early multimodal integration ii) action (or motor) primitives as the smallest structural chunks of embodied and cognitive recognition and representation iii) action primitives detection through change in low-level kinematic velocity-, acceleration- and direction-based features.

A number of segmentation approaches, such as change point detection, templates and heuristics-based methods, have been discussed. Motivated by the findings outlined above, the basis of our approach is a change point detection method that extends an earlier proposal by Fearnhead [26]. Our approach realizes a Bayesian change point detection framework for multimodal bimanual interaction with the goal to improve scalability and to overcome the limitations of the heuristic- and template-based approaches that strongly rely on domain-specific knowledge. Finally, for the purpose of identification of action primitives, a rationale for the use of the ordered means models [36], developed for higher-level modeling of incomplete, multimodal, and dynamic data, has been given.

# Chapter 3

# Experimental Setup and Scenario

In this chapter we present the data pool that serves as a basis for our empirical study of multimodal manual interaction. Based on the acquired data, the main goals that we pursue are: empirical assessment of the decomposition and recognition approaches, as well as the role of multiple modalities for both of the above issues (see Chapters 5 and 6). To realize these goals, on the one hand, we need to choose a representative manual interaction scenario, in which a human demonstrator conducts a sequence of actions on an object. On the other hand, we need suitable sensing devices to capture representative multimodal data.

We target a scenario that consists of a set of actions common in an everyday life. Importantly, most actions on an object are highly multimodal. Firstly, the human hand is involved in numerous perceptual modalities, such as sense of force, temperature, texture and proprioception. Secondly, the object's impact on the environment, as well as the change of its own state during an interaction may comprise multimodal phenomena, such as change of shape, content, etc. Nevertheless, the common approaches towards interaction data acquisition are coarse object- or hand-centered recordings. Often either the object's position and orientation, or the recordings of the position and orientation of the hand are represented by at most two trajectories (see e.g. [37, 20, 4]). In contrast to these approaches, in our work we aspire a comprehensive multimodal capture of manual interaction. In the following text we motivate and discuss the scenario and the captured modalities.

Obtaining ground truth for an interaction episode is a challenging task, and is still an open question. Traditionally, a time-costly manual annotation of interaction is carried out. In our work we propose an alternative time-saving ground truth acquisition method based on interaction triggering audio cues.

The rest of this chapter is structured as follows:

- Section 3.1 introduces the interaction scenario.

- Section 3.2 gives a detailed description of the hardware devices used in our experimental setup.

- Section 3.3 describes the ground truth acquisition, necessary for the quantitative assessment of the experimental results.
- Section 3.4 gives an overview of the relevant characteristics of the acquired data, motivating the methods of preprocessing, segmentation and representation described in the following chapters.
- Section 3.5 summarizes this chapter.

## 3.1 Scenario

As the basis for our empirical study we chose an interaction scenario that is typical and representative for a variety of daily manual actions. It involves a human demonstrator performing uni- and bi-manual actions with a gravel-filled plastic bottle[1] (see Figure 3.2). The scenario is inspired by a daily task of taking a bottle of juice, shaking it, opening it, pouring juice in a glass, and closing the bottle. Human demonstrators are instructed to conduct the following actions:

- pick up the bottle with both hands
- shake the bottle with both hands
- put down and release the bottle
- unscrew cap and release it
- pick up the bottle with the right hand
- pour from the bottle
- put down and release the bottle
- screw cap and release.

Instructions given to the human demonstrators prior to recording of the action sequence can be found in Appendix A.

## 3.2 Hardware Components

A typical manual interaction such as exemplified in our scenario, can be considered from at least four different perspectives:

1. self-relative perception of the hands of the interacting person
2. state of the interaction object
3. external observation of the interaction
4. interaction trigger (optional).

To acquire information corresponding to all four items requires to choose suitable sensing devices. In the following subsections we discuss our approach to their recording, whereby one subsection is dedicated to each of the four above items.

[1] The use of gravel filling instead of liquid is due to safety concerns.

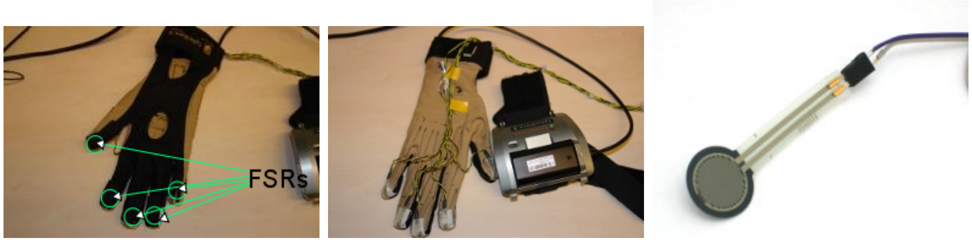


Figure 3.1: Proprioceptive hand sensors. Left: Immersion CyberGlove II with FSR sensors attached to the fingertips. FSR sensors are covered with a layer of foam to achieve a better distribution of force. Middle: Immersion CyberGlove II from the back. The cables connecting the FSR sensors with the Bluetooth communication module are mounted on the back of the hand to avoid movement restrictions during object manipulation. Right: FSR-402 sensor used on each fingertip for capturing of pressure.

### 3.2.1 Hand Sensors

Although temperature and texture might play a prominent role in a small number of manual interactions, proprioception is essential to most manual interactions. Therefore, in our setup, recording the dynamics of the hands during an interaction employs a number of proprioceptive sensors for capturing of the joint-angles of the fingers, the palm as well as pressure measurements at the fingertips of both hands.

High levels of finger activity are typical for a large set of actions associated with reaching, grasping, releasing, screwing or unscrewing movements, etc. In our scenario the finger-specific joint-angle trajectories of the hand are recorded by 22 proprietary resistive bend-sensing joint-angle sensors of the Immersion CyberGlove II [23]. A CyberGlove has three flexion sensors per finger, four abduction sensors, a palm-arch sensor, and sensors to measure wrist flexion and abduction.

Tactile sensor output plays an important role, because it characterizes most actions on an object, reflecting object's weight, orientation, as well as the grasping force, type of the grasp and the type of manipulation. The tactile pressure is measured with ten FSR-402 sensors, one sensor per fingertip. The resistance of the sensor changes when pressure is applied. Five sensors recording data for one hand are connected together to a micro-controller equipped device employing a bluetooth module for wireless communication. This device is referred to as “iHand” in the following.

The sensor setup for one hand is illustrated in Figure 3.1. The described array of sensors allows a finger-specific tracking of applied force and finger movements characterizing a large range of actions on objects.

### 3.2.2 Object Sensor

Analysis of audio signals in the past was commonly applied to language processing. However, it has recently become a source of information for activity recognition, e.g. [87]. Numerous interactions with an object are accompanied and characterized by sound. It carries a large amount of information about the object itself as well as the action that is performed with it.

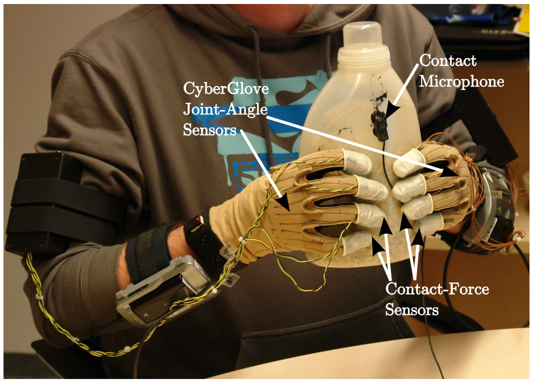


Figure 3.2: View from the camera: plastic bottle instrumented with a contact microphone; human demonstrator wearing CyberGloves with tactile sensors.

Examples are sounds coming from the kitchen that accompany cutting, shaking, stirring, cutlery scratching the plate, placing of the dishes on the table. Apart from the information about the kind of object we are interacting with, in many cases a human is able to recognize the action being conducted or even the person, conducting it. Such observations motivate the recording of the audio signal accompanying the interaction in order to obtain information about the state of the object as well as capture its impact on the surroundings (e.g. the table surface).

In our scenario the audio signal is recorded by a contact microphone mounted on the object of the interaction. The advantage of such an arrangement is that the sensor records only audio signal produced within the object or on its surface. It captures contact establishment with the hand, structural change (i.e. turning of the cap, opening of the lid) and interaction with its environment (i.e. during pushing). At the same time, the microphone filters out most of the sound coming from the environment and not associated with the object itself, like speech or environmental noise. In our setup we use the microphone AKG C411 L typically used for recording of music instruments.

### 3.2.3 External Setup View

To obtain reliable ground truth data, we use a video camera that records the human demonstrator conducting actions on the test object. In our experiments we use a Logitech Quick Cam Pro 900. The video and audio material recorded by the camera is used for manual annotation (see Subsection 3.3.1). Figure 3.2 shows an example view from the camera during a recording session.

### 3.2.4 Interaction Trigger

The interaction trigger component as part of the ground truth acquisition is described in detail in the following Subsection 3.3.2.

### 3.2.5 Overview of the Hardware Components

An overview of all above-mentioned hardware devices sorted accorded to the representing components – the self perception of the hand, the object, external observations, and interaction trigger - is presented in Table 3.1.

Table 3.1: Overview of hardware devices used during trial recording. The recording of the hand and the object are used for the algorithmic analysis of the interaction, segmentation and representation. Camera output and the recorded audio cue schedule serve as input for the generation of ground truth. All components are synchronized based on the respective timestamp logs.

| Name | Dimension | Frequency [Hz] | Component | Modality |
|---|---|---:|---|---|
| 2 × CyberGlove | 2 × 24 | 100 | self perception | joint-angles |
| 2 × iHand | 2 × 5 | 100 | self perception | tactile |
| contact microphone | 1 | 44100 | object state | audio |
| camera | 320 × 240 | 30 | external observation | video+audio |
| headphones+cues | 1 | | interaction trigger | timestamps |

The captured data used for algorithmic processing of the interaction, segmentation and recognition, is as follows (corresponding modality names appear in parentheses):

- microphone attached to the bottle (`a`).
- 2 × 24 joint-angles (`j`: both hands, `jl`: left hand, `jr`: right hand).
- 2 × 5 FSR pressure sensors attached to the fingertips of each CyberGlove (`t`: both hands, `tl`: left hand, `tr`: right hand).

This collection of sensors yields a 29-dimensional (24 + 5) representation for each hand in addition to a scalar audio signal.

## 3.3 Ground Truth Acquisition

To evaluate the algorithms that will be presented in the following chapters, we require reliable ground truth data.

To obtain such data, we have considered two methods: manual *annotation* of the sequences (see Subsection 3.3.1) and *automated cue-based* ground truth acquisition (see Subsection 3.3.2). The first method is used traditionally and involves manual annotation or hand labeling of the video recording of the executed sequence. The biggest disadvantage of this method is that it is time-costly. Therefore, we also propose the automatized cue-based method of ground truth acquisition.

The ground truth for an interaction is represented by a labeled segmentation, where timestamps of the beginning and the end of segments mark the action primitives, and labels indicate the respective type of an action primitive. The boundaries of action primitives are usually fuzzy and cannot be estimated precisely, therefore both methods of ground truth acquisition yield just an approximate description of the interaction structure.

In cases, when the trials are recorded with the audio-cue interaction triggers, we refer to them as **constrained**, and **unconstrained** otherwise. A detailed description of each type can be found in the following two subsections. We discuss the evaluation of segmentation with both methods in Section 5.4.

### 3.3.1 Manual Annotation

Manual annotation of action sequences is based on video and audio data obtained from a simple camera recording of the hands and the object during the interaction (see Subsection 3.2.3). For annotation we have designed a code book consisting of

- three modality-specific label collections: audio, joints-angles and tactiles,
- semantic label collection: the union of the modality-specific labels,
- cue label collection: annotation marking the same events as the automatized cue schedule (see Subsection 3.3.2).

In our work we consider annotation made only by a single annotator. Preliminary experiments have showed a good consistency of annotations made by three different annotators. This can be explained by the simple nature of annotation rules, marking i.e. “object contact” vs. “no object contact” regions (see Appendix B), and a clear and unambiguous content of the data (in contrast to data in e.g. [76]). Detailed descriptions of annotations for all label collections can be found in Appendix B.

### 3.3.2 Automated Annotation using Audio Cues

An automated method for ground truth acquisition that has been used in our experiments as an alternative to manual annotation, is interaction triggering audio cues. Such cues, similar to beep tones, are provided to the subject by headphones and indicate the beginning or the end of a particular action execution. The subject is asked to align her or his action execution with these cues. The usage of headphones excludes the cue signals from the sound recording by the microphones.

Each cue consists of a sequence of four beep sounds[2]: the first three are preparatory and allow the subject to anticipate the fourth signal (*main cue*) which notifies the associated event (beginning or end of action execution) to the subject. We write $c^{\alpha}_{i,j}$, $j \in \{1, 2, 3, 4\}$ to denote the point in time at which the $j$-th signal of the $i$-th cue is emitted in trial $\alpha$. We omit the superscript $\alpha$ if the trial is not important. Figure 3.3 illustrates the structure of the cues.

Audio cues are an automatization of ground truth acquisition with several disadvantages. They constrain the velocity of the execution to the time interval provided by the cues. Furthermore, the cues can provide only a partial description of the structure, if action primitives follow rapidly one after the other. Mistakes made by the human subject in the alignment or the action execution add up as temporal and structural errors in the statistical evaluation of segmentation.

[2] Similar to the structure of beep tones before the time announcement in Deutschland Funk.

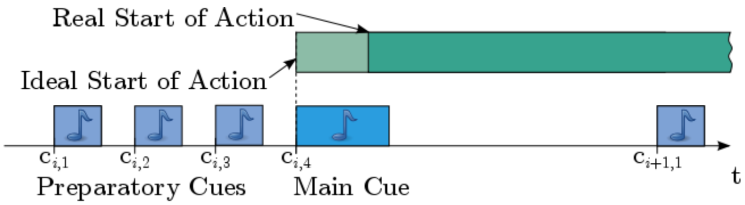


Figure 3.3: Example of temporal relations between cues and the actual action execution. The execution of an action by the subject is expected to start ( light green bar ) at the beginning of the main cue signal $c_{i,4}$, but the actual beginning of the execution usually deviates ( dark green bar ).

To achieve a rich variance of timing between trials, the desired duration of most action primitives was varied by superimposing Gaussian random variables $\eta_i \sim \mathcal{N}(0, 0.5\ s)$ on the mean values of the following duration variables as specified in the parentheses:

- pick up and hold the bottle with both hands (2 s + $\eta_1$)
- shake the bottle with both hands (0.7 s + $\eta_2$ )
- hold the bottle with both hands (0.3 s + $\eta_3$)
- put down the bottle, release and pause (1 s + $\eta_4$)
- unscrew the cap with both hands (1.2 s + $\eta_5$)
- release and pause (1 s + $\eta_6$)
- grasp and lift the bottle with right hand (2 s + $\eta_7$ )
- pour with right hand (1 s + $\eta_8$ + 1 s + $\eta_9$ )
- hold the bottle (0.3 s + $\eta_{10}$)
- put down the bottle, release and pause (1 s + $\eta_{11}$)
- screw the cap with both hands (1.2 s + $\eta_{12}$ )

The overall length of the time series of a trial accumulates on average to approximately 30 seconds.

## 3.4 Properties of Recorded Data

This section presents several preliminary experiments investigating the properties of the recorded data. The principal aim of the experiments is to explore, what kind of variability as well as invariance is characteristic for the acquired interaction data. Basis for conducting of such analysis is a set of action primitives identified according to the ground truth data[3].

[3] Here we use the previously described semantic label collection acquired by manual annotation (see Appendix B).

An example of multimodal trial time series, containing raw tactile, joint-angles and audio data, along with high-lighted action primitives, is presented in Figure 3.4. The first two rows present tactile time-series for each finger of the right and the left hands respectively. The third and the fourth row of Figure 3.4 show joint-angles data for both hands recorded during the same interaction sequence. The bottom row of Figure 3.4 shows the raw audio signal. The audio modality contains the sound that accompanies the actions on object, captured by the contact microphone.

We assume that the recorded data is influenced by many factors simultaneously, such as weight and orientation of the object, the way of grasping, the texture of the object surface, the degree of rigidness of the object, its velocity, shape, and the conducted action, etc. We subdivide the above-mentioned factors into three main categories: object-, human demonstrator-, and action-related. We argue that these factors make the absolute value of the output difficult to interpret.

Because the scenario is focused on a single target object, in the following sections we only investigate action- and human demonstrator-related variability. After introducing the mean and the variance calculation designed for comparing multidimensional time series representing action primitives, we illustrate and discuss different types of variability of the recorded interaction time series.

### 3.4.1 Mean and Variance Measures

In this paragraph we introduce the calculation of specific mean and variance measures, enabling us to build averages over subsequences corresponding to action primitives. These measures serve solely the illustration of the properties of the action primitives presented in the following paragraphs.

For a given modality-specific time series segment $y^{\alpha}_{s:t|\texttt{mod}}$, $s < t$, representing an action primitive $i$ in the trial $\alpha$, let $\mu^{\alpha}_{i}$ be the mean value over all modality dimensions $j$:

$$\mu^{\alpha}_{i} = 1/d \sum_{j=1}^{d} \mu^{\alpha}_{i,j}. \tag{3.1}$$

Here $d$ denotes the dimensionality of the modality $\texttt{mod}$ and $\mu^{\alpha}_{i,j}$ denotes the mean of data within action primitive $i$, i.e. $y^{\alpha}_{s:t|\texttt{mod}}$ in the dimension $j$. The variance of an action primitive $i$ in a trial $\alpha$ is calculated analogously:

$$\sigma^{\alpha}_{i} = 1/d \sum_{j=1}^{d} \sigma^{\alpha}_{i,j}. \tag{3.2}$$

For notational convenience we leave out the square in all $\sigma$ terms. Let further $mean(\mu^{\alpha}_{i})$ denote the mean of $\mu^{\alpha}_{i}$ over all trials $\alpha \in A$, with the corresponding variance $var(\mu^{\alpha}_{i})$. $mean(\sigma^{\alpha}_{i})$ and $var(\sigma^{\alpha}_{i})$ denote the mean value of $\sigma^{\alpha}_{i}$ and its variance over all trials $\alpha \in A$. Based on these action-specific mean and variance measures, in the following subsections we illustrate and discuss modality-, action-specific, inter- and intrapersonal variability. In the final paragraph we compare the variability of cue-based constrained vs. unconstrained trials.

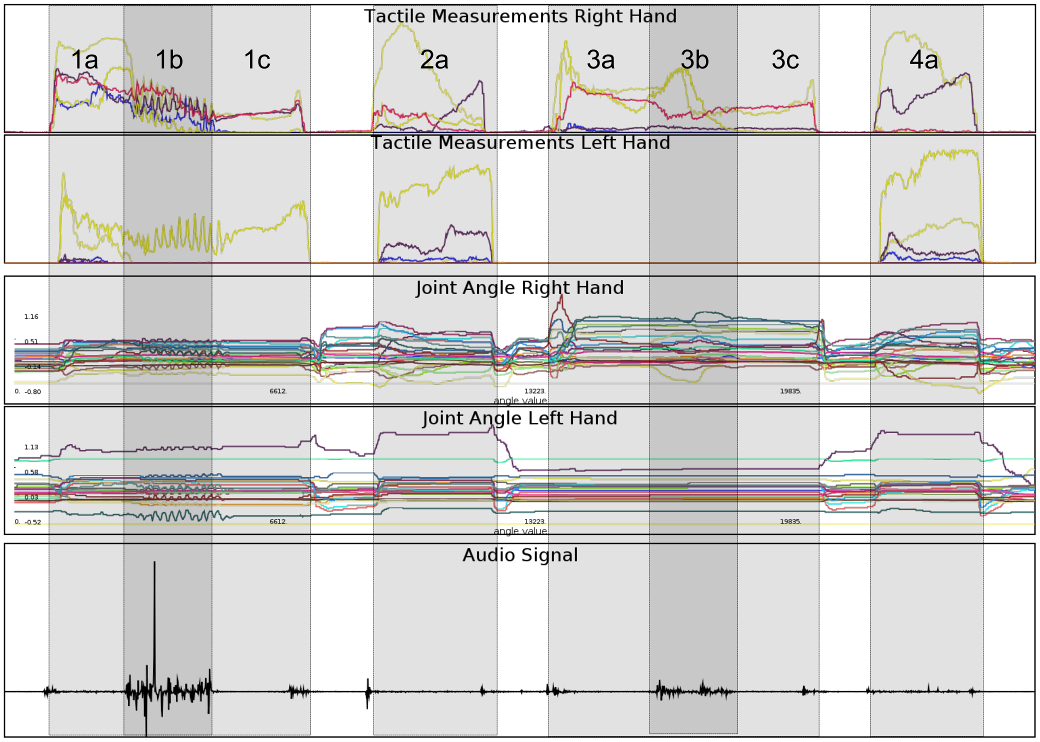


Figure 3.4: An example trial recording of raw multimodal time series. First two rows present tactile data for both hands. **First row**: five recorded measurements of the right-hand tactile sensors. Four non-zero regions 1-4 correspond to the regions of hand-object contact. **Second row**: five recorded measurements of the left-hand tactile sensors. Both subplots clearly show a difference in quality of the recorded time series for different fingers. This is due to occurrences of sensor slipping during the interaction. The third and the fourth rows present an example of joint-angle data for both hands: the right hand (**third row**) and the left hand (**fourth row**). The first change in the values of the angles corresponds to *grasping* (region 1a), after a pause follows *shaking* (1b) that can be recognized by a oscillating signal structure in most dimensions. After another pause the object is released (end of 1c). The second half of the plot starts with the right-handed activity (region 3), while the left hand remains static. After *grasping* (3a) and an idle phase, the following dynamics in the right hand correspond to *pouring* (3b) followed by *putting down* and a release (end of 3c). After a pause the right hand conducts screwing of the cap (4a), marked by dynamics in almost all dimensions, while the left hand remains idle after the grasp (throughout 4a). The **fifth row** shows the contact microphone signal. The regions corresponding to grasping (1a, 3a), shaking (1b), putting down (1c, 3c), screwing (2a, 4a) and pouring (3b) can be clearly differentiated from the idle regions.

### 3.4.2 Action-specific Variability

The purpose of this subsection is an exploratory study of the degree of variability for individual action primitives constrained to a particular modality. Based on the mean and variance values, we are looking for modality- and action-specific characteristics, as well as patterns within the whole interaction episode that may be employed in order to detect change in the respective modality (see Chapter 4).

Figures 3.5-3.7 demonstrate modality-specific values of mean and variance for each action primitive. Here, each modality – tactile, audio and joint-angles – is represented by one figure with two sub-figures. The top row of each figure illustrates modality-specific $mean(\mu_i^\alpha)$ and $var(\mu_i^\alpha)$. The bottom row illustrates modality-specific $mean(\sigma_i^\alpha)$ and $var(\sigma_i^\alpha)$. Averages are built over 40 constrained trials captured by three human demonstrators. Note that the $x$-axis in each figure is labeled according to the names in the semantic label collection. The description of each label is given in Appendix B.

Figure 3.5 illustrates the action-specific mean and variance values for tactile data $y_{|\mathtt{t}}$. The top sub-figure reflects the mean pressure that has been applied to the object during a particular action primitive, the bottom sub-figure demonstrates the corresponding variance. The figure shows high levels of mean for the action primitives associated with "object contact".

Figure 3.6 depicts the mean and variance of the action primitives calculated for the joint-angles data $y_{|\mathtt{j}}$. The top sub-figure shows regions of approximately constant mean and variance levels that approximately correspond to the action primitives within the "object contact" regions, in which the hand configuration does not change once the object has been grasped. The bottom sub-figure demonstrates the highest variance for the action primitives corresponding to high level of finger activity, e.g. *screwing, grasping, releasing.*

Figure 3.7 depicts the mean and variance of the action primitives w.r.t. audio data $y_{|\mathtt{a}}$. The top sub-figure shows high mean values and variance for loud actions[4]. The bottom sub-figure distinguishes similarly to the top sub-figure the loud actions, such as *shaking* and *putting down.* As expected, the silent action primitives, i.e. *lifting*, *releasing* or *holding* are characterized by low levels of mean and variance in both sub-figures.

Altogether the figures suggest that, based on the tactile modality, the action primitives can be assigned to "object contact" vs. "no object contact" categories. Based on the joint-angles, the action primitives can be characterized by high and low overall finger activity. Based on audio, the action primitives can be differentiated according to their loudness.

### 3.4.3 Inter- and Intrapersonal Variability

The goal of this section is to illustrate how action execution differs among human demonstrators (interpersonal variability) and how it differs among trials of the same human demonstrator (intrapersonal variability). Figures 3.8-3.10 compare action-specific means and variances for three human demonstrators. For each human demonstrator, $\mathtt{hd}_1, \mathtt{hd}_2$, and $\mathtt{hd}_3$, the averages are built over 10 constrained trials.

Figure 3.8 illustrates the intrapersonal mean and variance of the tactile modality for each of three human demonstrators. The top sub-figure shows that the level of force application is HD-specific for most action primitives. At the same time the plot demonstrates similar

[4] Action primitives associated with high levels of accompanying noise are referred to as "loud actions". Analogously, we refer to actions that are not accompanied by sound, as silent.

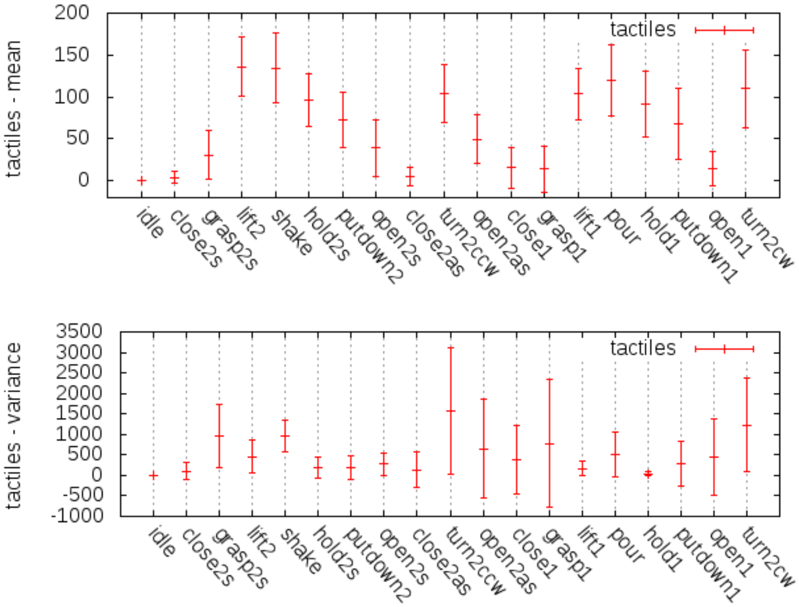


Figure 3.5: Action-specific variability for the tactile modality. Action-specific mean and variance are built over 40 trials of three human demonstrators. Top row: $mean(\mu_i^\alpha)$ and $var(\mu_i^\alpha)$. Between each *grasping* (labeled by e.g. *close2s, close2as, close1*) and each respective *releasing* (labeled by e.g. *open2s, open2as, open1*), we observe that the mean first goes up, reaches a maximum and then gradually falls. Action primitives like *lifting*, *pouring*, *shaking*, or *screwing* (labeled by *lift2, shake, turn2ccw, turn2cw, pour*) are characterized by the highest application of force. Close to zero are the action primitives like *idle*, *grasping* and *releasing* (labeled by e.g. *idle, close2s, close2as, open1*) that take place before and after the object interaction. Bottom row: $mean(\sigma_i^\alpha)$ and $var(\sigma_i^\alpha)$. The highest values mean and variance are reached by e.g. *screwing* and *unscrewing* (labeled by *turn2ccw, turn2cw*), action primitives that are characterized by the largest range of tactile values.

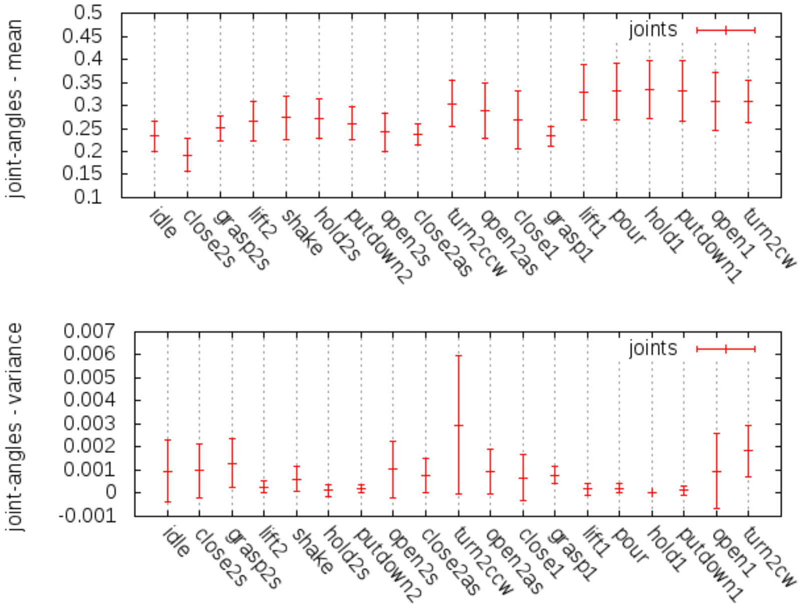


Figure 3.6: Action-specific variability for the joint-angles modality. Action-specific mean and variance are built over 40 trials of three human demonstrators. Top row: $mean(\mu_i^\alpha)$ and $var(\mu_i^\alpha)$. Constant levels of the mean and of the corresponding variance are associated with the regions of "object contact", during which the fingers have a constant configuration, e.g. *lift2, shake, hold2s, putdown2* and *lift1, pour, hold1, putdown1*. Bottom row: $mean(\sigma_i^\alpha)$ and $var(\sigma_i^\alpha)$. High level of variance and its variance is associated with the regions between the object contact regions, where no constant grasp is established: *screwing, unscrewing, releasing, grasping* (labeled by e.g. *close2s, grasp2s, open2s, close2as, turn2ccw, open2as, close1, open1*). *Unscrewing*, labeled by *turn2ccw*, demonstrates the highest mean value of the variance and its variance, implying the most variable styles of execution of this action.

patterns within the trial among all three HDs, e.g. the highest force application during *shaking, screwing, pouring* that gradually decreases in the following action primitives. This indicates that on average the human demonstrators apply individual levels of force with a similar trend throughout the sequence. The variance of the mean values varies strongly among the human demonstrators. The bottom sub-figure shows that the mean levels of variance are also specific for each of the three human demonstrators.

Figure 3.9 presents the intrapersonal mean and variance for three human demonstrators based on the joint-angles data. The top sub-figure shows, similar to the tactile modality, that means and variances differ among the human demonstrators, but the sequential patterns within the trial are comparable among the human demonstrators. The bottom sub-figure shows that the levels of variance differ among the human demonstrators.

For each action primitive Figure 3.10 shows the intrapersonal mean and variance levels of the audio signal for three human demonstrators. Both presented sub-figures demon-

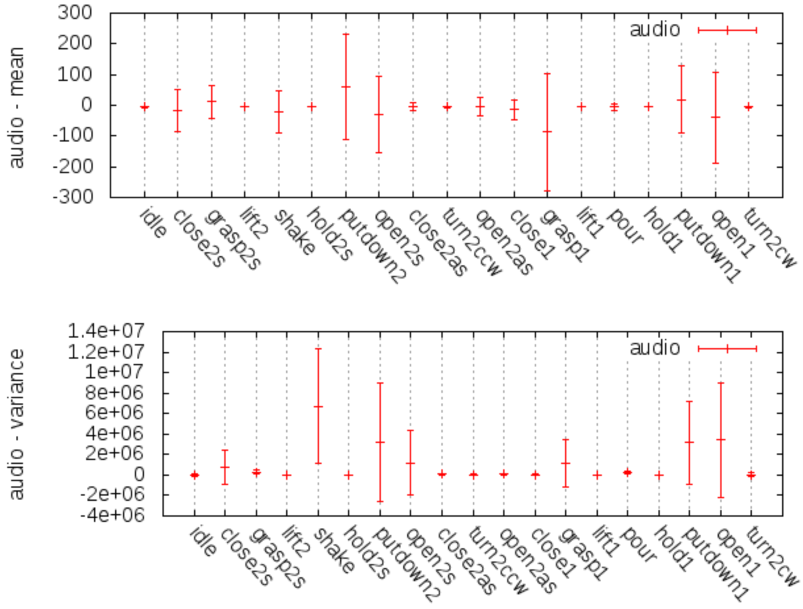


Figure 3.7: Action-specific variability for the audio modality. Action-specific mean and variance are built over 40 trials of three human demonstrators. Top row: $mean(\mu_i^\alpha)$ and $var(\mu_i^\alpha)$. Loud actions are characterized by a large variance of the mean values, e.g. *grasp1* and *put down*. Bottom row: $mean(\sigma_i^\alpha)$ and $var(\sigma_i^\alpha)$. Loud actions are characterized by high mean variance and its variance over trials.

strate different levels of mean and variance for the loud action primitives among the human demonstrators.

All three presented figures showed that execution of an action primitive for each human demonstrator may, to a large extent, be characterized by an individual level of intrapersonal variability. However, similar temporal patterns within some parts of the interaction episode (e.g. increase or decrease of the mean value) could be observed, especially for the joint-angles and the tactile modalities. A tentative hypothesis to explain the interpersonal difference are such factors as fitness, tiredness, or differences in the physiology of the hand among the human demonstrators.

### 3.4.4 Constrained vs. Unconstrained Trials

We have assumed that the data acquired in constrained and unconstrained scenarios (see Section 3.3) does not exhibit strong differences w.r.t. individual action primitives. In this subsection we compare the means and their variance for action primitives recorded in the constrained and the unconstrained scenarios. The averages for each case are built over 40 trials recorded by three human demonstrators.

Figure 3.11 contains a comparison of the modality-specific $mean(\mu_i^\alpha)$ and $var(\mu_i^\alpha)$ for

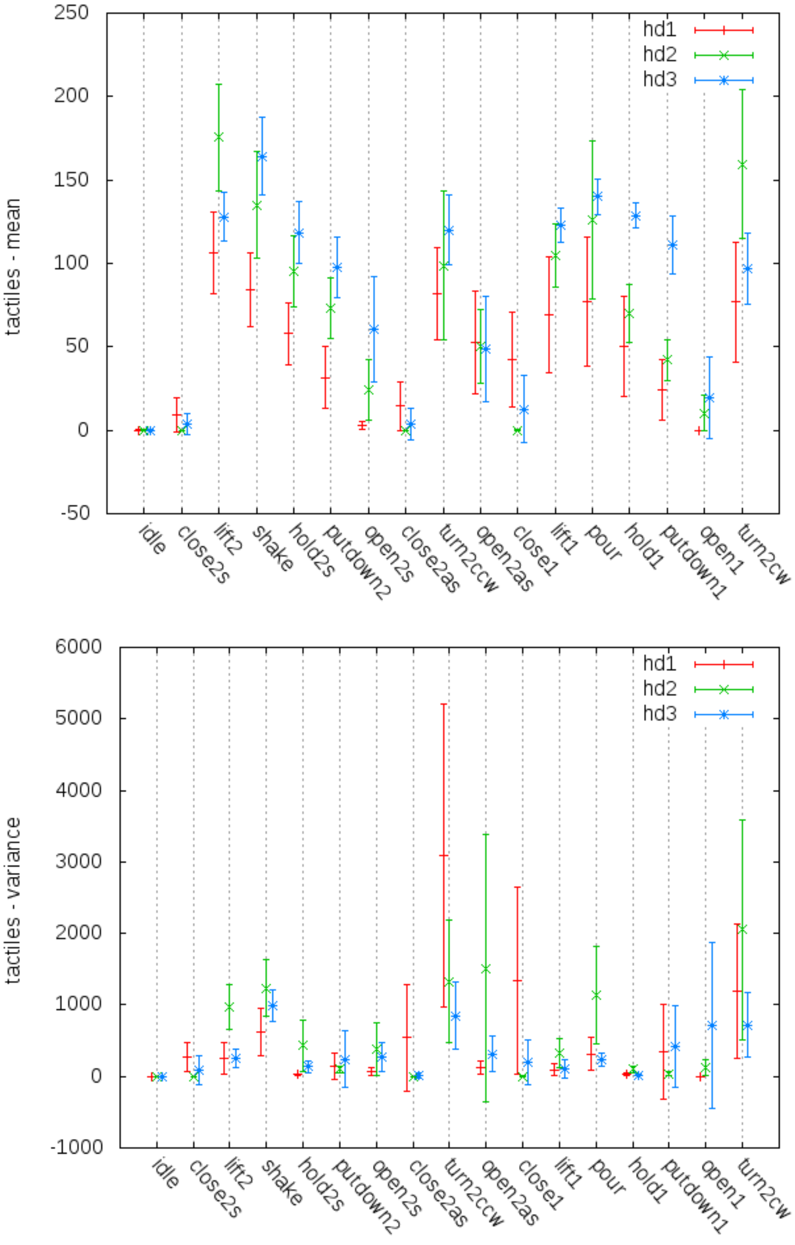


Figure 3.8: Comparison of action-specific mean and variance of the tactile signal for three human demonstrators $\texttt{hd}_1$ (red), $\texttt{hd}_2$ (green), and $\texttt{hd}_3$ (blue). Top row: $mean(\mu_i^\alpha)$ and $var(\mu_i^\alpha)$ for each human demonstrator. The top figure shows that the average level of force and its variance is the lowest for the $\texttt{hd}_1$ and the highest for $\texttt{hd}_3$ for almost all action primitives. We assume that the high mean value demonstrated by $\texttt{hd}_3$ is due to the fact that this individual is involved in extreme rock climbing. However, similar sequential patterns of the mean values can be observed among HDs. Bottom row: $mean(\sigma_i^\alpha)$ and $var(\sigma_i^\alpha)$ for each human demonstrator. $\texttt{hd}_3$ shows the smallest average levels of variance and its variance, implying the consistency of applied force over trials.

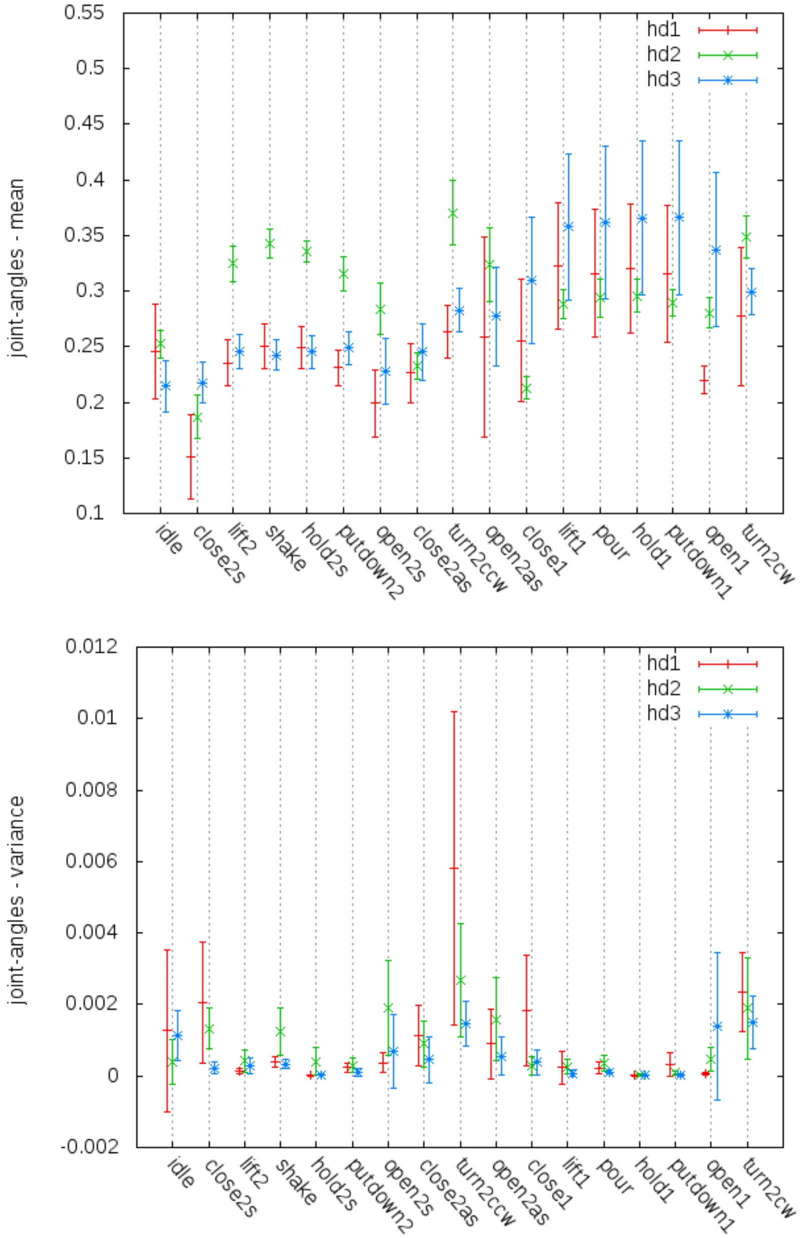


Figure 3.9: Comparison of action-specific mean and variance of the joint-angles signal for three human demonstrators $\texttt{hd}_1$ (red), $\texttt{hd}_2$ (green), and $\texttt{hd}_3$ (blue). Top row: $mean(\mu_i^\alpha)$ and $var(\mu_i^\alpha)$ for each human demonstrator. In the top sub-figure the plot shows similar dynamics of the mean values for all three HDs. Bottom row: $mean(\sigma_i^\alpha)$ and $var(\sigma_i^\alpha)$ for each human demonstrator. $\texttt{hd}_3$ shows the lowest average variance, $\texttt{hd}_2$ and $\texttt{hd}_1$ are comparably high. The variance is comparably high in the regions between "object contact", e.g. during *grasping, releasing* labeled by *close2s, open2s, close2as, open2as, open1, close1*. The variance is especially high during action primitives, such as *screwing* or *unscrewing*.

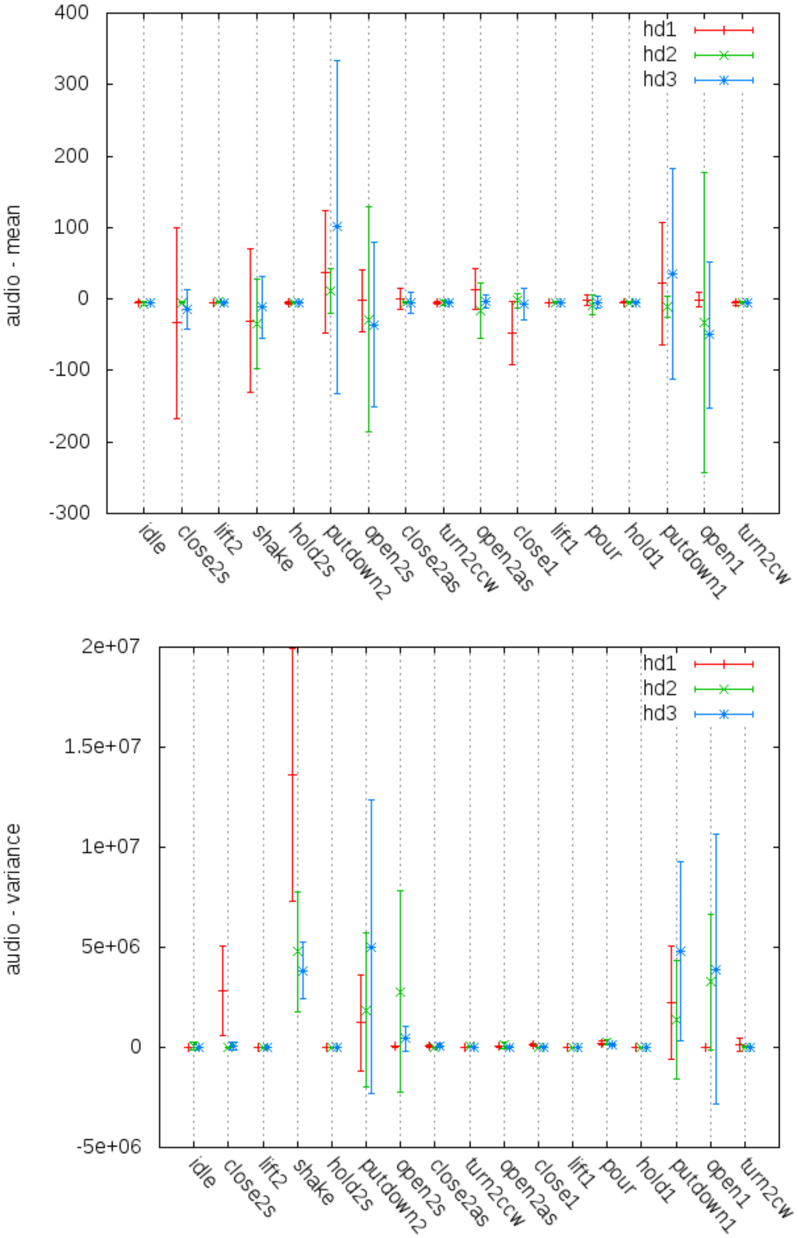


Figure 3.10: Comparison of action-specific mean and variance of the audio signal for three human demonstrators $\texttt{hd}_1$ (red), $\texttt{hd}_2$ (green), and $\texttt{hd}_3$ (blue). Top row: $mean(\mu_i^\alpha)$ and $var(\mu_i^\alpha)$ for each human demonstrator. Loud action primitives, like *grasping*, *shaking*, *putting down*, *releasing* labeled by *close2s, shake, putdown2, open2s, putdown1, open1* have high variance for all human demonstrators. Bottom row: $mean(\sigma_i^\alpha)$ and $var(\sigma_i^\alpha)$ for each human demonstrator. Loud action primitives are characterized by high levels of average variance.

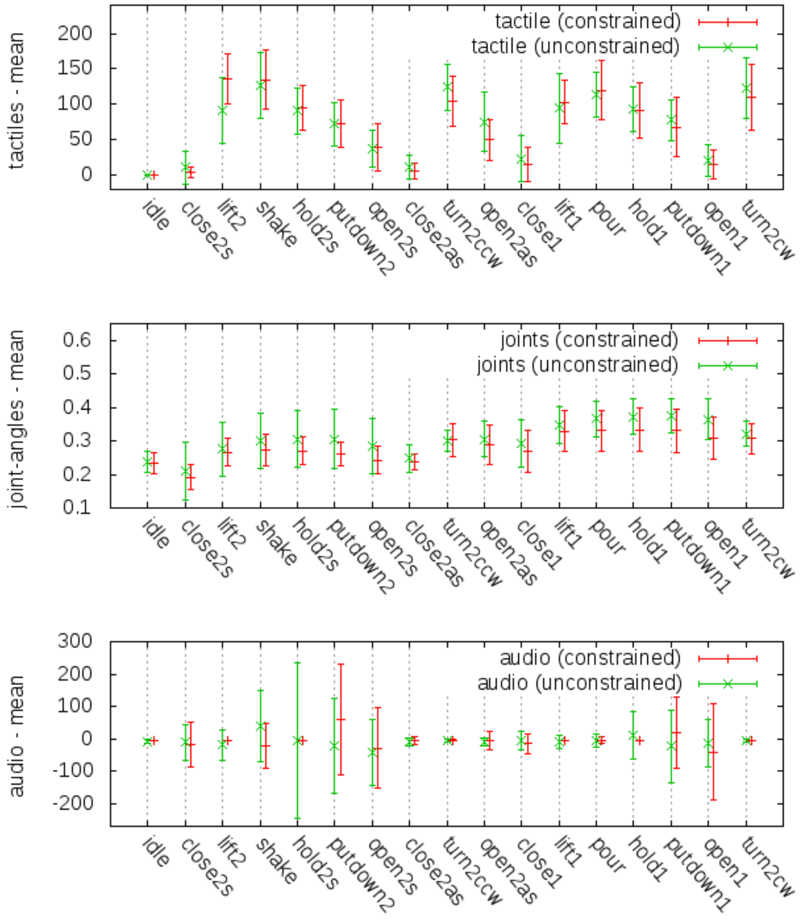


Figure 3.11: Comparison of action-specific means and variances of the modality-specific signals for constrained (red) and unconstrained (green) trials. Top row: $mean(\mu_i^\alpha)$ and $var(\mu_i^\alpha)$ for the tactile modality in constrained and unconstrained case. Middle row: $mean(\mu_i^\alpha)$ and $var(\mu_i^\alpha)$ for the joint-angles modality in constrained and unconstrained case. Bottom row: $mean(\mu_i^\alpha)$ and $var(\mu_i^\alpha)$ for the audio modality in constrained and unconstrained case.

constrained and unconstrained trials. The top row compares the application of force in both cases; the middle row compares the joint-angles dynamics and the bottom row illustrates the averages for the audio modality. All three sub-figures show large similarity of constrained and the unconstrained scenarios, in line with our assumptions.

## 3.5 Summary

In this chapter we have motivated and introduced a scenario and the required hardware setup for multimodal acquisition of manual interaction. The acquired data will be used for empirical studies with methods developed in the next chapters.

In the beginning of this chapter we have presented the recorded interaction sequence, consisting of a representative set of manual actions typical for a variety of daily scenarios. We have then presented a detailed description of the multimodal sensor setup employed for the recording of the following interaction components – the hands, the object, the external view of the interaction scene, and, optionally, the interaction triggers – aspiring a comprehensive multimodal manual interaction capture. The description of the hardware framework is followed by the description of two methods for ground truth acquisition: manual annotation and automated cue-based ground truth. Finally, based on the ground truth represented by the semantic label collection, the last section has presented several preliminary experiments investigating the action- and modality-specific characteristics of the multimodal data.

Taking the results of the preliminary experiments into consideration, firstly, different semantic properties of the recorded modalities motivate modality-specific modeling of action primitives. Secondly, the high levels of inter- and intrapersonal variability discourage from modeling of the absolute values corresponding to the action primitive subsequences.

The following Chapter 4 presents the underlying theory and a detailed discussion of the Bayesian segmentation framework employed in our work.

# Chapter 4

# Multimodal Interaction Decomposition: Theoretical Background

The purpose of this chapter is to provide the theoretical background necessary for the decomposition of interaction into action primitives. On its basis, the following Chapter 5 demonstrates the experimental results for multimodal decomposition, and Chapter 6 presents an approach towards identification of the resulting chunks.

Considering the state of the art methods (see Chapter 2) enormous advances have been made in interaction decomposition. However, most approaches are designed for domain-specific unimodal data, where either each action primitive is predefined, or a specific segmentation heuristic is applied. At the same time, the growing complexity of considered interaction scenarios (e.g. in cognitive robotics) requires an approach to decomposition that generalizes well to a wide range of actions and applications. Hence, in our work we aspire, firstly, a generic approach that scales well to a large number of actions and scenarios. Secondly, we aim at modularity w.r.t integration of multiple modalities.

The rest of the chapter contains a detailed description of our approach with regard to the above-mentioned issues:

- Section 4.1 introduces the employed generic decomposition framework handling the challenge of unknown interaction structure.

- Section 4.2 is dedicated to our approach to modeling of action primitives. Based on a series of simple stochastic models, our approach emphasizes scalability to a wide range of modalities, and scenarios.

- Section 4.3 describes our approach to bimanual and multimodal integration.

- Section 4.4 presents a summary of this chapter.

## 4.1 Decomposition Framework

Assuming that no action-specific or data-specific segmentation heuristics are available to guide the decomposition, the structure of the interaction is to be considered as unknown, i.e. action primitives constituting an interaction, as well as their number and locations are unknown. Inspired by recent psychological findings, central for our approach is the assumption that action primitives correspond to homogeneous regions within the time series. Thus, in our work the task of segmentation is reduced to finding such homogeneous regions.

The above-mentioned considerations and the demand for high scalability motivate application of a change point detection method. Due to the characteristics discussed previously in Section 2.4, we propose to employ the Bayesian multiple change point detection method introduced by P. Fearnhead [26, 27]. Importantly, this method has been previously applied to scalar and multivariate time series[1]. In the further subsection we introduce Fearnhead's segmentation method, and come back to our approach extending it for integration of multimodal and bimanual time series in Section 4.2.

### 4.1.1 Fearnhead's Algorithm

In his work, Fearnhead proposes a deterministic method that maximizes the posterior distribution of the number and location of change points w.r.t given observations. Central to the method is a dynamic programming algorithm based on the *filtering recursion*, an approach similar to the Viterbi algorithm [75] and methods for partition models [9, 8]. Hence, the method exhibits quadratic computational complexity in the number of data points $n$. However, an approximate version which exhibits linear complexity w.r.t. $n$ has demonstrated negligible errors [27].

The goal of the algorithm is, given an observation time series $y_{1:n}$ of length $n$ (representing an interaction in our use case)

$$y_{1:n} = (y_1, \dots, y_n), \tag{4.1}$$

to output a set of change points $\tau_i \in \mathbb{N}_0$:

$$\tau_0 = 0 < \tau_1 < \tau_2 < \ldots < \tau_m < n = \tau_{m+1}, \tag{4.2}$$

partitioning the data into $m+1$ subsequences corresponding to action primitives.

The following paragraphs describe the algorithm in detail, starting with the essential parts of the filtering recursion, the prior and the likelihood components used in our work for action primitive modeling (described in Subsections 4.1.1.1 and 4.1.1.2 respectively). Subsection 4.1.1.3 derives the filtering recursion. Finally, Subsection 4.1.1.4 presents Fearnhead's algorithm for change point detection based on the filtering recursion. In the following text we employ the same notation as in [27].

#### 4.1.1.1 Prior Distribution on Segment Length

The prior distribution on segment lengths employed in Fearnhead's approach is the first component of the action primitive modeling. We denote such prior by $g(l)$, where $l \in [1, n]$

[1] Tests have been conducted on e.g. well log data published by Ruanaidh et al. [44].

denotes the length parameter. There are various possibilities for the choice of prior. In [27] $g$ is specified by a negative binomial probability mass function:

$$g(l) = \binom{l-k}{k-1} p^k (1-p)^{l-k} \quad \text{and} \quad g_0(l) = \sum_{i=1}^{k} \binom{l-i}{i-1} p^i (1-p)^{l-i}/k,$$

where $g_0(l)$ is the probability mass function of the first change point after 0. "For small values of $p$ the negative binomial distribution can be thought of as a discrete version of the gamma distribution. Larger values of $k$ can reduce the number of very short segments" [27]. For the special case of $k = 1$ we receive a geometrical distribution, and the point process is Markov:

$$g_0(l) = g(l) = p(1-p)^{l-1}$$

for some probability $p$. In [26] a prior is alternatively specified by a function dependent on the length of the time series $n$ and the number of change points $m$:

$$p(m, n) = p^{m-1}(1-p)^{n-m}$$

for some probability $p$.

The distribution function of the distance between two successive points is calculated by:

$$G(l) = \sum_{s=1}^{l} g(s) \quad \text{and} \quad G_0(l) = \sum_{s=1}^{l} g_0(s).$$

Then the probability of $m$ change points occurring at positions $\tau_1, \ldots, \tau_m$ is given by the following product:

$$g_0(\tau_1) \left( \prod_{j=2}^{m} g(\tau_j - \tau_{j-1}) \right) (1 - G(\tau_{m+1} - \tau_m)),$$

where the first term corresponds to the prior of the first segment, the second term is a product of priors for all following segments until the last one, and finally, the third term corresponds to the prior of the last segment.

#### 4.1.1.2 Marginal Likelihood

Marginal likelihood of a subsequence in Fearnhead's approach estimates, how well a given subsequence can be described by a particular model without the knowledge of the model parameters. Let $y_{1:n}$ be the time series of observations and $0 < \tau_1 < \ldots < \tau_m < n$ denote an arbitrary segmentation of the time series. The observation data restricted to a time interval from $i$ to $k$, $i < k \in \{1, \ldots, n\}$ is denoted by

$$y_{i:k} = (y_i, \ldots, y_k).$$

The $j$-th segment consists of the observations from $\tau_{j-1} + 1$ to $\tau_j$.

Let a single parameter or a parameter vector $\theta_j$ specify the model associated with the $j$-th segment, $j = 1, \ldots, m + 1$ (models employed in this work are discussed in detail in Section 4.2). The priors for the parameters $\theta_j$ are denoted by $\pi(\theta_j)$ and are assumed to be statistically independent across segments. If time index $i$ is within the $j$-th segment, then the observation $y_i$ is distributed according to a density $f(y_i|\theta_j)$. Based on the independence assumption, the likelihood of the observations $y_{t:s}$ conditioned on the change point positions and the parameter $\theta$ results in:

$$\texttt{Pr}(y_{t:s}|t,\ s,\ \text{in the same segment },\theta) = \pi(\theta)\prod_{i=t}^{s} f(y_i|\theta).$$

Let $P(t,s)$ denote the marginal likelihood of time series data $y_{t:s}$, $t \leq s$, $\ t,s \in \{1,\dots n\}$ which is part of a single segment. Exploiting the independence assumptions the marginal likelihood can be calculated as follows:

$$\begin{aligned} P(t,s) &= \texttt{Pr}(y_{t:s}|t,s \text{ in the same segment}) && (4.3)\\ &= \int \prod_{i=t}^{s} f(y_i|\theta)\pi(\theta)d\theta, && (4.4)\end{aligned}$$

where $\theta$ denotes the model parameters, as described above.

#### 4.1.1.3 Filtering Recursion

For estimation of the maximum of the posterior distribution of segmentations w.r.t to the observations, described in the next section, Fearnhead introduces a set of auxiliary recursions.

For each $t \in [2,n]$ let $Q$ be defined as follows:

$$Q(t) = \texttt{Pr}(y_{t:n}|\text{change point at } t-1). \tag{4.5}$$

For $t = 1$, $Q(1) = \texttt{Pr}(y_{1:n})$. Further, according to the law of total probabilities and by dropping the conditional on the change point at $t-1$ for notational convenience, Equation 4.5 can be represented as an average over the next change point positions $s \geq t$:

$$\begin{aligned} Q(t) = &\sum_{s=t}^{n-1} \texttt{Pr}(y_{t:n}, \text{next change point at } s) && (4.6)\\ &+\texttt{Pr}(y_{t:n}, \text{no further change points}). && (4.7)\end{aligned}$$

Furthermore, assuming the change point process to be Markov, for each $s \in [t, n-1]$ the terms in Equation 4.6 can be calculated recursively as follows:

$$\begin{aligned} &\texttt{Pr}(y_{t:n}, \text{next change point at } s) && (4.8)\\ =&\texttt{Pr}(\text{next change point at } s)\texttt{Pr}(y_{t:n}|\text{next change point in } s) && (4.9)\\ =&\texttt{Pr}(\text{next change point at } s)\texttt{Pr}(y_{t:s}, y_{s+1:n}|\text{next change point in } s) && (4.10)\\ =&g(s+1-t)\texttt{Pr}(y_{t:s}|t,\ s \text{ in the same segment })\texttt{Pr}(y_{s+1:n}|\text{change point at } s) && (4.11)\\ =&g(s+1-t)P(t,s)Q(s+1). && (4.12)\end{aligned}$$

The transition from Equation 4.8 to Equation 4.9 is according to the definition of conditional probability. Equation 4.10-4.11, is firstly, according to the assumed Markov property of the change point process. Secondly, the definition of the prior distribution on segment length $g$ is applied, considering the conditional of the change point in $t-1$. Equation 4.11-4.12 is according to the definition of marginal likelihood and the definition of $Q$ (see Equation 4.5).

Analogously follows the calculation of the term in Equation 4.7.

$$\mathtt{Pr}(y_{t:n}, \text{no further change points}) = P(t,n) \cdot (1 - G_0(n-t)).$$

From the above derivation originating from [27] (Theorem I) follows for $t = 2, \ldots, n$:

$$Q(t) = \sum_{s=t}^{n-1} P(t,s)Q(s+1)g(s+1-t) + P(t,n)(1 - G_0(n-t)) \tag{4.13}$$

and

$$Q(1) = \sum_{s=1}^{n-1} P(1,s)Q(s+1)g_0(s) + P(1,n)(1 - G_0(n-1)). \tag{4.14}$$

#### 4.1.1.4 Change Point Detection Algorithm

In this section we describe the algorithm proposed by Fearnhead [26] for calculation of the maximum a posteriori (MAP) estimate of the segmentation. This method is based on a dynamic programming algorithm that first maximizes $Q$ for each $t \in \{1, \ldots, n\}$ yielding $Q^*$. In the second step, based on $Q^*$ the algorithm efficiently estimates the segmentation, optimal in the sense that a combination of a prior distribution on segmentations and the segment-wise likelihoods is maximized.

For a simple model specified by a marginal likelihood $P(t,s)$ and $t \in \{1, \ldots, n\}$, the recursive estimation of $Q^*$ is defined analogously to the above Equations 4.13 and 4.14:

$$Q^*(t) = \max\{\max_{t \leq s \leq n-1}(P(t,s)Q^*(s+1)g(s+1-t)), (P(t,n)(1 - G(n-t))\}. \tag{4.15}$$

If the model is a mixture model, the maximum is taken additionally over the set of model components $m \in \mathcal{M}$ of the mixture model $\mathcal{M}$:

$$Q^*(t) = \max_{m \in \mathcal{M}}\{\max_{t \leq s \leq n-1}(P_m(t,s)Q^*(s+1)g(s+1-t)), (P_m(t,n)(1 - G(n-t))\}, \tag{4.16}$$

where analogously to the above, $P_m(t,s)$ specifies the marginal likelihood function of the mixture model component $m \in \mathcal{M}$. $Q^*(n+1)$ is initialized as follows:

$$Q^*(n+1) = 1.$$

Further let $s^*(t)$ and $m^*(t)$ be the values that achieved the maximum. Then according to the algorithm proposed by Fearnhead the estimate of the change points $\tau_1^*, \ldots, \tau_m^*$ and the corresponding mixture model components models $m_1^*, \ldots, m_m^*$ can be obtained as presented in Algorithm 1. This procedure for estimation of change points and the corresponding mixture model components is employed in our experiments (see Chapter 5).

**Algorithm 1** Fearnhead's algorithm for calculation of change points and the corresponding models.

```
Require: $s^*(t)$ and $m^*(t)$
  $\tau_0^* \leftarrow 0$, $j \leftarrow 0$
  while $\tau_j^* < n$ do
      $\tau_{j+1}^* \leftarrow s^*(\tau_j^* + 1)$
      $m_{j+1}^* \leftarrow m^*(\tau_j^* + 1)$
      $j \leftarrow j + 1$
  end while
```

## 4.2 Modeling of Action Primitives

In this work we assume that a series of homogeneous regions corresponding to action primitives (see Chapter 2) constitute an interaction. Accordingly we assume, that each change point characterizes the beginning of an action primitive, and each homogeneous region is generated by a particular model.

Previous section presented a change point detection framework by Fearnhead, in which the subsequence modeling is defined by a marginal likelihood $P(t, s)$ (see Equations 4.15 and 4.16 respectively). In this section we propose a set of simple models that specify the calculation of marginal likelihood $P(t, s)$, and are later employed for unimodal, multimodal and bimanual modeling approaches of action primitives.

In the following subsections we first give an overview of different model types, used for the modality-specific likelihood calculation: a linear model, a constant model and a threshold model. Further we present a product and a mixture model that we propose for likelihood calculation in bimanual and multimodal time series.

### 4.2.1 Linear Models

Linear models have been used within regression analysis for signal segmentation (i.e. [74]). For two neighboring change points $\tau_i$ and $\tau_{i+1}$ the linear regression for the observations $y_{(\tau_i+1):\tau_{i+1}}$ associated to the $i$-th segment is given as follows:

$$y_{(\tau_i+1):\tau_{i+1}} = G_i^{(p_i)} \beta_i + \epsilon_{(\tau_i+1):\tau_{i+1}},$$

where $\epsilon_{(\tau_i+1):\tau_{i+1}}$ is a vector consisting of independent and identically distributed (i.i.d) Gaussian random variables $\epsilon_{i,j}$ modeling measurement noise with

$$\epsilon_{i,j} \sim \mathcal{N}(0, \sigma_i^2),$$

and $\beta_i$ is a vector of coefficients, describing the dependency between explanatory variables of the matrix $G_i^{(p_i)}$ and response of the $i$-th segment $y_{\tau_{(i+1)}:\tau_{i+1}}$; $G_i^{(p_i)}$ is the matrix consisting of basis components of the linear model; $p_i$ denotes the type of the model within the $i$-th segment.

In this paragraph we describe the selection of conjugate priors allowing analytical calculation of marginal likelihood (Equation 4.3) for each $s, t \in \{1, \ldots, n\}$, with $t < s$. For the $j$-th regression parameter of the $i$-th segment $\beta_{i,j}$ we assume a normal prior with mean 0 and an unknown variance $\sigma_i^2 \delta_j^2$ independent of all other regression parameters:

$$\beta_{i,j} \sim \mathcal{N}(0, \sigma_i^2 \delta_j^2). \tag{4.17}$$

For the variance of noise $\sigma_i$ we assume an Inverse-Gamma prior with global hyperparameters $\nu/2$ and $\gamma/2$:

$$\sigma_i^2 \sim \text{Inv-Gamma}(\nu/2, \gamma/2). \tag{4.18}$$

The priors are independent for different segments. This above choice of priors, given the Gaussian noise model, allows marginalisation of nuisance parameters $\beta_i$ and $\sigma_i$ [74]. In [74, 26, 77] it is proposed to estimate the parameters $\gamma$, and $\delta_1$, …, $\delta_{p_i}$ from data in order to increase the robustness of the prior.

Consider the observation data $y_{t:s}$, $s \geq t$, and a linear regression model of a fixed order $q$. Let $G$ be the $(s-t+1) \times q$ matrix of basis vectors for the $q$-th order linear regression model on this segment. Under the assumption of i.i.d. Gaussian noise, for model parameters $\theta = \{\sigma, \beta\}$ the likelihood of $y_{t:s}$ to form a segment is defined as follows:

$$P(y_{t:s}|\theta) = (2\pi\sigma^2)^{-((s-t)/2)} \times \exp\left(-\frac{\|y_{t:s} - G\beta\|^2}{2\sigma^2}\right). \tag{4.19}$$

Under the assumption of conjugate priors (see Equation 4.17 and 4.18) and the likelihood defined in Equation 4.19, the marginal likelihood is calculated analytically by integrating out the regression parameter $\beta$ and the variance $\sigma$ (see [74, 26]):

$$\begin{aligned} P(t,s) &= \int_\theta P(y_{t:s}|\theta)p(\theta)d\theta = \\ &= |M|^{1/2}(\gamma + \|y_{t:s}\|_P^2)^{-(\nu+s-t+1)/2} \times \frac{\Gamma\left(\frac{\nu+s-t+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \prod_{j=1}^{q} \delta_j^{-1}, \end{aligned}$$

with

$$\begin{aligned} M &= (G^T G + D)^{-1}, \\ P &= (I - GMG^T), \end{aligned}$$

and

$$\|y\|_P^2 = y^T P y,$$

where $I$ is a $(s-t+1) \times (s-t+1)$ identity matrix and $D$ defines the prior variance on the regression parameters:

$$D = Diag(\delta_1^2, \ldots, \delta_q^2).$$

Particularly interesting special cases of the linear model, a polynomial and an autoregressive model are used by Punskaya et al. [74] and Fearnhead [26] for segmentation.

#### 4.2.1.1 Polynomial and Autoregressive Models

The **autoregressive (AR) model** of order $n$ assumes that the observations are generated by an autoregressive process of order $n$. An example of a third-order model is:

$$G_{t:s}^{(3)} = \begin{pmatrix} y_{t-1} & y_{t-2} & y_{t-3} \\ y_t & y_{t-1} & y_{t-2} \\ \ldots & \ldots & \ldots \\ y_{s-1} & y_{s-2} & y_{s-3} \end{pmatrix}.$$

AR models of orders 1-3 are used in our work to describe the oscillating structure of the audio signal.

The **polynomial model** assumes that the relationship between the response and the explanatory variables can be modeled by a polynomial of a given degree $n$. Basis elements of this model are the orthogonal monomials $\{x^i\}_{0\leq i<n}$. In an example for $n = 3$, the matrix $G^{(3)}_{t:s} \in \mathbb{R}^{(s-t+1)\times 3}$ contains basis elements for the constant, linear and quadratic component:

$$G^{(3)}_{t:s} = \begin{pmatrix} 1 & x_t & x_t^2 \\ 1 & x_{t+1} & x_{t+1}^2 \\ \vdots & \vdots & \vdots \\ 1 & x_s & x_s^2 \end{pmatrix} .$$

Application of polynomial model for segmentation is very promising and therefore part of our future work.

### 4.2.2 Constant Models

A constant model estimates how well a segment $y_{t:s}$ of a scalar time series can be described by a constant function $f(x) = \mu$, where $\mu$ is the parameter of the likelihood function. Like in the previous section we assume conjugate priors in order to integrate out the parameter $\mu$ and analytically calculate the marginal likelihood $P(y_{t:s}|m_{\mathsf{c}})$, where we denote a constant model by $m_{\mathsf{c}}$. First of all the individual samples $y_k$ are assumed to be i.i.d. according to a Gaussian distribution:

$$y_k \sim \mathcal{N}(\mu, 1)\,,$$

where the mean $\mu$ is the unknown parameter of the constant model $m_{\mathsf{c}}$. We assume a Gaussian prior distribution for $\mu$ as well, i.e. $\mu \sim \mathcal{N}(\eta, u)$, where we set $\eta = \langle y \rangle$. This choice of $\eta$ yields the following simplified log marginal likelihood of $y_{t:s}$ for the constant model $m_{\mathsf{c}}$:

$$\begin{aligned} \log P(y_{t:s} \mid m_{\mathsf{c}}) &= \log \int P(y_{t:s} \mid \mu, m_{\mathsf{c}}) P(\mu \mid m_{\mathsf{c}}) d\mu && (4.20) \\ &= -u\mathrm{Var}(y_{t:s}) + C, && (4.21) \end{aligned}$$

where $C$ is a constant. Except for the choice of $\eta$, this likelihood calculation is identical to the one proposed by Fearnhead (see [27], Section 4.2). The result of the Equation 4.20 and 4.21 can be interpreted as follows: when the empirical variance of an approximately constant segment gets close to zero, thus maximizing the log likelihood independent of the exact value level $\mu$, the segment can be well approximated by a constant function. If the segment comprises two constant subsegments of different value level, the variance will become much larger, indicating that such a segment cannot be well fitted by a single constant model, but would be better fitted by two separate constant models. In our work we apply constant model to preprocessed hand-posture trajectories and an energy-based feature extracted from the audio signal.

### 4.2.3 Threshold Models

A threshold model is a binary model designed to roughly estimate whether the data of a segment mainly lies below or above a given threshold $\gamma$. The marginal likelihoods associated to these models, denoted by $m_{<\gamma}$ and $m_{>\gamma}$ resp., indicate how well the time series segment $y_{t:s}$ fits the assumptions of being below or above the threshold $\gamma$. With the independence assumption for the individual samples $y_k$, we define the improper marginal likelihood for $m_{<\gamma}$ as follows:

$$P(y_{t:s} \mid m_{<\gamma}) = \prod_{k=t}^{s} p(y_k|m_{<\gamma}), \tag{4.22}$$

$$\text{where} \quad p(y_k \mid m_{<\gamma}) = \begin{cases} 1, \text{ if } y_k < \gamma \\ p_o \text{ otherwise} \end{cases} \tag{4.23}$$

where $p(y_k|m_{<\gamma})$ is the probability, that a single sample $y_k$ fits the model assumption. The parameter $p_o$ determines the probability that $y_k$ does not fit the assumption. Denoting the segment length by $u = s - t$ and the number of not fitting samples by $n = |\{y_k > \gamma \ \mid \ t \leq k < s\}|$, and ignoring the constant normalization factor, we can derive the following, more compact formulas for both models:

$$P(y_{t:s} \mid m_{<\gamma}) = {p_o}^{n} \quad \text{and} \quad P(y_{t:s} \mid m_{>\gamma}) = {p_o}^{u-n} \tag{4.24}$$

As can be seen from Equation 4.24, the marginal likelihood becomes smaller, the more data points are on the wrong side of the threshold. In our work we apply the threshold model on preprocessed tactile data.

### 4.2.4 Product Models

A product model is a probabilistic model containing several component models, which are combined in a multiplicative way, assuming statistical independence of the individual models. In our work we assume that the multimodal data lies in a Cartesian product space of independent modality spaces. This allows an application of the product model to the multimodal sequences, where one modality or channel corresponds to one product model component. We consider a special case of a product model with weighted components. Assuming the independence of component models, marginal likelihood is calculated as a product of the individual model marginal likelihoods:

$$P(t, s) = \prod_{k \in K} P_k(t, s)^{w_k},$$

where $K$ denotes the number of model components, $P_k(t, s)$ denotes the marginal likelihood of the $k$-th model and $w_k \in \{w_1, \ldots, w_K\}$, $w_k \in [0, \infty[$ denote the weights. The values of $w_k$-s determine the weights of the individual likelihood terms in the product. Within our approach the parameter $w_k$ determines the influence of the $k$-th modality on the product likelihood. In case of $w_k = 0$ the corresponding likelihood term $P_k(t, s)^{w_k}$ is set to 1 and is therefore neutral to the product.

### 4.2.5 Mixture Models

A mixture model is a probabilistic model consisting of $K$ different components, whose probability density functions are additively weighted to form the mixture probability density function of the model.

Let a mixture model $m_{\texttt{mix}}$ consist of $K$ model components with possibly vector-valued parameters $\theta_1, \ldots, \theta_K$ associated with each model component. Let $\phi_k$ denote the mixture weight, i.e. prior probability $p(k)$ of a particular mixture component $k$: $p(k) = \phi_k$. With $P_k(t, s)$ denoting the marginal likelihood of $y_{t:s}$ for the model component $k$, marginal likelihood for a mixture model can be defined as follows:

$$P(t, s) = \sum_{1 \leq k \leq K} P_k(t, s)\phi_k.$$

In our work a mixture model component is a product model, used for likelihood calculation of a multimodal segment. We describe the application of the modeling approach to the segmentation of multimodal time series in Chapter 5.

## 4.3 Multimodal Bimanual Segmentation Approaches

In this section we introduce our approach to multimodal integration within a decomposition framework.

As previously described in the beginnig of the chapter, a common approach towards integration of multiple modalities primarily conducts a modality-specific segmentation followed by a heuristic-based merge of the segment borders over all modalities. In order to overcome this limitation, in this section we propose an extension of Fearnhead's algorithm for bimanual and multimodal segmentation. This accounts for one of the main contributions of this work.

We begin by briefly recapitulating the necessary notation. For a given time series $y_{1:n}$ we use the notation $y_{|\texttt{mod}}$ to indicate the restriction to a modality $\texttt{mod} \in \{\texttt{t}, \texttt{j}, \texttt{a}\}$, where $\texttt{t}$, $\texttt{j}$, $\texttt{a}$ designate tactile, joint and audio modality respectively. An additional restriction of the modality-specific time series to the left or the right hand is denoted for each modality with an additional index $\texttt{l}$ or $\texttt{r}$, i.e. $y_{|\texttt{tl}}$ or $y_{|\texttt{tr}}$ to refer to the tactile data for the left and right hands resp. In the following text we use the term *channel* to address a time series restricted in this way. Application of a feature extractor $f$ is denoted by $f(y_{1:n})$.

The rest of the section is structured as follows. In Subsection 4.3.1 we motivate and describe our approach towards decomposition modeling of bimanual unimodal data. The following two Subsections 4.3.2 and 4.3.3 describe two approaches to multimodal segmentation: a *hierarchical* and a *parallel* approach respectively.

### 4.3.1 Bimanual Segmentation Approach

Each unimodal recording of a bimanual interaction, such as hand-posture trajectories, acceleration or pressure has two hand-specific channels. In the next paragraphs we discuss, what method should be applied in order to estimate a common segmentation for both channels.

In order to motivate our approach, we first consider an example of a typical object interaction, during which a human demonstrator grasps the object and then releases it. During a typical grasping, illustrated in Figure 4.1 based on the tactile sensor output, both

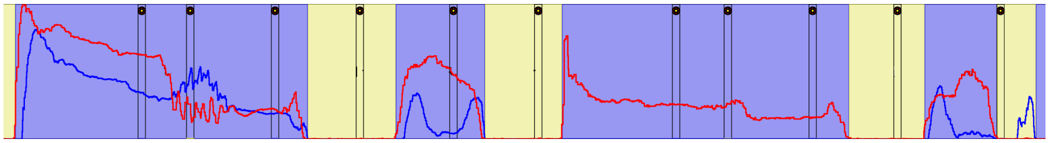


Figure 4.1: An example of a trial with synchronous bimanual movement, small asynchrony is reflected in the tactile output. All three bimanual object contact regions (blue regions 1,2 and 4) demonstrate that in the beginning and in the end of the region, both hands establish or loose contact asynchronously.

hands hardly ever establish object contact precisely at the same time. Based on these observations, we assume that such asynchrony is inherent to some synchronous bimanual actions, such as "grasping" or "releasing". In the case when both channels are considered separately, the output of the segmentation (e.g. for bimanual grasping) would contain two temporally close segment borders for each hand relating to the beginning of the same semantic action – the bimanual grasping. In order to avoid oversegmentation and to have a common border for both hands for one bimanual action an additional fusion step employing a heuristic algorithm would be needed.

Our approach aims to solve the asynchrony-related oversegmentation problem without a heuristic-based merge. For this purpose we propose a joint modeling approach applied within the segmentation framework of Fearnhead.

Consider an exemplary basic binary partition of the respective activities of both hands during an interaction with an object into "object contact" and "no object contact" (see Figure 4.1). We denote the corresponding models by $m_L$, $m_R$ ("object contact" with left or right hand respectively) and $m_l$, $m_r$ ("no object contact" with left or right hand respectively). As both hands can act independently of each other, the above classification yields a combination of four possible states: contact for both hands, contact for the left hand only, contact for the right hand only, no contact for both hands. Due to the independence assumption the hand-specific models (e.g. $m_l$ and $m_R$) in our approach are combined multiplicatively to yield a joint bimanual model, denoted by $m_{lR}$. We denote the composite models that correspond to each of these states by $m_{LR}$, $m_{Lr}$, $m_{lR}$, $m_{lr}$. Hence, the overall model is then a mixture model consisting of four product model components. The corresponding mixture model likelihood under the assumption of uniform component priors is as follows:

$$P(y_{t:s|\mathtt{t}}) = \frac{1}{4}P(y_{t:s|\mathtt{t}} \mid m_{lR}) + \frac{1}{4}P(y_{t:s|\mathtt{t}} \mid m_{Lr}) + \frac{1}{4}P(y_{t:s|\mathtt{t}} \mid m_{LR}) + \frac{1}{4}P(y_{t:s|\mathtt{t}} \mid m_{lr}). \quad (4.25)$$

Importantly, in order to estimate common segmentation of bimanual channels, we propose to employ the above mixture model within the segmentation framework by Fearnhead. Note, that within a product model, any two suitable models of the same type could be employed for bimanual unimodal segmentation. In contrast to a heuristic-based approach, the presented method not only allows simultaneous processing of data for both hands, but also encompasses a prior distribution on segment lengths, a mechanism preventing the oversegmentation and, therefore, making the method particularly suitable for modeling of action primitives.

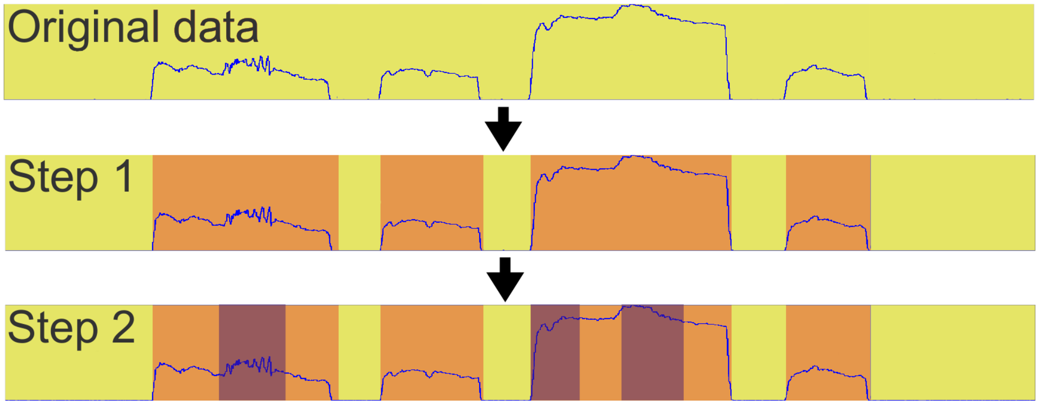


Figure 4.2: An schematic example of a two-stage hierarchical segmentation; original data (first row) is segmented in the step 1 resulting in yellow and orange segments (second row); in step 2 the orange segments generated in step 1 are subsegmented (third row).

### 4.3.2 Hierarchical Segmentation Approach

The main concept of the hierarchical segmentation is an iterative refinement of the semantic decomposition structure by conducting a series of segmentation and subsegmentation steps. In each iteration of the algorithm, a different feature or modality serves as a basis for the segmentation. Figure 4.2 illustrates the hierarchical segmentation approach on an example of two subsequent segmentation steps.

For an arbitrary time series $y_{1:n}$ an outline of the procedure is presented in Algorithm 2. Here, in each iteration step $i \in \{1, \ldots, S\}$ denoting the level of the segmentation hierarchy, the algorithm explores the structure of a given feature time series $f_i(y_{|\mathtt{mod}_i})$ by applying Fearnhead's algorithm. Essential for the multimodal integration is the fact that in each iteration step $i$ the hierarchical segmentation method only refines the segmentation structure previously generated for the time series $y_{1:n}$ in the step $i-1$. And finally, the new segmentation generated for the current level $i$ is applied globally to $y_{1:n}$. If we denote the segmentation on the level $i$ by $\Xi_i$, then from the design of the algorithm follows:

$$\Xi_1 \subset \ldots \subset \Xi_S. \tag{4.26}$$

Filtering is an optional operation that allows to select or postprocess segments according to their model description. Note, that the number of segmentation steps, the feature extractors and the model sets $\mathcal{M}_i$ used in each step $i$ have to be specified in advance based on the prior knowledge.

The experimental evaluation of the hierarchical segmentation method will be presented in Section 5.4. In this section we illustrate application of the method to bimodal data, consisting of the tactile and the audio modalitie, and a detailed discussion of the employed models.

**Algorithm 2** Outline of the hierarchical segmentation approach. Note: we use set and tuple notation interchangeably for sets whose elements can be ordered.

```
Require: $(\lambda_1, \dots, \lambda_S)$    ▷ Level-specific priors on segment lengths
Require: $(\mathcal{M}_1, \dots, \mathcal{M}_S)$    ▷ Level-specific model definition
  $\Xi_0 \leftarrow (\tau_0^1 = 1, \tau_0^2 = n), n_0 \leftarrow 2$    ▷ Initialize the sequence of change points
  $M_0 \leftarrow ()$.    ▷ and the corresponding model descriptors
  for $i = 1, \dots, S$ do
    $C_i \leftarrow \left(c_i^j = y_{\tau_{i-1}^j:\tau_{i-1}^{j+1}} \mid 1 \leq j < n_{i-1}\right)$    ▷ Partition $y_{1:n}$ into a sequence of chunks
                                                              ▷ according to $\Xi_{i-1}$
    if $M_{i-1} \neq ()$ then    ▷ Optional filtering
      $C'_i \leftarrow \left(c_i^j \in C_i | c_i^j \text{ selected by } M_{i-1}\right)$    ▷ Perform filtering of $C_i$ based on $M_{i-1}$
    else
      $C'_i \leftarrow C_i$
    end if
    for each chunk $y_{\tau_{i-1}^j:\tau_{i-1}^{j+1}} \in C'_i$ do
      $(\Xi_i^j, M_i^j) \leftarrow \text{fearnhead}(f_i(y_{\tau_{i-1}^j:\tau_{i-1}^{j+1}|\mathtt{mod}_i}), \lambda_i, \mathcal{M}_i)$    ▷ Apply Fearnhead's
                                   ▷ segmentation to chunk restricted to modality $\mathtt{mod}_i$
    end for
    $\Xi_i \leftarrow \Xi_{i-1} \cup \Xi_i^1 \cup \dots \cup \Xi_i^{n_{i-1}} = (\tau_i^1, \dots, \tau_i^{n_i})$
    $M_i \leftarrow (M_i^1, \dots, M_i^{n_i})$.
  end for
```

### 4.3.3 Parallel Segmentation Approach

The *parallel approach* is a generalization of the unimodal bimanual approach (see Section 4.3.1) for multiple modalities. Similar to the unimodal bimanual approach, the algorithm estimates a segmentation for the complete multimodal time series $y_{1:n}$ in one pass and is characterized by a mixture model $\mathcal{M}_{\mathtt{mix}}$, whose components are product models (see Equation 4.25). In contrast to the unimodal bimanual integration, the mechanism of the multimodal integration is realized with the product models that may contain modality-specific components of *different types*. An essential feature of the parallel approach is the weight vector controlling the influence of the individual modalities within the product model. The procedure is outlined in Algorithm 3. Hence, given a mixture model, and the weight vector, the procedure calculates a multimodal segmentation in one pass by applying Fearnhead's algorithm.

An empirical evaluation of the algorithm will be presented in Section 5.5. This section presents a detailed description of the corresponding mixture model, a discussion of the parameter choice ($\lambda$ and the weight vector), and an application of the algorithm for all three modalities.

## 4.4 Summary

This chapter has presented the theoretical framework aiming at decomposition of interaction into action primitives. The proposed approach addresses the following three challenges:

**Algorithm 3** Outline of the parallel segmentation approach.

```
Require: $(w_1, \dots, w_r)$                                          ▷ Weight vector
Require: $\mathcal{M}_{\texttt{mix}}$                                 ▷ Mixture model
Require: $\lambda$                                                    ▷ Prior on segment lengths
  $\hat{y}_{1:n} \leftarrow$ synchronize$(f_1(y_{1:m_1|\texttt{mod}_1}), \dots, f_r(y_{1:m_r|\texttt{mod}_r}))$   ▷ Synchronize the multimodal
                                                                      ▷ feature time series.
  $(\Xi, M) \leftarrow$ fearnhead$(\hat{y}_{1:n}, \lambda, \mathcal{M}_{\texttt{mix}})$   ▷ Apply Fearnhead's segmentation
```

unknown interaction structure, modeling of action primitives, and integration of multiple modalities.

Central for addressing the first challenge, the unknown structure of interaction, has been the assumption that action primitives correspond to homogeneous regions within the time series. Chapter 2 motivates the application of the Bayesian change point detection framework introduced by P. Fearnhead [26, 27]. Briefly, it is a deterministic method that maximizes the posterior distribution of the number and locations of change points w.r.t. observations.

In order to address the second challenge, the decomposition approach builds upon various models of homogeneity, determining the semantics of the resulting action primitives. Models proposed in Section 4.2 include simple models for unimodal segmentation (threshold, constant, linear), as well as product and mixture models for segmentation of multimodal and bimanual time series.

Section 4.3 proposes an extension of Fearnhead's procedure for multimodal and bimanual time series, the third addressed challenge. Here, two proposed methods, parallel and hierarchical segmentation, present individual mechanisms of multimodal and bimanual integration. While the parallel approach integrates over all modalities in one step, the hierarchical approach is based on iterative modality-specific semantic refinement. Chapter 5 presents results of the experimental evaluation for uni- as well as multimodal interaction decomposition.

Chapter 5

# Multimodal Interaction Decomposition: Experimental Results

This chapter contains a significant part of the experiments investigating interaction decomposition into action primitives with the methods derived in the previous chapter. Within the chapter, the analysis is organized according to the usage of modalities: a series of unimodal segmentation experiments, followed by the bimodal segmentations, and, finally, segmentation based on the complete time series consisting of three captured modalities is evaluated.

Each of the three modalities – tactile, joint-angles, and audio – has its individual semantics. Therefore, the goal of the preliminary unimodal experiments is to investigate the semantic relevance of the proposed modality-specific decomposition approach. The major part of the experiments explores and illustrates the multimodal methods introduced in Chapter 4. Their main target is to assess, how well the multimodal approaches can integrate unimodal segmentations.

Ground truth plays a vital role in the evaluation of segmentation quality, however its acquisition in the context of interaction identification is still an open question. Chapter 3 already presented and discussed the traditional ground truth acquisition methods, manual annotation, and the proposed alternative, automated cue-based ground truth. To investigate the advantages and the disadvantages of these two types of ground truth is a further objective of our study. Building upon the ground truth, we propose four segmentation quality measures, estimating the structural and temporal correctness of the segmentation.

The rest of the chapter is structured as follows:

- Section 5.1 presents the data pool, consisting of multimodal time series recorded by four human demonstrators.
- Section 5.2 introduces an evaluation method for assessing the quality of the generated segmentations, including its temporal and structural quality.

- Section 5.3 presents the unimodal experiments, and consists of a tactile, a joint-angles and an audio subsection, each containing a description of preprocessing, modeling and segmentation experiments. The section is concluded by an overview of the modality-specific segmentations generated by each of the three modalities.

- Sections 5.4 and 5.5 present the results of the multimodal approaches, based on two and three modalities respectively. Section 5.6 compares the quality of segmentations generated by both approaches.

- Section 5.7 presents a summary of all results.

## 5.1 Data Pool

This section presents the assembled data pool that has served as a basis for all conduced experiments.

As described in Section 3.2, the data pool is recorded with the help of multiple sensing devices, such as CyberGloves, iHands, a camera, and a contact microphone. Accordingly, the captured time series consist of joint-angle and tactile trajectories, audio, video, and an optional audio cue schedule. The interaction scenario recorded by four human demonstrators $\mathtt{hd}_i, i \in \{1, \ldots, 4\}$ with one test object has been previously outlined in Section 3.1. The scenario consists of an action sequence with a filled non-rigid plastic bottle. To prepare the data recording, each human demonstrator was given a sheet describing the interaction (see Appendix A). Although the structures of all trials should be identical except for timing differences, it turned out to be rather difficult for the human demonstrators to perform a large number of trials without errors. As a result, some trials exhibit structural differences like missing or additional tactile contacts or repeated actions. No correction of these irregularities has been conducted.

According to the two types of ground truth acquisition we differentiate between constrained and unconstrained trials (see Section 3.3). Unconstrained trials are recorded without providing audio cues to the subject and therefore with their natural execution speed. Constrained trials are recorded with audio cues controlling the beginning and the end of action execution. Figure 5.1 shows an example of a constrained trial (after preprocessing) and a corresponding cue-based ground truth. For some cues the figure shows a temporal deviation of the actual action execution timing from the cue signal, typical for the constrained scenario.

Table 5.1 presents an overview of the data set and its characteristics: recorded modalities, type of scenario, and characteristics of the human demonstrators. The column "cues" indicates whether data has been recorded with or without audio cues, corresponding to constrained and unconstrained scenario respectively. The columns "l-r" and "gender" state whether the human demonstrator is left- or right-handed ("l" and "r" respectively) and the gender.

## 5.2 Measures of Segmentation Quality

Following a decomposition of an interaction episode, the key question is: how to measure the quality of the resulting segmentation?

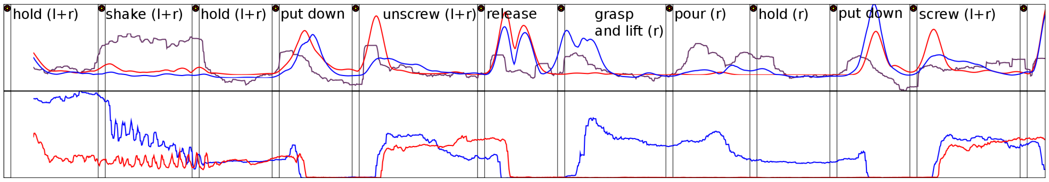


Figure 5.1: Illustration of a trial time series after preprocessing. Ground truth is automatically generated from the recorded audio cues that are drawn as black frames with an ● indicator. Upper plot: preprocessed joint-angle trajectories for the left (red) and right (blue) hand as well as the audio signal (brown). Lower plot: cumulative tactile feedback for the left (red) and right (blue) hand. The actions "unscrew" and "screw" show an example of bad alignment of a human demonstrator's actions to the cues. The captured tactile data shows a temporal offset between the actual start and the end of the action and the scheduled cues.

Table 5.1: Overview of the recorded trials in 7 experimental conditions; abbreviations are used to mark the recorded modalities: joint-angles (j), audio (a), tactiles (t), video (v); cues: denote whether during the recording audio cues have been emitted to mark the beginning or the end of the action execution; hd: denotes the human demonstrator; the column "l-r" denotes whether a human demonstrator is left- or right-handed.

| Condition | Modalities | Cues | hd | l-r | Gender | #Trials |
|---:|---|---|---|---|---|---:|
| 1 | j, t, a, v | yes | $\mathtt{hd}_1$ | r | m | 10 |
| 2 | j, t, a, v | no | $\mathtt{hd}_1$ | r | m | 10 |
| 3 | j, t, a, v | yes | $\mathtt{hd}_2$ | l | m | 20 |
| 4 | j, t, a, v | no | $\mathtt{hd}_2$ | l | m | 20 |
| 5 | j, t, a, v | yes | $\mathtt{hd}_3$ | l | m | 10 |
| 6 | j, t, a, v | no | $\mathtt{hd}_3$ | l | m | 10 |
| 7 | j, t, a | yes | $\mathtt{hd}_4$ | r | m | 30 |

In this section we describe the measures designed to evaluate the quality of the generated segmentation. For this purpose we compare the set of generated change points denoted by

$$\Xi := \{\tau_1, \ldots, \tau_M\} \tag{5.1}$$

with the set of ground truth change points denoted by

$$C := \{c_1, \ldots, c_m\}. \tag{5.2}$$

Note that the change points are ordered: $\tau_i < \tau_j$ and $c_i < c_j$ for $i < j$. In order to evaluate the overall correctness of the generated change points $\Xi$ w.r.t. the ground truth change points $C$, it is necessary to consider several aspects: how close to each other the generated and the ground truth change points lie, what is the ratio between the number of generated and the ground truth change points, what is the ratio of generated segment lengths and the ground truth segment lengths? Therefore, inspired by Basseville [10], who introduced several

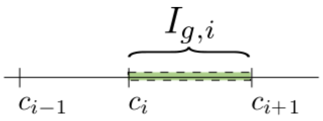


Figure 5.2: An example of a search range $I_{g,i}$ for calculation of segmentation granularity.

similar intuitive segmentation quality measures for online change detection, we consider the following four performance measures:

- segmentation granularity $\mu_g$,
- temporal accuracy $\mu_t$,
- overlap ratio $\mu_r$,
- missing segments ratio $\mu_m$.

The four segmentation measures – to be defined in more detail shortly – assess the quality and structure of the generated segmentation, considering the generated segmentation between two neighboring ground truth change points, segment's temporal alignment with the ground truth and its length, in case it has been detected. The following subsections describe the above measures and their calculation in detail.

### 5.2.1 Segmentation Granularity Index

The segmentation granularity index $\mu_g$ is the average number of change points that have been generated between the two neighboring ground truth segment borders. This index is a quality measure for the correctness of segmentation granularity. For a trial $\alpha$ and a ground truth change point $c_i$ we define $\epsilon^{\alpha}_{g,i}$ by the number of change points generated within the interval

$$I_{g,i} := [c_i, c_{i+1}] \tag{5.3}$$

between neighboring trial-specific ground truth segment boundaries $c_i$ and $c_{i+1}$:

$$\epsilon^{\alpha}_{g,i} := |\{\tau \in \Xi \mid \tau \in I_{g,i}\}|. \tag{5.4}$$

Figure 5.2 shows an example of an interval $I_{g,i}$. $\mu_{g,i}$ is an action-specific segmentation granularity measure that is defined for a given ground truth change point $c_i$ by the average of $\epsilon^{\alpha}_{g,i}$ over trials $\alpha \in A$. The segmentation granularity index $\mu_g$ is an average over all ground truth change points $c_i \in C$:

$$\mu_{g,i} := \frac{1}{|A|} \sum_{\alpha \in A} \epsilon^{\alpha}_{g,i} \tag{5.5}$$

$$\mu_g := \frac{1}{m} \sum_{1 \leq i \leq m} \mu_{g,i}. \tag{5.6}$$

Corresponding to $\mu_{g,i}$ and $\mu_g$ are the variances $\sigma_{g,i}$ and $\sigma_g$:

$$\sigma_{g,i} := \frac{1}{|A|} \sum_{\alpha \in A} (\epsilon^{\alpha}_{g,i} - \mu_{g,i})^2 \tag{5.7}$$

$$\sigma_g := \frac{1}{m} \sum_{1 \le i \le m} (\mu_{g,i} - \mu_g)^2. \tag{5.8}$$

A segmentation that has a perfect granularity w.r.t. the ground truth yields granularity index $\mu_g$ equal one. Deviation from this value indicates either an undersegmentation, insufficient number of generated segments ($\mu_g < 1$) or an oversegmentation ($\mu_g > 1$).

### 5.2.2 Temporal Accuracy Index

The temporal accuracy index measures how precise the timing of the generated segment borders is with respect to a given ground truth set $C$. For a trial $\alpha$ and a ground truth change point $c_i$ we define $\epsilon^{\alpha}_{t,i}$ as the distance between the generated and the expected change point. In order to find the generated change point corresponding to the ground truth change point $c_i$, we define a search interval around it. For a predefined constant $\epsilon$, let $I_{t,i}$ be defined as follows:

$$I_{t,i} := [c_i - \epsilon, c_i + \epsilon] \cap [c_{i-1}, c_{i+1}]. \tag{5.9}$$

Therefore, the interval $I_{t,i}$ is an $\epsilon$-neighborhood of $c_i$ limited from the sides by $c_{i-1}$ and $c_{i+1}$. Let $\Xi' := I_{t,i} \cap \Xi$ (cf. Equation 5.1) be the subset of all change point positions $\Xi$ that lie within the search range $I_{t,i}$. If $\Xi' \neq \emptyset$, then the change point

$$\tau^*_i := \arg \min_{\tau \in \Xi'} |c_i - \tau| \tag{5.10}$$

is the closest to the ground truth $c_i$ and determines the calculation of the temporal error:

$$\epsilon^{\alpha}_{t,i} :=| c_i - \tau^*_i |. \tag{5.11}$$

Because in this case it is not essential, whether the generated change point lies to the left or to the right from the ground truth $c_i$, we only consider the absolute value of the error in the above Equation 5.11.

The averages $\mu_{t,i}$, $\mu_t$, and variances $\sigma_{t,i}$ and $\sigma_t$ are calculated similar to Equations 5.5 - 5.8. Importantly, only the ground truth change points for which $\Xi' \neq \emptyset$ are used for averaging. The case, when $\Xi' = \emptyset$ for a given $c_i$, increases the missing segments index (see Subsection 5.2.4).

Altogether, the better the quality of the segmentation, the closer is the temporal accuracy index $\mu_t$ to zero.

### 5.2.3 Segment Overlap Ratio

In order to estimate the segment overlap ratio $\mu_r$, we first consider how much one generated segment $p := [\tau^*_i, \tau^*_{i+1}]$ (see Equation 5.10) overlaps with the corresponding ground truth segment $I_{g,i}$ (see Equation 5.3). For calculation of the segment overlap ratio $\epsilon^{\alpha}_{r,i}$, the quotient of generated segment length $|p| = (\tau^*_{i+1} - \tau^*_i)$ to ground truth segment length $|I_{g,i}|$ is calculated as

$$\epsilon_{r,i} := \min\{1, |p|/|I_{g,i}|\}. \tag{5.12}$$

The averages $\mu_{r,i}$, $\mu_r$, and variances $\sigma_{r,i}$ and $\sigma_r$ are calculated analogously to Equations 5.5 - 5.8. The higher the overlap ratio $\mu_r$, the better is the generated segmentation.

### 5.2.4 Missing Segments Index

The missing segments index is the percentage of change points that have not been detected. $\epsilon^{\alpha}_{m,i}$ is 1 in case no segment has been detected within the interval $I_{t,i}$ (see Equation 5.9) and 0 otherwise. The averages $\mu_{m,i}$, $\mu_m$, and variances $\sigma_{m,i}$ and $\sigma_m$ are calculated analogously to Equations 5.5 - 5.8. The lower the index $\mu_m$, the better is the segmentation.

## 5.3 Unimodal Segmentation

As a first study using the previously introduced data pool and the quality measures described in Section 5.1 and Section 5.2, we focus on the case of unimodal segmentation.

Essentially, segmentation of unimodal data into action primitives involves a choice of segmentation semantics, i.e. a decision about what kind of change to detect. The differences in the semantic content of each recorded modality (discussed previously in Section 3.4) lead to a modality-specific choice of the segmentation semantics. In order to represent different unimodal semantics, we propose to employ a previously introduced set of simple models: the threshold model, the constant model and th AR model (see Section 4.2). After a modality-specific segmentation semantics has been linked to a suitable model, preprocessing is necessary to produce an appropriate input for the chosen model.

After a brief parameter overview in Subsection 5.3.1, the following Subsections 5.3.2, 5.3.3 and 5.3.4 are dedicated to one of the three modalities: tactile, audio, and joint-angles. For each unimodal approach, we discuss the motivation behind the adopted segmentation semantics, the modeling and the corresponding preprocessing approach. Subsection 5.3.2, describing the unimodal approach to the tactile modality, demonstrates a particularly detailed example of the quantitative effect of the central parameters on the segmentation. Similarly detailed accounts for unimodal segmentation experiments investigating the dependency of the central parameters on the segmentation for the audio and the joint-angles modalities can be found in Appendix C.

### 5.3.1 Parameter Overview and Evaluation Issues

All of our studies are characterized by a small set of parameters: two "global" parameters $\lambda$ and $s$ relevant for all modalities, and further six remaining parameters characterizing modality-specific processing. According to the above-mentioned considerations, the parameters can be further categorized into preprocessing-, modeling- and segmentation-relevant parameters. Table 5.2 presents a parameter overview, arranged according to these categories.

The first global parameter $0 < \lambda < 1$ defines the prior distribution on segment lengths within the Bayesian framework:

$$p(t) = \lambda(1-\lambda)^{(t-1)}. \tag{5.13}$$

Here parameter $t$ denotes the length of a segment, i.e. values of the parameter $\lambda$ closer to 1 favor smaller segments. The number of generated segments is a monotonously increasing function of $\lambda$.

Table 5.2: Overview of the segmentation parameters for audio (a), joint-angles (j) and tactile (t) modalities.

| Name | Parameter type | Modalities | Description |
|---|---|---|---|
| $\lambda$ | segmentation | a, j, t | prior on segment lengths distribution |
| $\gamma$ | model | t | threshold parameter |
| $p_o$ | model | t | $p(y_k)$ being an outlier w.r.t. model |
| $s$ | preprocessing | a, j, t | subsampling rate of the time series |
| $\rho$ | preprocessing | a | signal range |
| $c$ | preprocessing | a | filtering threshold of high amplitude values |
| $w$ | preprocessing | a | width of the sliding window (variance calculation) |
| $\sigma$ | preprocessing | j | Gaussian smoothing parameter |

The second global parameter $s$ defines the subsampling rate of the time series. Higher subsampling rate leads to reduction of the number of data points, which in turn results in information loss, making the calculation less computationally expensive, less sensitive to small artifacts, but also less precise. The remaining local parameters, presented in Table 5.2 will be discussed in the corresponding modality-specific subsections below.

In the three following subsections we will explore the influence of the global parameters $\lambda$ and $s$ on the segmentation based on the quality measures $\mu_t$, $\mu_g$, $\mu_r$, and $\mu_m$ (see Section 5.2). In order to calculate the above measures, the search around each ground truth change point for a corresponding generated change point has to be constrained (see Equation 5.9). Consequently, the generated segments that are not included in the calculation of $\mu_t$, $\mu_r$, increase the missing segments index $\mu_m$. Hence, the above quality measures only serve as an approximation to the resulting segmentation quality. Therefore, in the following sections we will mainly discuss the relative influence of the parameters on the segmentation quality, such as improvement or deterioration of a quality measure, rather than aiming to find an optimal set of parameters based on the conducted experiments.

Another important evaluation issue arises in the cases, in which the video-based annotation for the ground truth change points cannot be clearly seen in the corresponding modality data and consequently do not get detected for a large range of parameter values. In this case, for an appropriate choice of values for e.g. $\lambda$ and $s$, it is necessary to compromise between the values of the segmentation granularity $\mu_g$ and the missing segments index $\mu_m$: the growing granularity goes along with the falling missing segments index. In the case if we try to get a "difficult" change point detected and reach a zero missing segments index, we may, at the same time, drive our procedure into a high oversegmentation. Such a situation corresponds to the case, when the additionally generated segment border does not only correspond to the ground truth, but also lie between the ground truth segment borders. Altogether, if a further increase of, e.g. $\lambda$, might decrease the missing segments index, it is necessary to ensure that the corresponding granularity stays beyond 1. A similar tradeoff has to be made between the missing segments ratio $\mu_m$ and the overlap ratio $\mu_r$. A decreasing number of missing segments indicates an increasing number of generated segments. This may result in a decrease of the segment overlap ratio, that should, in the optimal case, be close to 1.

In all experiments, averages for the calculation of each segmentation index are built over 20 trials recorded by one human demonstrator in a constrained scenario. Because the aim of following Subsections 5.3.2-5.3.4 is to explore the appropriateness of the proposed modeling approach for the unimodal semantics, and to present an exploratory study of the influence of the main parameters, no corroboration of the findings with statistical confidence measures will be provided.

### 5.3.2 Tactile Modality

The tactile modality reflecting the force applied to the object, presents a rich source of information about the interaction with an object. The absolute values of the tactile modality data integrate many factors, such as object weight and orientation, the executed action and the human demonstrator. At the same time, the data exhibits a high degree of inter- and intrapersonal variability (illustrated in Section 3.4.1). Hence, in our work the semantics, chosen for the tactile modality is the binary "object contact" vs. "no object contact" for both hands (see Section 4.3.1). Such segmentation semantics is certainly one of the simplest possible, however, it is advantageous due to its invariance on the object, person or on the type of the grasp. In our future work we would like to conduct experiments with a more informative modeling approach.

In order to receive the above-mentioned segmentation semantics for the tactile data, we propose to employ binary threshold models (see Section 4.2.3). The jointly modeled bimanual segmentation can be conducted with the help of a mixture model containing multiplicatively combined simple threshold models, one for each hand (see Section 4.3.1).

The complete set of simple models employed for unimanual modeling consists of the following four: $m_l$, $m_L$, $m_r$ and $m_R$, where $m_l$ and $m_r$ denote "no object contact" and $m_L$ and $m_R$ - "object contact" for the left and the right hands respectively. In the following we will also refer to $m_l$ and $m_r$ models as "off-models", and to $m_L$ and $m_R$ as "on-models". The following table shows an overview of these simple models:

| Notation | Hand | Description |
|---|---|---|
| $m_l$ | left | no contact |
| $m_L$ | left | contact |
| $m_r$ | right | no contact |
| $m_R$ | right | contact |

The mixture model for bimanual activity (see Section 4.3.1, Equation 4.25) must consider all possible combinations for both hands and, therefore, contains four states: "no contact for both hands" ($m_{lr}$), "contact for left hand only" ($m_{Lr}$), "contact for right hand only" ($m_{lR}$), and "contact for both hands" ($m_{LR}$). The marginal likelihood, e.g. $P(y_{s:t} \mid m_{lR})$ is computed as a product of the individual likelihoods:

$$P(y_{s:t} \mid m_{lR}) = P(y_{s:t} \mid m_l) \cdot P(y_{s:t} \mid m_R).$$

Table 5.3 shows an overview of the resulting four mixture model components.

Preprocessing of the tactile time series is based on the assumption that for segmentation of an interaction in our scenario no finger-specific force measurements are needed. Hence, the tactile values for each hand are summed up to yield a cumulative tactile force. This approach aims to compensate for different limitations of the hardware and to reduce the

Table 5.3: Overview of the four product model components of the mixture model.

| Notation | Description |
|---|---|
| $m_{lr}$ | no contact for both hands |
| $m_{lR}$ | contact for the right hand only |
| $m_{Lr}$ | contact for the left hand only |
| $m_{LR}$ | contact for both hands |

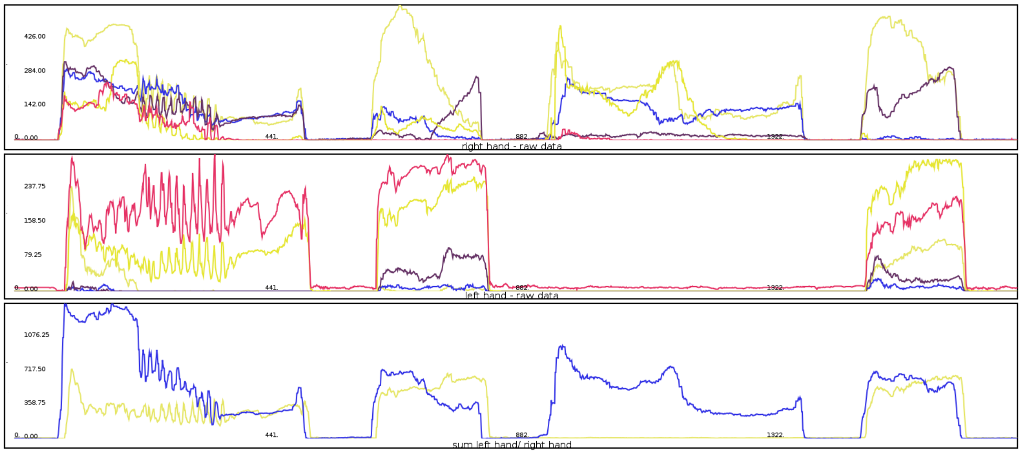


Figure 5.3: Preprocessing of the tactile time series: first and second rows: raw signal of the left and right hands $y_{tl}$ and $y_{tr}$ respectively; Each one of the two rows shows five trajectories, one for each finger. The tactile feedback is susceptible to different levels of noise or missing sensor input. Some regions of the plot demonstrate the problem of the data recording, i.e in the left hand only two fingers are recorded properly in this trial. One of the fingers in the left hand exhibits a higher level of noise, i.e. in the region of "no contact", there is a high level of signal. Third row: summed up data per hand (blue and yellow).

dimensionality from five dimensions to one dimension for each hand. The top two rows of Figure 5.3 show the raw sensor recordings of the left and the right hands $y_{tl}$ and $y_{tr}$. The third row of Figure 5.3 shows the summed values of the individual finger sensors for the left and the right hand.

In a threshold model, the main parameter $\gamma$ (see Section 4.2.3) defines the value of the threshold for recognizing contact. For ideal non-noisy data the parameter $\gamma$ could be set to zero. In our experiments it is chosen empirically to cut off the sensor noise. The further parameter $p_o$ is a constant probability of a data sample $y_k$ to be an outlier w.r.t to the model (see Equation 4.24). Larger values of $p_o$ make the model less sensitive towards outliers (the values below or above the threshold for “on” and “off” models respectively). In our experiments we used $p_o = 0.3$ for $m_l$ and $m_r$ models and $p_o = 0.7$ for $m_L$ and $m_R$.

The following subsections are dedicated to experiments investigating the temporal and structural accuracy, and robustness of the described segmentation approach w.r.t. different values of the prior segmentation length $\lambda$, subsampling rate $s$, and the threshold value $\gamma$. The tactile label collection from manual annotation (see Section 3.3.1) is employed as the ground truth. This labelling corresponds to the four mixture model states presented in the above Table 5.3.

#### 5.3.2.1 Threshold Parameter $\gamma$

The threshold parameter $\gamma$ controls the separation of data into "above" and "below" the threshold, corresponding to "contact" vs. "no contact" regions. Due to the unknown level of noise in the data we estimate the value of this parameter experimentally. In this paragraph we investigate the influence of the threshold parameter $\gamma \in \{10, 20, 30, 40, 50\}$ on the generated segmentation, while the remaining parameters stay constant with $\lambda = 10^{-4}$ and $s = 30$.

Table 5.4: Overview of the segmentation of the tactile modality for different values of the threshold parameter $\gamma$.

| $\gamma$ | $\mu_t$ | $\mu_g$ | $\mu_r$ | $\mu_m$ |
|---|---|---|---|---|
| 10 | 0.15 | 0.88 | 0.95 | 0.07 |
| 20 | 0.14 | 0.86 | 0.96 | 0.02 |
| 30 | 0.14 | 0.87 | 0.97 | 0.01 |
| 40 | 0.14 | 0.87 | 0.97 | 0.01 |
| 50 | 0.14 | 0.87 | 0.97 | 0.01 |

The results of the experiment are summarized in Table 5.4. For $\gamma = 10$, Table 5.4 shows a slightly larger granularity $\mu_g$ corresponding to a smaller segment overlap ratio $\mu_r$, and a larger missing segments index $\mu_m = 0.07$. We believe that the influence of noise leads to a small increase of the number of generated segments that, however, lie slightly outside of the search ranges and therefore can not be detected. For $\gamma > 10$ there is hardly any change in the values of the segmentation indices, implying that an increase of $\gamma$ within the test range has no significant effect on the generated segmentation, neither structural nor temporal. However, for a further growing value of $\gamma$, the number of segments will gradually decrease. For a sufficiently large $\gamma$, for which all data lies beneath the threshold, no change points will be generated.

#### 5.3.2.2 Subsampling Rate $s$

The subsampling rate $s$ determines the accuracy with which high frequency structures are be represented. Here we examine the effect of this parameter on the quality of the segmentation with threshold models. We conduct an evaluation for $s \in \{10, 30, 50, 70\}$ with remaining parameters constant: $\lambda = 10^{-4}$, $\gamma = 15$. We expect that fewer segments are generated for growing values of the subsampling rate[1].

Table 5.5 presents the summary of the experimental results, corroborating the expected effect: subsampling rate $s$ is inversely correlated with the granularity of segmentation $\mu_g$. In

[1] Subsampling rate $s = 10$ yields a time series frequency of approximately 20 Hz.

Table 5.5: Overview of the segmentation of the tactile modality for different subsampling rates.

| $s$ | $\mu_t$ | $\mu_g$ | $\mu_r$ | $\mu_m$ |
|---|---|---|---|---|
| 10 | 0.13 | 0.90 | 0.94 | 0.05 |
| 30 | 0.13 | 0.87 | 0.95 | 0.07 |
| 50 | 0.13 | 0.87 | 0.95 | 0.06 |
| 70 | 0.12 | 0.78 | 0.97 | 0.41 |

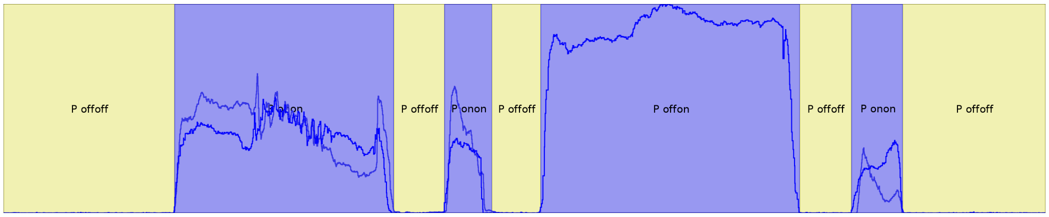


Figure 5.4: Example of a trial segmentation based on tactile modality with threshold models; A combination of parameters $s = 30$ and $\lambda = 10^{-4}$ yields segmentation structure corresponding to ground truth.

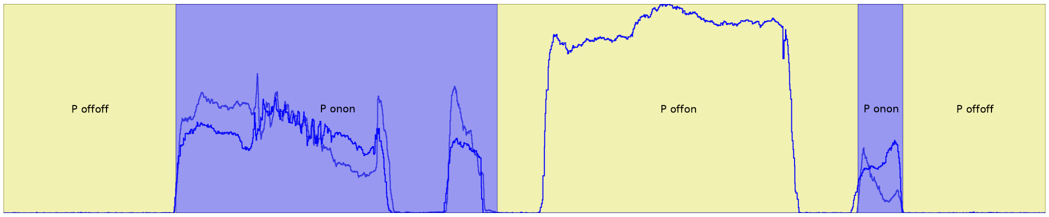


Figure 5.5: Example of a trial segmentation based on tactile modality with threshold models; A combination of parameters $s = 70$ and $\lambda = 10^{-4}$ yields undersegmentation.

contrast to the previous experiment, the increase of granularity is accompanied by decrease of the missing segments index $\mu_m$. This indicates that for falling $s$ the newly generated segments lie within the search ranges of the corresponding ground truth change points and can be detected.

Figures 5.4 and 5.5 illustrate two trial segmentations. Figure 5.4 shows an example of a generated segmentation corresponding to the ground truth segment structure ($s = 30$). Figure 5.5 shows an example with $s = 70$, corresponding to sample frequency of approximately 3 Hz. For this rate, the "off"-samples become outliers of the "on"-model. Therefore, the first two blue segments, showed in Figure 5.4, are merged to one in Figure 5.5. Within the test range, the missing segments rate increases from $\mu_m = 0.05$ to $\mu_m = 0.41$, indicated a strong influence on the segmentation.

Table 5.6: Overview of segmentation indices for the tactile modality for different values of $\lambda$.

| $\lambda$ | $\mu_t$ | $\mu_g$ | $\mu_r$ | $\mu_m$ |
|---|---|---|---|---|
| $10^{-9}$ | 0.12 | 0.77 | 0.96 | 0.42 |
| $10^{-8}$ | 0.13 | 0.75 | 0.95 | 0.4 |
| $10^{-7}$ | 0.13 | 0.88 | 0.95 | 0.11 |
| $10^{-6}$ | 0.13 | 0.87 | 0.95 | 0.07 |
| $10^{-5}$ | 0.13 | 0.87 | 0.95 | 0.06 |
| $10^{-4}$ | 0.13 | 0.87 | 0.94 | 0.06 |
| $10^{-3}$ | 0.13 | 0.88 | 0.94 | 0.06 |
| $10^{-2}$ | 0.13 | 0.88 | 0.95 | 0.06 |
| $10^{-1}$ | 0.13 | 0.94 | 0.93 | 0.05 |

#### 5.3.2.3 Prior Length Parameter $\lambda$

The parameter $\lambda$ influences the a-priori probability distribution of the segment length: the smaller the value of $\lambda$ the larger and fewer are the generated segments. In this experiment we investigate the influence of $\lambda$ on the segmentation of tactile data. We have conducted an evaluation for $\lambda \in \{10^{-9}, \ldots, 10^{-1}\}$, while other parameters remained constant: $s = 30$, $\gamma = 15$.

Table 5.6 presents the experimental results and demonstrates the anticipated effect of growing value of $\lambda$ on the segmentation: a strong increase of the granularity $\mu_g$ coupled with a decrease of the missing segments rate $\mu_m$. Larger values of $\lambda$ cause generation of smaller segments corresponding to a rise of segmentation granularity $\mu_g$ and at the same time a decrease of $\mu_m$, similar to the previous experiment. For too small values of the parameter, e.g. $\lambda < 10^{-8}$, the ratio of missing segments rises to about 40% and the granularity falls to $\mu_g = 0.77$ making the results comparable to the results for $s = 70$ in the previous subsection. Within the test range, $\lambda$ does not have a strong influence on $\mu_t$ or $\mu_r$. This implies that in the case a change point has been generated for different values of $\lambda$, the change of the parameter's value has no significant effect on the position of the change point.

#### 5.3.2.4 Tactile modality: Summary

The proposed method for tactile segmentation generated a robust segmentation into four different types of object contact state for both hands.

Table 5.7: Influence of parameters on the segmentation indices for segmentation based on the tactile modality.

| Parameter | Direction of Change | $\mu_t$ | $\mu_g$ | $\mu_r$ | $\mu_m$ |
|---|---|---|---|---|---|
| Threshold $\gamma$ | ↑ | No significant effect | ↓ | ↑ | ↓ |
| Subsampling rate $s$ | ↑ | No significant effect | ↓ | ↑ | ↑ |
| Prior distribution $\lambda$ | ↓ | No significant effect | ↓ | ↑ | ↑ |

Table 5.7 sums up the results of the experiments demonstrating the influence of the three parameters $\lambda$, $s$ and $\gamma$. Parameters $\lambda$ and $s$ have both, quantitatively and qualitatively comparable effect on the segmentation: larger $\lambda$ and smaller $s$ cause larger segmentation granularity and smaller missing segments index. There is hardly an influence on temporal precision for the generated segments.

For very small values of $\lambda \leq 10^{-9}$ and large values of $s \geq 70$ the values of all indices are approximately equal and show a drastic increase of missing segments ratio and a considerable fall of segmentation granularity. Sufficiently large values of $\gamma > 10$ result in segmentation that is robust against noise. Due to application of Fearnhead's method, incorporating a model of the segment length, the segmentation has demonstrated robustness against noise even for small values of $\gamma$, otherwise leading to severe oversegmentation.

### 5.3.3 Audio Modality

In our experiments, the audio modality is recorded by a contact microphone and captures object-centered acoustic noise caused by an interaction. The attachment of a microphone enables to capture sounds generated by the object during manipulation (see Section 3.2.2) and filter almost all environmental noise. During a manipulation, i.e "pushing", "shaking", "pouring" or "stirring", the structure of the raw audio signal has an approximately homogeneous oscillatory structure. Our approach builds upon the assumption that within an interaction such homogeneous areas in the audio signal correspond to action primitives. In the next two paragraphs we show two modeling approaches, implementing the above-mentioned change detection with an autoregressive and a constant model (previously described in Sections 4.2.1.1 and 4.2.2 respectively).

Due to the sensitivity of the contact microphone, not only audio-related action primitives, e.g. "shaking", "pouring" or "putting down", but also tactile-related action primitives, e.g. "grasping" or "screwing" can be detected in the audio signal. Therefore for evaluation of segmentation we use the combination of audio and tactile label collections generated from the manual annotation (see Appendix B).

#### 5.3.3.1 Autoregressive Model

Autoregressive processes have proven to be well suited for modeling of audio signals due to their stationary and oscillatory structure. Preprocessing consists of several empirically established steps (see Figure 5.6) that aim at improving of audio signal segmentation with AR models. The first row of the figure shows an example of the raw audio signal after subsampling. The second row illustrates the effect of a preprocessing step during which a predefined percentage $c$ of the highest values of the signal are cut out. This step aims at homogenizing the signal by getting rid of high signal peaks that may correspond to random acoustic noise produced by the object apart from the interaction, e.g. cracking of the plastic bottle. The next preprocessing step that takes the "pseudo" square-root of the resulting signal, is calculated as follows:

$$y_i^{'} = \begin{cases} \sqrt{y_i}, & \text{if } y_i \geq 0 \\ -\sqrt{|y_i|}, & \text{if } y_i < 0 \end{cases} \tag{5.14}$$

The output is illustrated in the third row of the same figure. Preliminary experiments (not included in this work) have showed that these preprocessing steps improve segmenta-

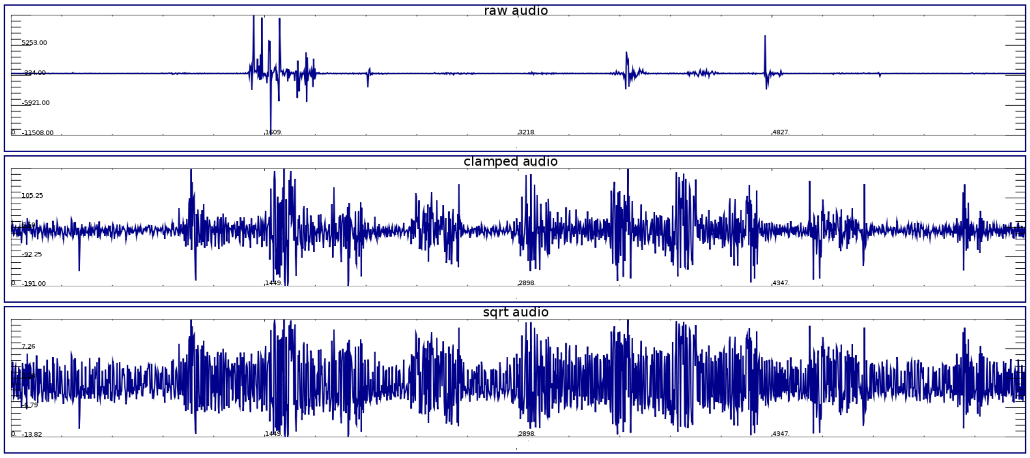


Figure 5.6: Audio preprocessing for AR modeling. First row: raw audio signal; second row: clamped audio signal, $c = 5$; third row: square root of the clamped audio signal.

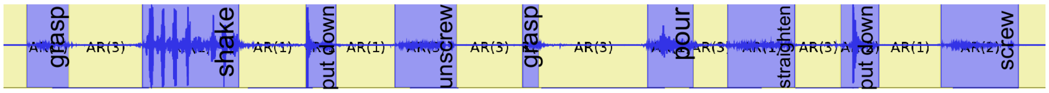


Figure 5.7: An example of an audio signal segmentation with AR models of order 1,2 and 3. The parameters are set to $s = 5$, $\lambda = 10^{-5}$ and $r = 12$. The generated segmentation corresponds among others to action primitives such as "grasp", "shake", "put down", "screw" and "pour".

tion with AR-models. The amplitude of the audio signal influences the value of segment marginal likelihoods estimated within Fearnhead's algorithm. For this reason in the third preprocessing step the signal is scaled to a range defined by the parameter $\rho$.

In the following we report on the experiments investigating the unimodal segmentation with a mixture of AR models of order 1, 2 and 3[2], whereby the theoretical background of the AR modeling have been previously discussed in Sections 4.2.1.1. Figure 5.7 shows an example of the resulting segmentation. The figure shows a decomposition of interaction into segments corresponding to action primitives, such as “grasp”, “shake”, “put down”, “screw”, “pour” and “unscrew”.

A detailed description of the experiments demonstrating the influence of the global parameters $s$ and $\lambda$, similar to the previous section, can be found in Appendix C.1.1. In this paragraph we solely present the main results summed up in Table 5.8.

Table 5.8 shows that within the test range the parameters $\lambda$ and $s$ are, similar to the previous section, inversely related. Both, an increase of $\lambda$ and a decrease of $s$, increase the granularity of the segmentation $\mu_g$ and decrease the missing segments ratio $\mu_m$. Hence, with both parameters, the structural accuracy of the segmentation can be improved. However, in contrast to the decrease of $\lambda$, an increase of the subsampling rate has an additional negative

[2] As a basis for our implementation we have used the code available online [25]

Table 5.8: Overview of the parameter influence for audio segmentation with a mixture of AR models.

| Parameter | Direction of Change | $\mu_t$ | $\mu_g$ | $\mu_r$ | $\mu_m$ |
|---|---|---|---|---|---|
| Subsampling rate $s$ | ↑ | ↑ | ↓ | ↓ | ↑ |
| Prior distribution $\lambda$ | ↓ | No significant effect | ↓ | ↓ | ↑ |

effect on the temporal error. We believe that this is due to the information loss connected with the subsampling.

Our experiments have showed that the parameters have a very strong influence on the AR-based modeling (the detailed descriptions can be found in Appendix C.1.1), leaving only a small range in which the generated segmentation is close to optimal. An alternative modeling approach is discussed in the next paragraph.

#### 5.3.3.2 Constant Model

In this paragraph we present the application of constant models (see Section 4.2.2) for segmentation of the audio signal. We assume that such models capitalize on the previously stated property that action primitives correspond to regions of the interaction in which the audio signal has an approximately homogeneous oscillatory structure and constant amplitude. To make such regions detectable for the constant model, we calculate signal variance with a sliding window. In the resulting time series the regions of the constant amplitude yield approximately constant output that we use as an input to Fearnhead's segmentation.

In contrast to applying a mixture of three AR models to approximately raw audio signal (see Subsection 5.3.3.1), we aspire to improve the segmentation of the audio signal based on the above preprocessing. We assume that detecting constant regions in a feature time series smoothed by a sliding window, is less sensitive to change of parameters, such as the subsampling rate $s$, than the AR models, where too large subsampling may easily eliminate the segmentation-relevant structures (see Appendix C.1.1).

Figure 5.8 illustrates the preprocessing: the first row shows audio signal after subsampling, the second row illustrates the variance computed in a sliding window of width $w = 20$. The value of the parameter $w$ is a trade-off: increasing of the window width results in a smoothing effect, causing less precise segmentation; decreasing of the window width results in a time series similar to the original oscillating structure and unsuitable for fitting with a constant model.

A detailed description of the experiments investigating the influence of the global parameters $s$ and $\lambda$ on the generated segmentation can be found in Appendix C.1.2. The evaluation of the segmentation has been conducted analogously to Subsection 5.3.3.1. In this paragraph we solely show the main results summed up in Table 5.9.

Table 5.9 shows that within the test range the parameters $\lambda$ and $s$ have an analogous effect on the generated segmentation as described in all previous experiments in this section. However, the model has showed to be more robust towards changes of parameters $\lambda$ and $s$, and the temporal structure of the generated segmentation turns out to be more precise in comparison with the AR model (for further details see Appendix C.1).

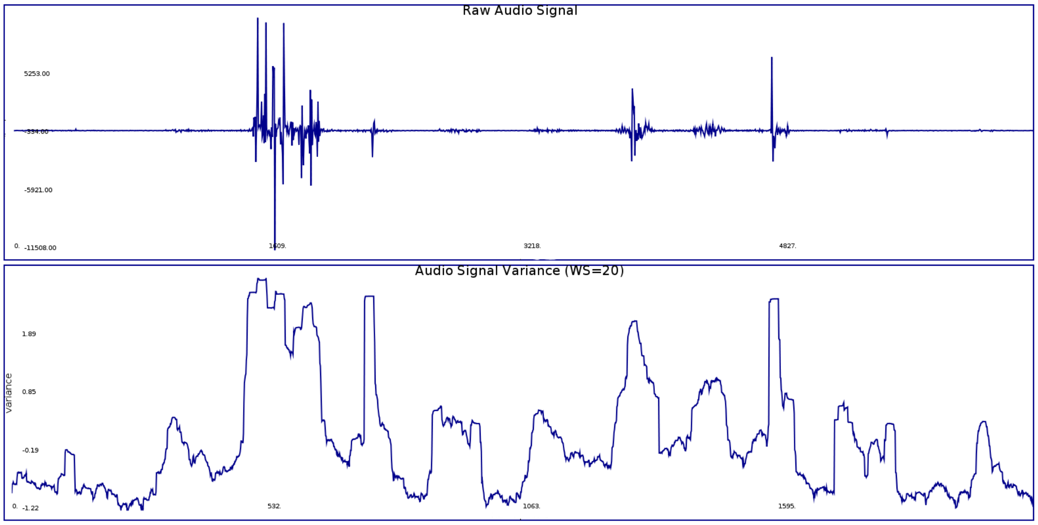


Figure 5.8: Audio preprocessing for constant models. First row: raw signal after subsampling; Second row: variance calculated within a sliding window of width $w = 20$.

Table 5.9: Overview of the parameter influence for audio segmentation with constant models.

| Parameter | Direction of Change | $\mu_t$ | $\mu_g$ | $\mu_r$ | $\mu_m$ |
|---|---|---|---|---|---|
| Subsampling rate $s$ | ↑ | ↑ | ↓ | No sign. effect | ↑ |
| Prior distribution $\lambda$ | ↓ | No sign. effect | ↓ | No sign. effect | ↑ |

#### 5.3.3.3 Audio Modality: Summary

The segmentation of the signal captured with the contact microphone has been evaluated with respect to a set of action primitives, consisting of "grasp", "shake", "put down", "screw", "unscrew", "hold" and "pour".

For both models, the AR and the constant, an increase of $\lambda$ and a decrease of $s$ increase the granularity of the segmentation $\mu_g$ and decrease the missing segments ratio $\mu_m$. An increase of the subsampling rate has an additional negative effect on the temporal error and on the overlap ratio indices. On the other hand, higher values of $s$ reduce the computation time.

For the same values of $\lambda$ and $s$, e.g. $\lambda = 10^{-6}$ and $s = 10$, segmentation with AR models produces larger temporal error, almost two times smaller segmentation granularity, and approx. 20% larger undetection rate in comparison to constant modeling. The experiments have showed that the parameters have a stronger influence on the segmentation with AR-models in comparison to the constant model, which is more robust. The disadvantage of the constant modeling is the need for an extra preprocessing step, the variance calculation.

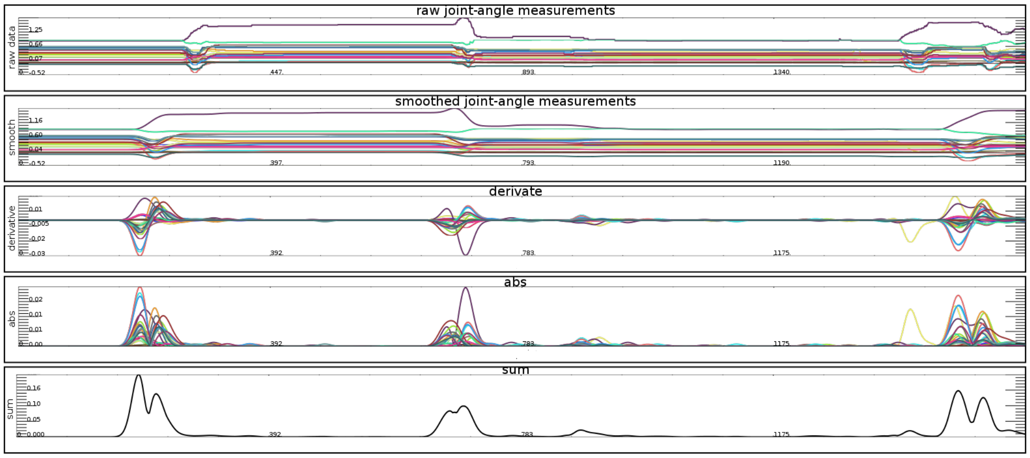


Figure 5.9: Preprocessing of joint-angles. First row: raw signal of all joint-angles (right hand) after subsampling; second row: smoothed signal; third row: temporal derivative of the signal; fourth row: absolute value of the signal; fifth row: summed up signal used as an input to Fearnhead's algorithm.

### 5.3.4 Joint-angles Modality

The joint-angles modality directly correlates with the hand posture. Joint-angles data is recorded by two Immersion CyberGloves and consists of joint-angles of the fingers and the palm for both hands (see Section 3.2.5). We assume that most action primitives involving finger movement, are characterized by an approximately constant *overall level of the finger activity*, a central concept for our approach to the hand posture segmentation. As an example, consider such primitives as *grasping* or *releasing* an object, *screwing* or *unscrewing* a lid. Hence, in our approach to segmentation of the joint-angles, we are looking for constant regions in the overall level of the finger activity (see a detailed description below).

Based on this concept, our preprocessing approach consists in reducing the 24-dimensional joint-angle trajectories of each hand to a scalar time series defined to represent the overall level of finger activity. We calculate it by summing up the absolute values of time derivatives over all dimensions of the hand posture time series $y_{jl}$ and $y_{jr}$. Therefore, the preprocessing consists of the following steps (see Figure 5.9):

1. the raw joint-angles data is subsampled; the subsampling rate is controlled by the parameter $s$;

2. each dimension of the input data is smoothed with Gauss smoothing. This preparatory step aims at improving of the quality of discrete derivative;

3. time derivative for each input dimension is calculated. Third row of the plot shows regions of high levels of finger-specific activity corresponding to grasping and releasing the object. The level of finger activity in the other regions is approximately zero;

4. the absolute values of all joint velocities are accumulated, the resulting trajectory is whitened for normalization.

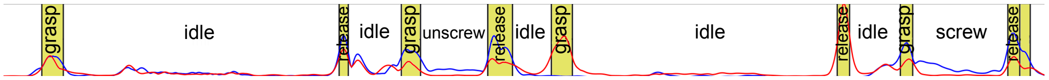


Figure 5.10: An example of a trial segmentation based on preprocessed bimanual joint-angle time series for subsampling rate $s = 10$, $\lambda = 10^{-15}$. The segmentation results in interaction decomposition into segments corresponding to "idle", "release", "grasp", "screw" and "unscrew".

In order to obtain a decomposition into different constant levels of the overall finger activity, we model the preprocessed scalar trajectories of the left and the right hand[3] with constant models denoted by $m_{\mathtt{jl}}$ and $m_{\mathtt{jr}}$ respectively. Our approach to bimanual modeling, similar to the tactile modality, assumes independence of both dimensions and uses a product model as follows:

$$P(y_{t:s}|m_j) = P(y_{t:s}|m_{jl}) \cdot P(y_{t:s}|m_{jr}) \tag{5.15}$$

Figure 5.10 illustrates an example of segmentation carried out on a preprocessed bimanual joint-angles signal using a mixture of constant models. The resulting segmentation is characterized by the regions of high and low overall finger activity for either of the hands. High overall finger activity corresponds to action primitives such as “grasp” or “release” taking place before and after the action on object is conducted. The regions during the grasp itself, i.e. during *shaking* or *pouring*, when hardly any finger dynamics can be observed, are marked by *idle*.

For segmentation evaluation we have used the joint-angles label collection (see Appendix B). A detailed description of the experiments investigating the influence of the global parameters $s$ and $\lambda$ on the generated segmentation can be found in Appendix C.2. The segmentation of the joint-angles modality has been evaluated with respect to action primitives, such as “grasp”, “release”, “screw”, “unscrew” and “idle”. In this paragraph we solely present the qualitative influence of the parameters as showed in Table 5.10.

Table 5.10: Influence of the parameters on the segmentation of bimanual joint-angles data with constant models.

| Parameter | Direction of Change | $\mu_t$ | $\mu_g$ | $\mu_r$ | $\mu_m$ |
|---|---|---|---|---|---|
| Subsampling rate $s$ | ↑ | ↑ | ↓ | ↑ | ↑ |
| Prior distribution $\lambda$ | ↓ | No significant effect | ↓ | ↑ | ↑ |

The table illustrates the common effect of both parameters $\lambda$ and $s$ on the segmentation granularity and $\mu_g$ and the missing segments index $\mu_m$. The decrease of $\lambda$ and the increase of subsampling rate $s$ both decrease the granularity $\mu_g$, increasing the $\mu_r$ and the missing segments index $\mu_m$. Similar to the audio modality, an increase of the value of $s$ has a negative influence on the temporal error $\mu_t$.

[3] In some experiments we use the cumulative trajectory of both hands to represent a cumulative finger activity for both hands.

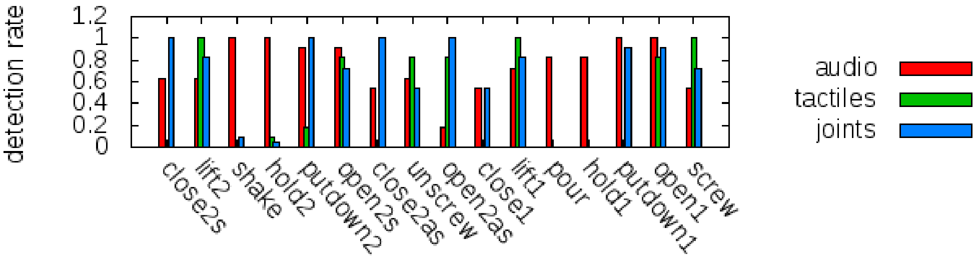


Figure 5.11: Comparison of action-specific detection rates $(1 - \mu_{i,m})$ for tactile (green), audio (red), and joint-angles (blue) modalities.

### 5.3.5 Comparison of Unimodal Segmentations

Based on the results of the previous three sections, here we present a comparison of unimodal segmentations for all three modalities: tactile, audio and joint-angles. The main purpose is to illustrate and discuss the semantic differences and similarities between the unimodal approaches.

We have employed threshold models for the tactile modality, and simple constant models for both, the audio and the joint-angles modalities[4]. The evaluation is based on the semantic label collection generated by manual annotation, containing a superset of all action primitives (see Appendix B). The measure of segmentation quality is based on the action-specific detection rate $(1 - \mu_{i,m})$. For the calculation of $\mu_{i,m}$ we have segmented ten trials of one human demonstrator recorded in a constrained scenario with a constant set of parameters, $\lambda = 10^{-5}$, $s = 8$. Figure 5.11 illustrates the results of the evaluation for unimodal segmentations. Primarily, the figure demonstrates the crucial role of the audio modality for detection of *pouring*, *shaking* and *holding*. Joint-angles modality is essential for detecting action primitives related solely to the dynamics of the hand, i.e. *closing* or *opening* of the hand during grasping and releasing of the object. Corresponding to this, the joint-angles modality along with the tactile modality detects the beginning and the end of an “object contact” region, i.e. *lifting*, *screwing* or *unscrewing*. Due to the fact that the joint-angles dynamics is weak in some trials, the tactile modality is more robust in detection of *lifting* or *unscrewing*.

Altogether, this experiment demonstrates that each modality is particularly suitable for segmentation of a specific subset of action primitives within an interaction episode. This fact is one of the main reasons for a multimodal approach to segmentation aspiring to integrate segmentations generated by unimodal approaches.

## 5.4 Bimodal Segmentation: Hierarchical Approach

Section 5.3 showed that unimodal segmentation approaches tend to be highly suitable only for different subsets of action primitives. In this section we proceed to the first multimodal

[4] During the preprocessing of the joint-angles modality, all existing dimensions have been summed up yielding a scalar time series for both hands.

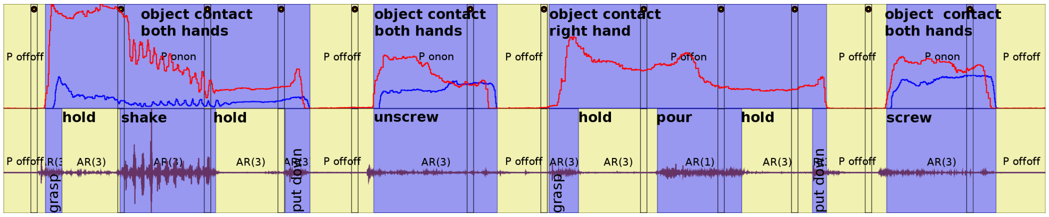


Figure 5.12: An example trial segmentation with two-stage hierarchical method, based on the tactile and the audio modalities. Step 1 (first row): tactile segmentation in "object contact" vs. "no object contact" regions. Step 2 (second row): subsegmentation step based on audio modality.

decomposition approach, the *hierarchical segmentation*. In order to integrate the unimodal segmentations, the hierarchical segmentation performs a series of sequential segmentation steps, aiming to refine the semantic structure of the time series in each of them (see Subsection 4.3.2). The goal of the section is to investigate the integration of segmentations of two semantically very different modalities: the audio and the tactile modalities.

### 5.4.1 Method and Model Overview

For this purpose, given the input time series $y_{1:n}$, the segmentation is first performed on the **tactile** modality $y_{|\mathtt{t}}$, followed by the subsegmentation step conducted on the **audio** modality $y_{|\mathtt{a}}$ (as described in the previous Subsections 5.3.2 and 5.3.3 respectively).

Figure 5.12 illustrates an example of both sequential segmentation steps. The first segmentation step (Figure 5.12 - first row) performs a rough joint analysis of the tactile signals of both hands. This step yields contact assignments identifying parts of the time series that are directly associated with object contact or interaction (see Section 5.3.2). In this step the interaction episode is divided into “no object contact” and “object contact” regions for both hands. The assignment of product models in $\mathcal{M} = \{m_{lr}, m_{LR}, m_{lR}, m_{Lr}\}$ (see Section 5.3.2) to the segments can be exploited for *filtering* and *postprocessing* to exclude joint-angles and tactile modalities (`jl`, `tl` for left hand; `jr`, `tr` for right hand) of “inactive” hands from subsequent processing steps (e.g. clustering, see Section 6.4). For example, the assignment of $m_{lR}$ to a segment $y_{t:s}$ leads to the corresponding data fragment $y_{t:s|\mathtt{jl},\mathtt{tl}}$ being excluded. When the model $m_{lr}$ is assigned, the segment in question can be ignored entirely.

For the application of Fearnhead’s method in this stage, we set the value of the prior parameter $\lambda^{\alpha} = 1/n^{\alpha}$ for a trial $\alpha$ of length $n^{\alpha}$. Although this choice conceptually corresponds to a single expected segment, it turned out to be suitable for small numbers of segments as well. This has been confirmed by the experimental evaluation.

In the subordinate second segmentation step (Figure 5.12 - second row), all segments produced and not discarded in the previous step are sub-segmented. Here the audio signal in the sub-segments is assumed to be produced by autoregressive (AR) models of order 1, 2 or 3: $\mathcal{M}_{\text{sub}} = \{AR(1), AR(2), AR(3)\}$ [27]. Thus, the sub-segmentation is formed by selecting segments that exhibit homogeneous oscillatory properties within the audio modality (see Section 5.3.3.1). In contrast to the procedure outlined in the previous paragraph, the value

Table 5.11: Model overview for two-stage segmentation

| Stage | Mixture model components | # Components | Notation |
|---|---|---|---|
| 1 | product of threshold models | 4 | $m_{lr}$, $m_{LR}$, $m_{lR}$, $m_{Lr}$ |
| 2 | autoregressive models | 3 | AR(1), AR(2), AR(3) |

of the segment length distribution parameter $\lambda_{\text{sub}}$ is fixed and determined empirically.

The sequential application of segmentation and selection steps yields a set of segments that are characterized by constant contact topology in respect to the tactile hand activity as well as homogeneous characteristics of the audio signal. The overview of the models used in both step is displayed in Table 5.11.

Next sections verify the proposed methods by evaluating the quality of segmentation for different types of scenarios, ground truth, and for different human demonstrators. The goal of the experiments is to investigate the generated segmentation with respect to a set of action primitives combining the semantic structure of both, the tactile and the audio modalities.

### 5.4.2 Segmentation of Constrained vs. Unconstrained Trials

In the first experiment we compare the segmentation of constrained vs. unconstrained trials. As previously discussed, constrained trials are recorded with the help of audio cues that are emitted to mark the beginning and/or the end of actions and, therefore, control the action execution speed. Unconstrained trials are recorded with natural speed. After the discussion in Subsection 3.5, we may assume that the main difference between the two trial types is the average length of action primitives, estimated to be approximately two times higher in a constrained scenario in comparison to an unconstrained scenario.

By comparing the generated segmentation for both types of trials, we mainly aim to investigate the impact of the length of action primitives on the resulting segmentation In order to make the segmentations of both trial types comparable, the same set of parameters has been used in both cases. An outline of the parameters can be found in Table 5.12.

Figure 5.13 presents four histograms illustrating the quantitative results of the experiment. The plots display action-specific segmentation indices $\mu_{t,i}$, $\mu_{g,i}$, $\mu_{r,i}$ and $\mu_{m,i}$ (see Section 5.2) in the constrained and unconstrained scenario with the corresponding variances. All four sub-figures show a great similarity in comparison of constrained vs. unconstrained segmentation indices. The first histogram illustrates the temporal error $\mu_{t,i}$ (see Figure 5.13, first row - left), ranging from ca. 0.1 to 0.25 seconds and shows a high accuracy of the generated segmentation w.r.t. the ground truth. The variation $\sigma_{t,i}$ is negligibly small and does not reach values higher than 0.02.

The segmentation granularity histogram in Figure 5.13 (first row - right) shows that the index values are comparable in both cases, apart from *grasp+lift1* and *grasp+lift2*, which show an oversegmentation of c.a. 2.5. This can be explained by the structural difference in both trial types: in the unconstrained case, the *grasping* is followed directly by shaking. In the constrained case there is a pause following *grasping*, before *shaking* or *pouring* starts. Thus, an additional segment is generated in these cases. Because both types of trials have the same annotation structure, higher segmentation granularity index values are calculated

Table 5.12: Parameter overview for constrained and unconstrained segmentations.

| Parameter | Value |
|---|---|
| tactile subsampling rate $s$ | 5 |
| threshold $\gamma$ | 15 |
| tactile $\lambda$ | $1/n^{\alpha}$ |
| audio subsampling rate $s$ | 20 |
| audio range $\rho$ | 10 |
| audio $\lambda$ | $10^{-4}$ |
| number of constrained trials | 40 |
| number of unconstrained trials | 40 |
| number of HDs | 3 |
| ground truth type | annotation |
| label collection | cues |
| segmentation method | hierarchical |

for the constrained case. Note that both action primitives *put down 1* and *put down 2* have a comparably high segmentation granularity value in both setups. This is due to the fact that in the annotation *put down* marks only the beginning of the action primitive. Both above-mentioned cases of oversegmentation can be seen on an example of Figure 5.12 in the beginning of the section. For *put down* two change points corresponding to the structure of the audio signal are generated; similarly, corresponding to the rise and fall of the audio volume, two segments are generated after the beginning of grasping in both cases, while only one segment border is present in the annotation.

The overlap ratio histogram Figure 5.13 (second row - left) shows that the values of the overlap index are very close for the constrained and the unconstrained scenarios. Exceptions are *grasp+lift1* and *grasp+lift2* already discussed in the previous paragraph. These actions differ in both scenarios, due to an extra generated segment in the constrained scenario. Following this, the overlap ratio in the constrained scenario is small, while in the unconstrained scenario it is large in the case of this action primitive. Strong oversegmentation in the segmentation granularity diagram correlates with low overlap ratio in the segment overlap ratio histogram. *Put down* exhibits small overlap in both scenarios. The reason is the same as for the high oversegmentation: the ground truth annotation only marks the starting point of this point-event, therefore the end of the search interval for *put down* ends with the beginning of the next action, i.e. *screw* or *unscrew*.

Figure 5.13 (second row - right) compares the missing segment index in both scenarios. The histogram shows how many segment borders present in ground truth are missing in the generated segmentation. In case of *hold1* and *hold2* the ratio is 0.25 in the unconstrained scenario, corresponding to 25 percent missing segment borders. We explain this by still present audio signal after the actions *shake* and *pour* have been conducted and before the object has been *put down*. Thus the data has not been recognized as audio pause and no corresponding segment has been generated. In the case of the unconstrained scenario only for *pour* 5% of segments corresponding to this action primitive have not been detected.

Altogether, the above experiments have demonstrated satisfactory results, showing that,

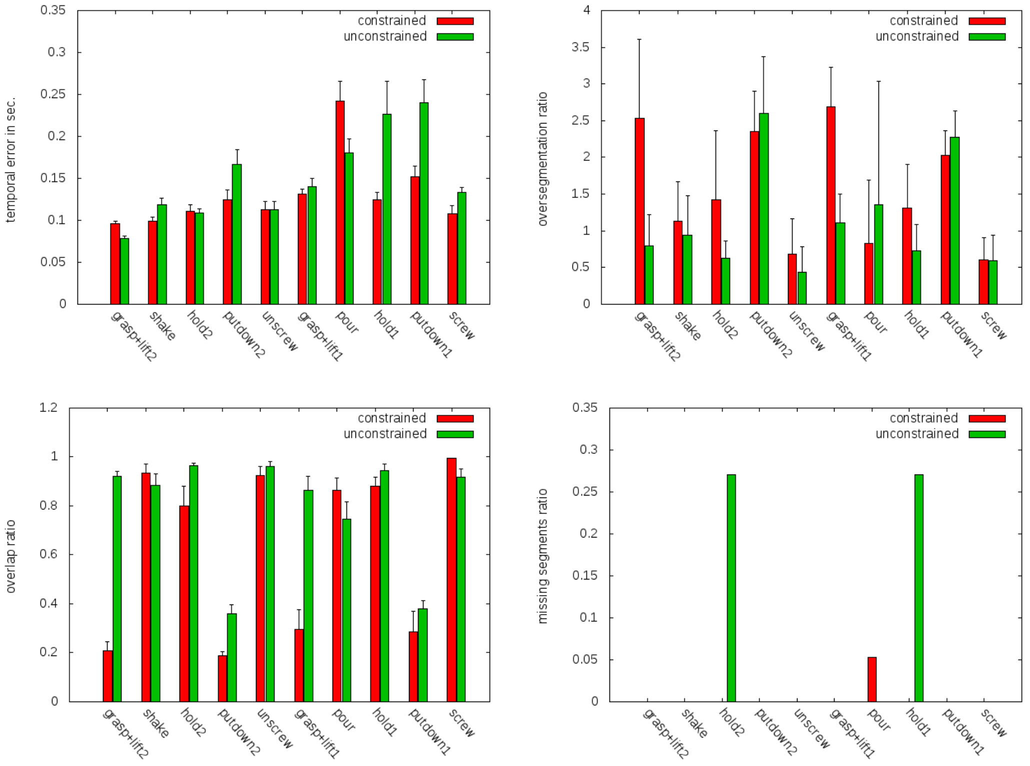


Figure 5.13: Comparison of segmentation results for constrained (red) and unconstrained (green) scenarios; bars indicate averages built over all available constrained and unconstrained trials respectively recorded by three human demonstrators; error bars indicate the corresponding variances. First row: action-specific temporal error $\mu_{t,i}$ (left) and action-specific segmentation granularity $\mu_{g,i}$ (right); second row: action-specific overlap ratio $\mu_{r,i}$ (left) and action-specific missing segments $\mu_{m,i}$ (right).

despite a large difference in length of the action primitives, hierarchical approach produced comparable segmentation in both scenarios. This can be explained by a stronger influence of the model likelihood vs. the prior determined by the parameter $\lambda$ on the segmentation. Temporal error in both scenarios is on a very low level of 0.1 to 0.25 seconds. We further argue that the action primitives characterized by high oversegmentation and low overlap ratio partially result from the incompleteness of the ground truth annotation, e.g. in the cases of *put down* or *grasp*. The ratio of detected segments in both scenarios is similar, apart from one type of action primitive, *holding* ($\mu_m \approx 0.25$) and *pouring* ($\mu_m \approx 0.05$) in the unconstrained and the constrained scenarios respectively. The first result may be explained by a slightly higher semantic granularity of the constrained scenario w.r.t. the ground truth.

### 5.4.3 Manual Annotation vs. Cue-based Ground Truth

The goal of this experiment is to compare the results of the segmentation evaluation based on two types of ground truth: manual annotation vs. automated cue-based ground truth. As previously mentioned, video-based manual annotation of the trials is time consuming, while cue-based ground truth can be generated automatically. Due to this advantage, in this experiment we want to examine, how well the cue-based ground truth performs, in comparison to the traditional video-based annotation method.

For this purpose, we evaluate segmentation of the same set of trials with both, the cue-based and the annotation ground truth. Due to the fact that a human demonstrator aligns her or his action to the cues during the cue-triggered recording, we expect higher levels of temporal error and higher levels of undetected segment borders for the cue-based evaluation method vs. annotation. Table 5.13 presents an overview of the parameters and the experimental data pool.

Table 5.13: Overview of experiment comparing segmentation evalution with hand-labeled annotation vs. automatically acquired ground truth.

| Parameter | Value |
|---|---:|
| tactile subsampling rate $s$ | 5 |
| threshold $\gamma$ | 15 |
| audio subsampling rate $s$ | 20 |
| audio range $\rho$ | 10 |
| prior distribution $\lambda$ | $10^{-4}$ |
| number of HDs | 3 |
| number of trials | 40 |
| ground truth type 1 | audio cues |
| ground truth type 2 | label collection cues |
| scenario | constrained |
| segmentation method | hierarchical |

Similar to the previous section, Figure 5.14 shows histograms comparing four action-specific segmentation indices produced by the evaluation with both ground truth types. The temporal error histogram (see Figure 5.14, first row - left) shows that in the automated case the error ranging from 0.3 to 0.55 seconds is considerably bigger than in the annotation case, ranging from 0.1 to 0.25 seconds. The difference of c.a. 0.3 seconds is attributed to the timing error of the HDs when aligning their actions to the emitted audio cues. The variance of the temporal error in the automated case is also larger than the one in the annotated case. This is due to the fact that timing precision varies across HDs. Figure 5.14 (first row - right) shows a histogram comparing the segmentation granularity. In both cases the action-specific segmentation granularity index is very similar. The same applies to the overlap ratios (see Figure 5.14, second row - left). The missing segments histogram is depicted in the second row on the right. Higher level of undetected change points is showed for the automated case, implying that the generated segment borders and the corresponding cues lie too far apart and this distance is larger than the $\epsilon$ defining the search range. *Hold1* has a particularly high level of undetected segments, which is probably caused by the fact

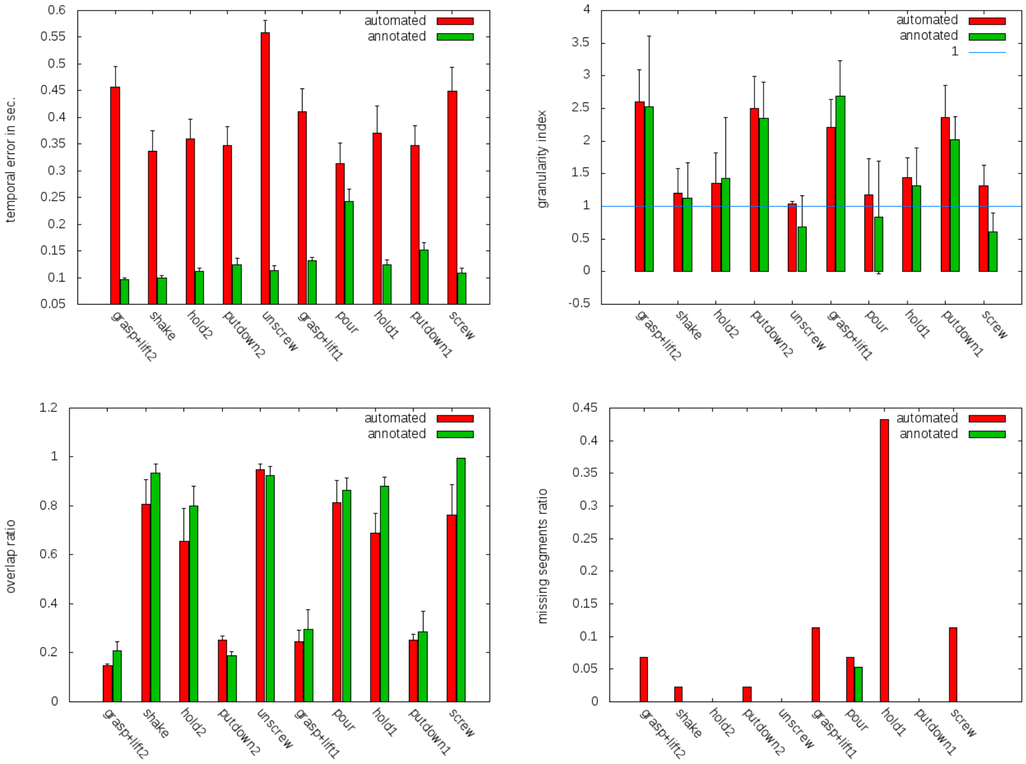


Figure 5.14: Comparison of evaluation results produced from segmentation evaluation with hand-labeled vs. automatically generated ground truth; Red bars represent evaluation with automated ground truth and green bars - with annotation. First row: action-specific temporal error $\mu_{t,i}$ (left), action-specific segmentation granularity $\mu_{g,i}$ (right); Second row: action-specific overlap ratio $\mu_{r,i}$ (left), action-specific missing segments index $\mu_{m,i}$ (right).

that it is difficult to align the end of pouring to the finishing cue. Therefore the generated change points lie too far away from the audio cue and are not detected within the small $\epsilon$-environment of the search range.

Altogether, despite the difference between the temporal error indices and, in some cases, a higher undetection rate $\mu_{m,i}$, we can recommend the usage of the cue-based ground truth for the segmentation evaluation. As a side effect of the evaluation, we have received an approximate estimation of the average alignment error for each action primitive (see Figure 5.14 - first row). It ranges from 0.1 to 0.5 second.

### 5.4.4 Segmentation Evaluation for Three Human Demonstrators

The goal of this experiment is to compare the segmentation results produced for different human demonstrators, and to examine whether the segmentation is invariant to different HDs. For this purpose we have generated segmentations for trials captured by three HDs.

Table 5.14 shows an overview of the experiment and the segmentation parameters.

Table 5.14: Overview of experiment comparing segmentation quality for three different HDs.

| Parameter | Value |
|---|---:|
| tactile subsampling rate $s$ | 5 |
| threshold $\gamma$ | 15 |
| audio subsampling rate $s$ | 20 |
| audio range $\rho$ | 10 |
| prior distribution $\lambda$ | $10^{-4}$ |
| number of HDs | 3 |
| number of constrained trials | 40 |
| ground truth type | annotation |
| label collection | cues |
| segmentation method | hierarchical |

Figure 5.15 shows histograms comparing the individual segmentation indices. All histograms show comparable segmentation for the three human demonstrators. The first histogram (first row - left) shows similar action-specific temporal error levels. For the $\mathtt{hd}_1$ in some cases the error is c.a 0.1 seconds less than for $\mathtt{hd}_2$ and $\mathtt{hd}_3$. No explanation could be found for this result. The variance is negligibly small, apart from *pour*, where for $\mathtt{hd}_2$ it reaches $\approx 0.03$. The other action-specific temporal errors lie below 0.2 seconds. The second histogram in the first row compares the segmentation granularity index for the three HDs. This histogram, as well as the one illustrating the overlap ratio (second row left), show comparable results for all three HDs. The fourth histogram depicts the missing segments index. Only for *pour* the missing segment index is about 0.1 for HD2 and HD3.

Altogether, for all three tested human demonstrators, despite the interpersonal variance, the generated segmentations yielded comparable structural and temporal quality.

## 5.5 Multimodal Segmentation: Parallel Approach

In this section we empirically investigate the second multimodal approach, the parallel segmentation (described in Section 4.3.3). The goal of the experiments is to integrate segmentation for all three recorded modalities. Before the main experimental part, in the following subsection we first outline the employed models and parameters, and demonstrate the resulting segmentation on several examples.

### 5.5.1 Method and Model Overview

As previously discussed, the starting point for the multimodal approach is a choice of semantics and a corresponding simple model for each modality. Based on the results of the previous unimodal studies (see Section 5.3), we have chosen to employ two kinds of simple models, constant and threshold, for the modeling of the following four channels: tactile - left hand, tactile - right hand, audio and joint-angles. An outline of unimodal assignments can be found in Table 5.15.

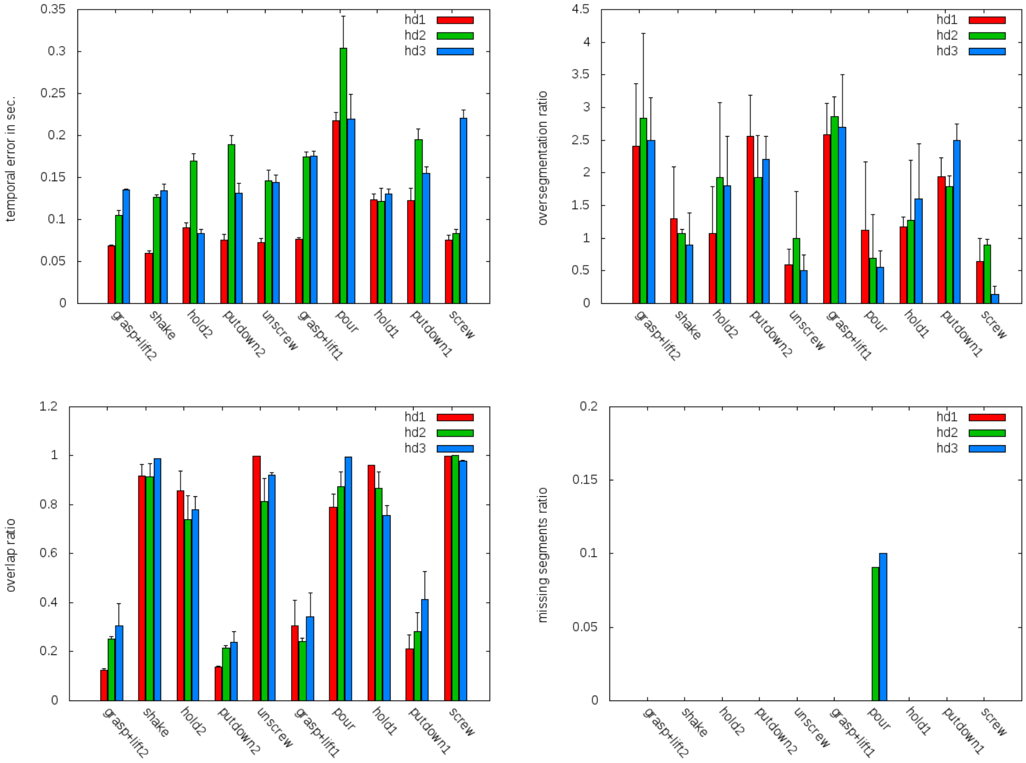


Figure 5.15: Comparison of segmentation quality for different HDs: $\texttt{hd}_1$ (red), $\texttt{hd}_2$ (green) and $\texttt{hd}_3$ (blue). First row: action-specific temporal error $\mu_{t,i}$ (left), action-specific segmentation granularity $\mu_{g,i}$ (right); second row: action-specific overlap ratio $\mu_{r,i}$ (left), action-specific missing segments index $\mu_{m,i}$ (right). Averages are built over all trials available for the corresponding HD.

Table 5.15: Overview of the model asignments.

| Sensor channel | Model | Notation |
|---|---|---|
| tactile sum - left hand | threshold | $m_l$, $m_L$ |
| tactile sum - right hand | threshold | $m_r$, $m_R$ |
| audio | constant | $m_a$ |
| overall activity for both hands | constant | $m_j$ |

In order to realize multimodal integration, the four channel-specific models are incorporated in a product model. This can be viewed as an extension of the bimanual tactile product model (see Section 4.3.1) with an audio and a joint-angle component. According to the modality-modal assignment in the table above, each product model consists of two threshold and two constant models to accommodate the tactile (left and right hand), the audio, and the joint-angles modalities. All possible assignment combinations yield the following four product models:

| Notation | Components |
|---|---|
| $m_{lr}$ | $m_l$, $m_r$, $m_a$, $m_j$ |
| $m_{LR}$ | $m_L$, $m_R$, $m_a$, $m_j$ |
| $m_{Lr}$ | $m_L$, $m_r$, $m_a$, $m_j$ |
| $m_{lR}$ | $m_l$, $m_R$, $m_a$, $m_j$ |

To control the influence of each modality on the joint likelihood, and to influence the value of likelihood itself (in comparison to the prior), we employ a weighted product model, in which the sum of the weights is not equal to 1. The likelihood of e.g. $m_{lR}$ is calculated as follows:

$$P(y_{t:s} \mid m_{lR}) = P(y_{t:s} \mid m_l)^{w_t} \cdot P(y_{t:s} \mid m_R)^{w_t} \cdot P(y_{t:s} \mid m_j)^{w_j} \cdot P(y_{t:s} \mid m_a)^{w_a}.$$

Here the weight vector is denoted by $(w_t, w_a, w_j)$ with tactile ($w_t$), audio ($w_a$), and joint-angles ($w_j$).

Finally, as an input to Fearnhead's algorithm servers a mixture of the above four product models:

$$P(y_{t:s}) = 1/4P(y_{t:s}|m_{lr}) + 1/4P(y_{t:s}|m_{lR}) + 1/4P(y_{t:s}|m_{Lr}) + 1/4P(y_{t:s}|m_{LR}), \quad (5.16)$$

Examples in Figure 5.16 illustrate the impact of the weight vector $(w_t, w_a, w_j)$ on the segmentation. The figure shows segmentations for the following weight combinations:

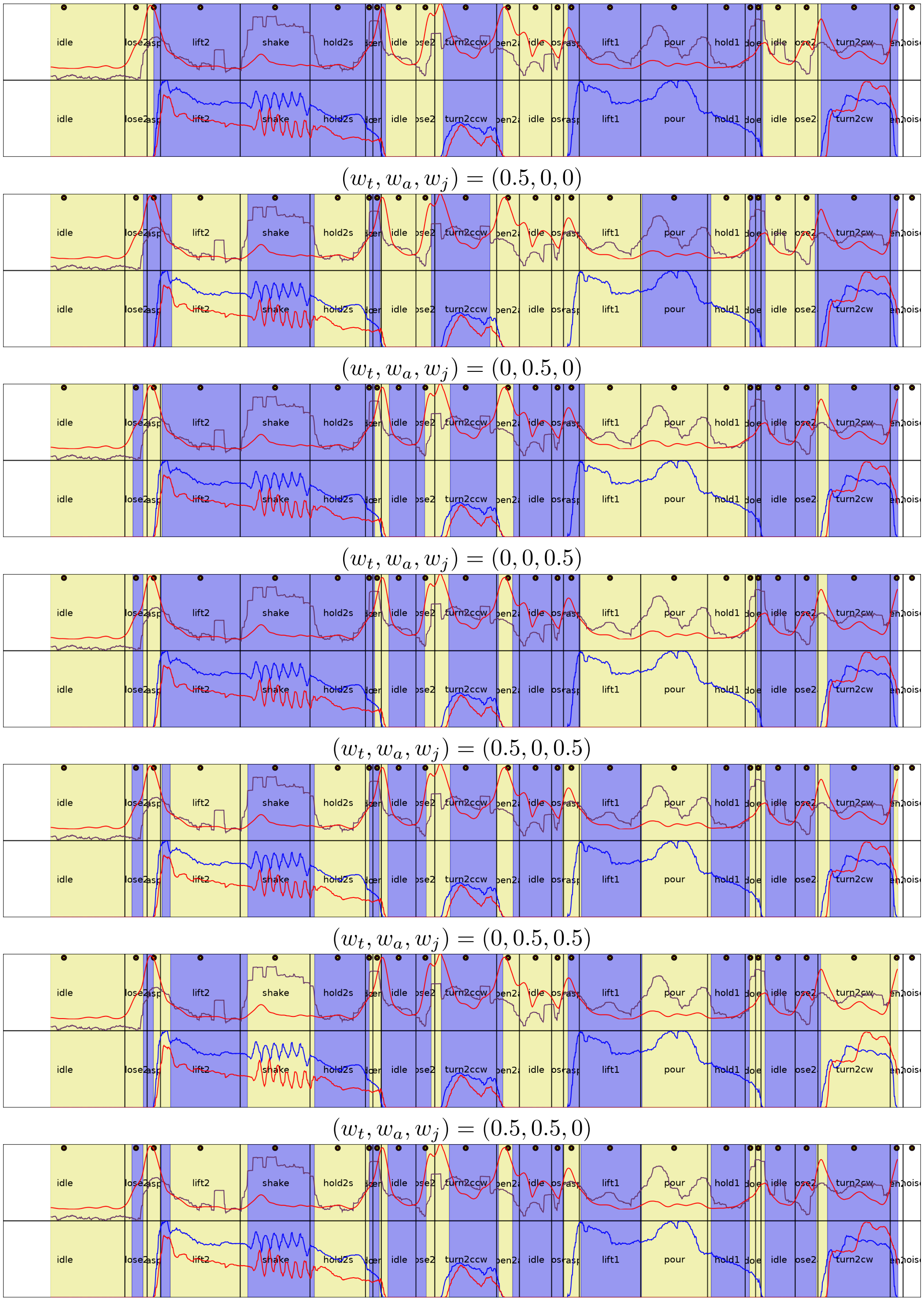
$$(w_t, w_a, w_j) \in \{0, 0.5\}^3 \backslash \{(0, 0, 0)\}.$$

In the case a weight is set to zero, the likelihood of this modality does not affect the product model likelihood. In the following enumeration we describe each weight combination (see the corresponding rows of Figure 5.16):

1. $(w_t, w_a, w_j) = (0.5, 0, 0)$ (first row): segmentation based only on tactile modality resulting in “object contact” and “no object contact” regions (blue and yellow regions respectively).

2. $(w_t, w_a, w_j) = (0, 0.5, 0)$ (second row): segmentation based only on audio channel selects regions of homogeneous amplitude. The generated segments of high amplitude are colored in blue: *grasping*, *shaking*, *putting down*, *screwing*, *pouring*, and *unscrewing*. Yellow regions correspond in this figure to low amplitude: *holding* and *idle*.

3. $(w_t, w_a, w_j) = (0, 0, 0.5)$ (third row): segmentation based on joint-angle modality. The sub-figure shows the regions of high and low overall hand activity. Regions of high level of hand activity (correspond to yellow color) are typically generated before and after an object manipulation, i.e. *grasping* and *releasing* an object.



$(w_t, w_a, w_j) = (0.5, 0, 0)$

$(w_t, w_a, w_j) = (0, 0.5, 0)$

$(w_t, w_a, w_j) = (0, 0, 0.5)$

$(w_t, w_a, w_j) = (0.5, 0, 0.5)$

$(w_t, w_a, w_j) = (0, 0.5, 0.5)$

$(w_t, w_a, w_j) = (0.5, 0.5, 0)$

$(w_t, w_a, w_j) = (0.5, 0.5, 0.5)$

Figure 5.16: An example segmentation for different weight combinations; $\lambda = 10^{-5}$, $s = 15$.

4. $(w_t, w_a, w_j) = (0.5, 0, 0.5)$ (fourth row): segmentation of tactile and joint-angles modality. The differences between the fourth and the third row are very subtle. The impact of the tactile modality can be seen in the next to last blue region: its left border has slightly moved to the right, where it better (compared to row 3) fits the beginning of "no object contact" region.

5. $(w_t, w_a, w_j) = (0, 0.5, 0.5)$ (fifth row): segmentation based on a combination of joint-angles and audio modality. In this sub-figure, the change points combine the set generated by audio and the set generated by joint-angles.

6. $(w_t, w_a, w_j) = (0.5, 0.5, 0)$ (sixth row): segmentation based on audio and tactile modalities. This example shows a combination of audio and tactile change points.

7. $(w_t, w_a, w_j) = (0.5, 0.5, 0.5)$ (seventh row): in this example the segmentation resulting for $(w_t, w_a, w_j) = (0.5, 0.5, 0.5)$ is coincidentally equal to the one resulting for the combination $(w_t, w_a, w_j) = (0, 0.5, 0.5)$.

The segmentations presented in Figure 5.16 show that the weight combinations in this example form equivalence classes. One class is $\{(0, 0, 0.5), (0.5, 0, 0.5)\}$, another is $\{(0.5, 0.5, 0.5), (0, 0.5, 0.5)\}$.

The following subsections contain empirical studies investigating the multimodal segmentation generated by the parallel approach. Subsections 5.5.2 and 5.5.3 are dedicated to examining the influence of the granularity parameter $\lambda$ and modality weighting vector $(w_t, w_a, w_j)$ on the multimodal segmentation. Subsection 5.5.4 compares the segmentation generated for constrained and unconstrained trials for a fixed set of parameters.

### 5.5.2 Parameter Influence: Granularity and Modality Weighting

In this experiment we conduct a systematic study of the influence of two main parameters: prior on segment lengths $\lambda$ (short: granularity), and the modality weighting vector $(w_t, w_a, w_j)$. Our goal is to discuss the influence of the parameters on the resulting segmentation, and to determine a range for which an application of the parallel approach results in a semantic segmentation, integrating all three unimodal segmentations.

For this purpose we conduct a grid search for parameters $w_i, i \in \{a, j, t\}$ with $w_i \in \{1/10, 2/10, \ldots, 1\}$ and $\lambda \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-8}\}$ based on ten trials. The value of the subsampling rate parameter $s$ remains fixed with $s = 7$. Note that in contrast to the hierarchical approach, within the parallel approach one value of $\lambda$ has to be applied to all modalities at once. The following table shows an overview of the experiment:

| Parameter | Value |
|---|---|
| $w_i$ | $\{1/10, 2/10, \ldots, 1\}$ |
| $\lambda$ | $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-8}\}$ |
| ground truth type | annotation |
| label collection | semantic |
| number of trials | 10 |
| number of HDs | 1 |
| scenario | unconstrained |
| segmentation approach | parallel |

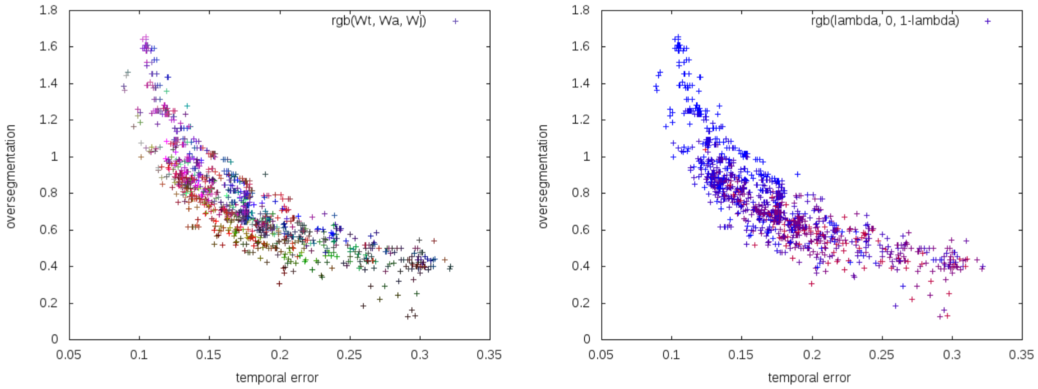


Figure 5.17: Dependency between segmentation granularity index $\mu_g$ and temporal error $\mu_t$ for different combinations of weight parameters $(w_t, w_a, w_j)$ and $\lambda$; averages $\mu_t$ and $\mu_g$ are built over 10 constrained trials and over all actions. Color encodes the values of the weight parameters: $(w_t, w_a, w_j)$ corresponds to (R,G,B) (left); strength of red and blue color encode the value of the parameter $\lambda$ (right).

Figures 5.17 - 5.21 present selected experimental results. Each figure contains two sub-figures with a different color coding of the same data. The left sub-figure presents the following color coding of the weight combination:

- red channel: tactile weight $w_t$ (more red corresponds to higher weight $w_t$)
- green channel: audio weight $w_a$ (more green corresponds to higher weight of $w_a$)
- blue channel: joint-angles weight $w_j$ (more blue corresponds to higher weight of $w_j$).

The sub-figure on the right color-codes parameter $\lambda$, whose value is used to interpolate between *red* and *blue*. Larger values of $\lambda$ correspond to a stronger blue component, smaller values of $\lambda$ correspond to a stronger red component. Individual figures are discussed in detail in the following paragraphs.

Figure 5.17 presents the dependency between the temporal error index $\mu_t$ and the segmentation granularity index $\mu_g$ for different values of $\lambda$ and different weight combinations. The left sub-figure shows that oversegmentation ($\mu_g > 1$) corresponds to larger values of all weight components identifiable through a lighter point color. The figure shows that a high ratio of red and blue (tactile and joint-angles) yields segmentation granularity close to 1. Green color (largely audio) corresponds to undersegmentation with $\mu_g \approx 0.5$. On both ends of the point cloud there is a saturation effect: temporal error $\mu_t$ stays on the constant level of approx. 0.1 seconds and does not fall further; segmentation granularity $\mu_g$ does not fall beneath the level of approx. 0.3 for the tested weight combinations. The right side of Figure 5.17 shows the same dependency between the segmentation indices with color encoding of $\lambda$. The figure shows that larger values of $\lambda$ correspond to large values of segmentation granularity and small temporal error.

Figure 5.18 illustrates the dependency between temporal error $\mu_t$ and the overlap ratio $\mu_r$. There is a clear positive correlation between the overlap ratio and the temporal error

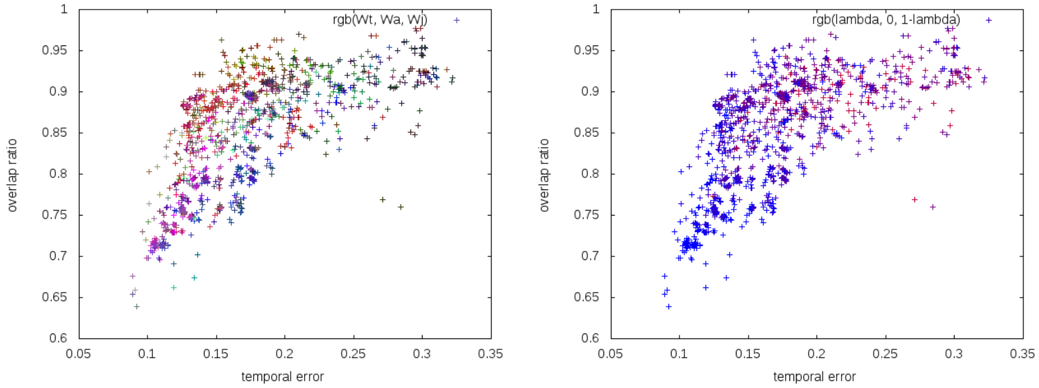


Figure 5.18: Dependency between temporal error $\mu_t$ and overlap ratio $\mu_r$ for different combinations of weight parameters $(w_t, w_a, w_j)$ and $\lambda$; averages $\mu_t$ and $\mu_g$ are built over 10 constrained trials and over all actions. Color encodes the values of the weight parameters $(w_t, w_a, w_j)$ (left); strength of red and blue color encodes the value of the parameter $\lambda$ (right).

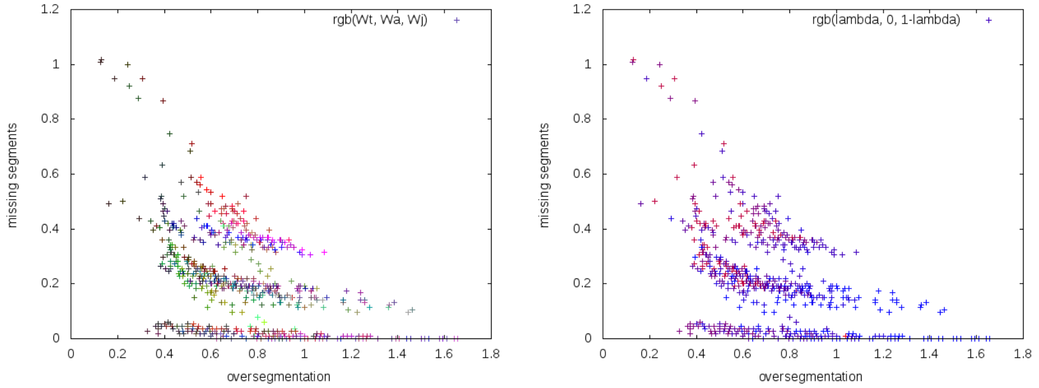


Figure 5.19: Dependency between segmentation granularity $\mu_g$ and the missing segments index $\mu_m$ for different combinations of weight parameters $(w_t, w_a, w_j)$ and $\lambda$. Averages $\mu_g$ and $\mu_m$ are built over 10 constrained trials and over all actions. Color encodes the values of the weight parameters $(w_t, w_a, w_j)$ (left); strength of red and blue color encodes the value of the parameter $\lambda$ (right).

within the left half of the plot corresponding to temporal error $\mu_t < 0.2$. For further increase of the temporal error there is no clear correlation with the overlap ratio. The color clusters in the left sub-figure are similar to the ones in the previous figure: a high ratio of blue (joint-angles) corresponds to the lower overlap and lower temporal error. Green (audio) and brown (mixture of all three) correspond to the highest level of the overlap ratio. The right sub-figure shows that both, small temporal error and small overlap ratio correspond to a larger values of the parameter $\lambda$.

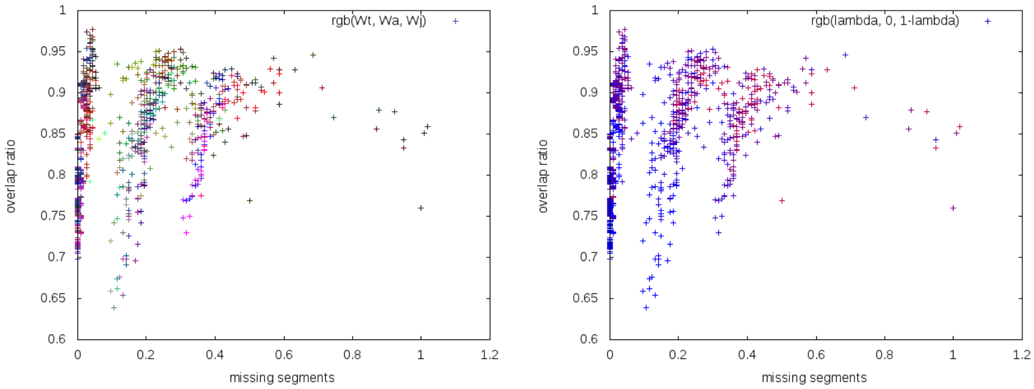


Figure 5.20: Dependency between missing segments rate $\mu_m$ and the overlap ratio $\mu_r$. Averages $\mu_m$ and $\mu_r$ are built over 10 constrained trials and over all actions. Color encodes the values of the weight parameters $(w_t, w_a, w_j)$ (left); strength of red and blue color encodes the value of the parameter $\lambda$ (right).

Figure 5.19 shows the dependency between segmentation granularity index $\mu_g$ and missing segments index $\mu_m$. We can approximately isolate the red-blue, the green-blue and the mixed clusters in the left part of the figure. We explain the existence of such clusters by the type of action primitives containing in the trial. By considering one modality, i.e. audio, only the modality-specific action primitives (i.e. *pouring*) get detected, independently of how large the weight of this component is or the corresponding value of $\lambda$. The clusters stretch from $\mu_g \approx 0.4$ to $\mu_g \approx 1.6$. The red-blue cluster corresponds to the highest rate of missing segments. This implies that considering either the *tactile* modality (red) or the *joint-angle* modality (blue) yields a high level of missing segments ratio of approx. 0.4 to 0.6 corresponding to an segmentation granularity index between 0.6 and 1. The second cluster is dominated by green and brown, corresponding to a strong influence of *audio* and *tactile* modalities within the weight combination and a small contribution of the *joint-angles* modality (blue). This cluster corresponds to a smaller level of $\mu_m$ from approx. 0.1 to slightly below 0.4. The part of the cluster where green prevails corresponds to the smallest segmentation granularity $\mu_g \approx 0.4$. This means that with a large influence of *audio* and a comparatively small influence of other modalities we achieve segmentation where at least 30% of the segment borders are missing. The third cluster is a mixed-color cluster containing combinations, where all three weights have positive values. The further on the right the points lie, the lighter are the corresponding point colors, implying that the contribution of each modality is increasing. The right part of Figure 5.19 illustrates that the color changes from red to blue corresponding to increase of the value of $\lambda$ from left to right. The sub-figures demonstrate that both, higher values of weights as well as larger $\lambda$ increase the granularity of segmentation.

Figure 5.20 shows the dependency between the missing segments index $\mu_m$ and overlap ratio $\mu_r$. The structure of the dependency as well as the coloring of the clusters is similar to the previous figure. The combination of the highest overlap ratio with the lowest missing segments index is reached within a mixed-color cluster. The right sub-figure shows that

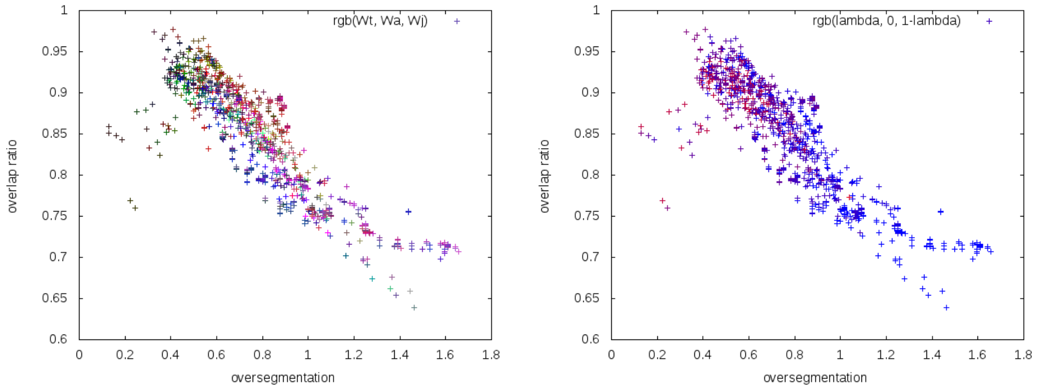


Figure 5.21: Dependency between overlap ratio $\mu_r$ and the segmentation granularity $\mu_g$. Averages $\mu_r$ and $\mu_g$ are built over 10 constrained trials and over all actions. Color encodes the values of the weight parameters $(w_t, w_a, w_j)$ (left); strength of red and blue color encodes the value of the parameter $\lambda$ (right).

larger values of $\lambda$ correspond to small overlap ratio.

Finally, Figure 5.21 shows the negative dependency between $\mu_r$ and $\mu_g$. From left to right, with growing value of segmentation granularity and with falling overlap ratio the points in the point cloud become lighter, meaning that the values of weights for all modalities are high in the corresponding combinations. Similar to previous figures, green (segmentation based on audio) corresponds to undersegmentation and high overlap ratio, while blue and red (tactile and joint modality) correspond to higher values of oversegmentation and lower overlap ratio. In the right sub-figure we see that larger values of $\lambda$ correspond to large values of $\mu_g$ and small values of $\mu_r$.

Figure 5.22 illustrates the range of influence of parameter $\lambda$ on the segmentation on the example of three values: $\lambda \in \{10^{-4}, 10^{-6}, 10^{-8}\}$. As previously showed, larger values of $\lambda$ correspond to larger segmentation granularity and smaller overlap ratio. The figure illustrates that the point clouds corresponding to different values of $\lambda$ form two-dimensional clusters. Altogether, for increasing $\lambda$ the figures illustrate the following:

$$\lambda \uparrow \;\; \Leftrightarrow \;\; \mu_g \uparrow, \mu_t \downarrow, \mu_r \downarrow, \mu_m \downarrow . \tag{5.17}$$

The left sub-figures in the upper row shows that a good rate of granularity index close to one can be primarily achieved by large values of $\lambda$. At the same time large values of $\lambda$ correspond to smaller overlap ratio, illustrated in the right figure of the upper row. Both sub-figures in the bottom row show that independent of the weight combination, for small value of $\lambda = 10^{-8}$ it is not possible to reach a value close to zero for the missing segments index.

Altogether, the above experiments have provided us with a basis for the choice of values for the central parameters $(w_t, w_a, w_j)$ and the prior length parameter $\lambda$. Firstly, it has been showed that in comparison with other modalities, the joint-angles modality tends to stronger oversegmentation. To avoid oversegmentation $\lambda$ should be set to a value not larger than $10^{-3}$, and the corresponding weight $w_j$ to a value $\leq 0.5$. Secondly, a high

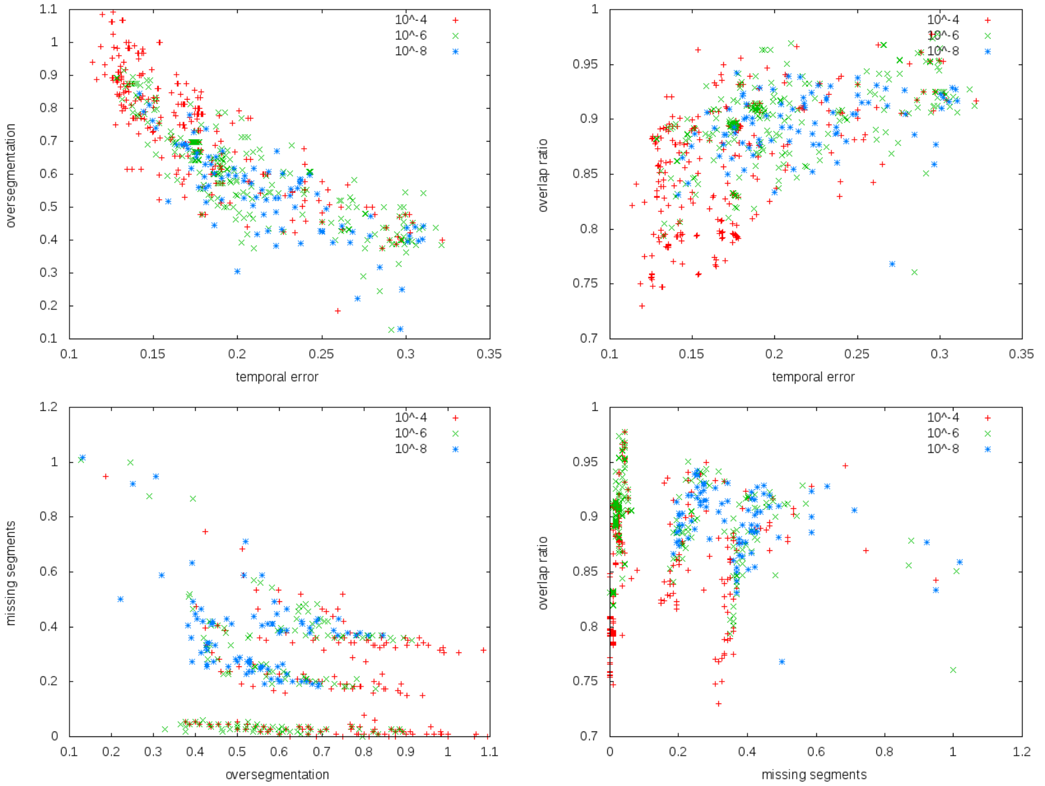


Figure 5.22: Dependency between segmentation indices for different values of $\lambda$. Color encodes three different values of $\lambda$: $10^{-4}$ (red), $10^{-6}$ (green), and $10^{-8}$ (blue). The dependency between temporal error $\mu_t$ and $\mu_g$ (first row - left); dependency between temporal error $\mu_t$ and the overlap ratio $\mu_g$ (first row - right); dependency between segmentation granularity $\mu_g$ and missing segments $\mu_m$ (second row - left); dependency between missing segments $\mu_m$ and overlap ratio $\mu_r$ (second row - right).

missing segments index is associated with segmentation based on single modalities. An appropriate granularity close to one, along with a low missing segments index close to zero have been achieved by considering all three modalities. In this case, in order not to yield oversegmentation, the sum of weights $w_a + w_j + w_t$ should not exceed 1.5. At the same time, the value of the parameter $\lambda$ should be chosen from the range $[10^{-3}, 10^{-7}]$. Thirdly, we assume that the temporal error ranging from 0.1 to 0.3 seconds is negligible and therefore, does not need to be considered during the choice of the parameter values.

### 5.5.3 Parameter Influence: Constrained vs. Unconstrained Scenario

Similar to the previous subsection, in this subsection we compare segmentation results for constrained vs. unconstrained scenarios. In comparison with the constrained trials, the unconstrained trials are characterized by approximately two times smaller length of action primitives on average, as well as a smaller structural granularity. Hence, the goal of this experiment is to investigate how these two properties influence the resulting segmentation. We believe that due to the greater length and semantic granularity, for a given weight vector $(w_t, w_a, w_j)$ and a fixed value of $\lambda$, the segmentations generated for constrained trials will exhibit a higher granularity and a lower overlap ratio, in comparison to the unconstrained trials. An overview of the experiment is presented in the following table:

| Parameter | Value |
|---|---|
| $w_i$ | $\{1/10, 2/10, \dots, 1\}$ |
| $\lambda$ | $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-8}\}$ |
| ground truth type | annotation |
| label collection | semantic |
| number of unconstrained trials | 10 |
| number of constrained trials | 10 |
| number of HDs | 1 |
| scenario | constrained and unconstrained |
| segmentation approach | parallel |

Figure 5.23 illustrates the segmentation results based on the segmentation indices. Primarily, all four sub-figures show that the correlation between the segmentation indices in both scenarios have a similar structure. However, the point clouds of both scenarios differ by a two-dimensional offset. The left sub-figure in the first row illustrates that for the same value of $\mu_t$, the constrained trials have a higher segmentation granularity. The right sub-figure in the first row shows that the overlap ratio in the constrained case is smaller compared with the unconstrained scenario for a fixed value of $\mu_t$. The left sub-figure in the second row shows that for a constant level of missing segments ratio $\mu_m$, the level of segmentation granularity is higher in the constrained case. The right sub-figure in the second row shows that for a constant level of missing segments, the overlap ratio $\mu_r$ is smaller in the constrained scenario.

Altogether, a comparison of constrained and unconstrained segmentations has showed larger granularity and smaller overlap ratio for the constrained trials, in line with our assumption. These results mainly suggest that in the parallel approach, the value of the parameter $\lambda$ has to be adjusted to the execution speed. Based on an action-specific comparison, the next subsection describes these results in more detail.

### 5.5.4 Segmentation of Constrained vs. Unconstrained Trials

Based on trials recorded by three human demonstrators, this section illustrates an action-specific comparison of constrained vs. unconstrained segmentations for a fixed set of parameters: weight vector $(w_t, w_a, w_j) = (0.5, 0.3, 0.2)$ and $\lambda = 10^{-3}$. Table 5.16 gives an overview of the experiment.

Figure 5.24 presents four histograms, each of them comparing one of the four segmentation indices in constrained and unconstrained scenario. Higher segmentation granularity in the constrained scenario, and a higher temporal error in the unconstrained scenario demonstrated in the previous section can be clearly observed in these action-specific plots.

The first sub-figure in Figure 5.24 shows the action-specific temporal error, that ranges in both cases between 0.05 and 0.3 seconds. The level of error in the constrained scenario

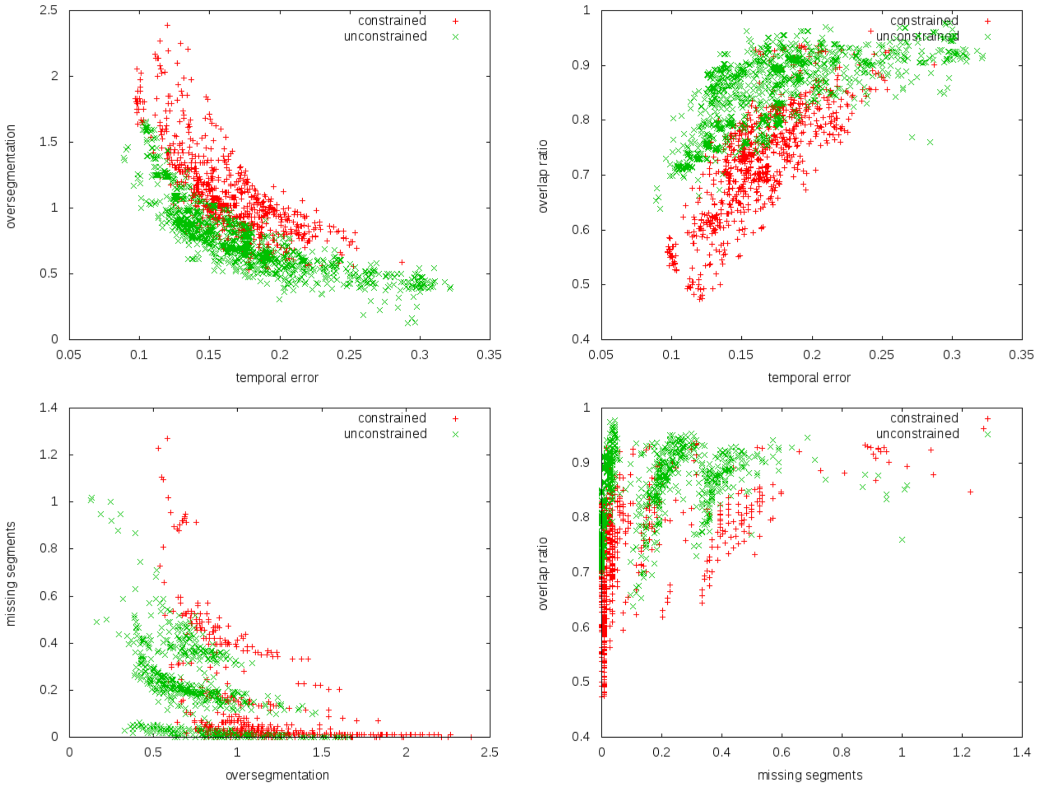


Figure 5.23: Comparison of segmentation for constrained (red) vs. unconstrained scenario (green). First row: dependency between $\mu_t$ and $\mu_g$ (left); dependency between $\mu_t$ and $\mu_r$ (right); Second row: dependency between $\mu_g$ and $\mu_m$ (left); dependency between $\mu_m$ and $\mu_r$ (right). The averages are build over all available trial in the constrained and unconstrained scenario, respectively.

is at most 0.2 seconds lower than in the unconstrained scenario. The right sub-figure in the first row shows action-specific segmentation granularity, whose value in the unconstrained case is slightly smaller than in the constrained case. This implies that more segments could be detected in the constrained case, which is possibly due to almost a double length of the constrained trials in comparison to the unconstrained trials. Oversegmentation of $\approx 2$ in cases of *lift* is due to the differing trial structure in the constrained vs. unconstrained case. Corresponding to this, the values of overlap ratio $\mu_{r,i}$ for *lift* are smaller in the constrained case. The other values are similar for both scenarios. The bottom right sub-figure compares the missing segments index $\mu_{m,i}$, whose results are similar apart from *put down1* and *put down2*. In the unconstrained scenario, *putting down* has a high undetection rate in comparison to the constrained scenario. Especially high undetection rate of *put down1* can be explained by the fact that it is conducted directly after *pouring* and is therefore likely to be fused with this segment on the basis of persisting audio signal. The other primitives exhibit the rate of missing segments index close to zero.

Table 5.16: Overview of the experiment: constrained and unconstrained segmentation for fixed parameters.

| Parameter | Value |
|---|---|
| $(w_t, w_a, w_j)$ | $(0.5, 0.3, 0.2)$ |
| $\lambda$ | $10^{-3}$ |
| $s$ | 7 |
| ground truth type | annotation |
| label collection | semantic |
| number of unconstrained trials | 40 |
| number of constrained trials | 40 |
| number of HDs | 3 |
| scenarios | constrained and unconstrained |
| segmentation approach | parallel |

Altogether, the segmentation generated for both, constrained and unconstrained trials is largely comparable. The difference can be observed in the segmentation granularity, where the constrained scenario generally exhibits higher action-specific values. We argue that this is mainly due to lower execution speed of the constrained trials, characterized by a higher number of action primitives, i.e. pauses, that are missing between such action primitives as "grasp" and "shake", or "grasp" and "pour" in faster executed unconstrained trials. Apart from one type of action primitive, *put down*, the low undetection rate indicates satisfactory segmentation results for both scenarios.

## 5.6 Comparison of Parallel and Hierarchical Segmentation

In this section we compare the segmentation quality generated by both proposed multimodal approaches: the parallel and the hierarchical. The goal of the experiment is to examine which approach generates segmentation closer to the ground truth. The data pool of the experiment consists of constrained and unconstrained trials recorded by three human demonstrators. Segmentation has been carried out based on the audio and the tactile modalities. Table 5.17 presents an overview of the experiment and the applied parameters.

Figure 5.25 illustrates the results of the comparison based on the action-specific values of the segmentation indices $\mu_{t,i}$, $\mu_{g,i}$, $\mu_{r,i}$, and $\mu_{m,i}$. All four sub-figures demonstrate a very large similarity of results in both cases. In the case of one type of action-primitive, *lifting*, the hierarchical approach produces higher segmentation granularity in comparison to the parallel approach. The hierarchical approach processes the audio modality individually in the second segmentation step. Therefore, we believe that an additional segment is generated based on the low level of noise accompanying grasping of the non-rigid test object. We presume that in the parallel approach this audio artifact is dominated by the homogeneity of the tactile modality, therefore no segment is generated. Corresponding to the oversegmentation of *lifting* is the small overlap ratio (bottom row, left). The overlap ratio of the other action primitives is very close to one. Missing segments index (bottom

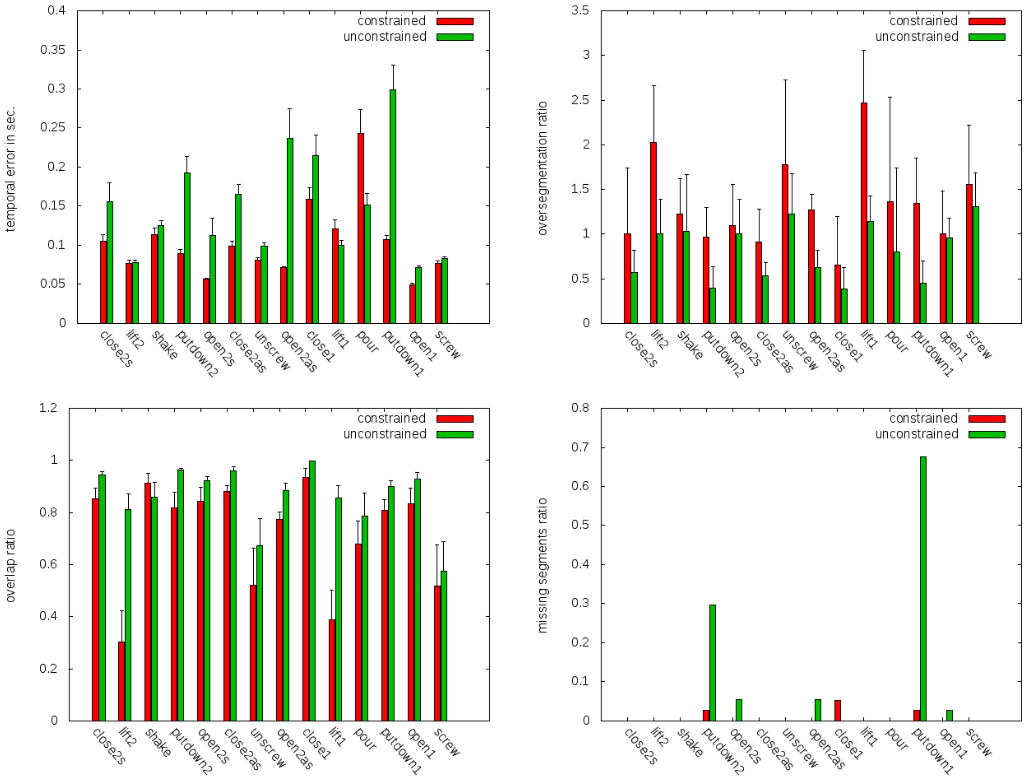


Figure 5.24: Comparison of action-specific segmentation of trials in constrained (red) and unconstrained (green) scenarios. The weight vector is $(w_t, w_a, w_j) = (0.5, 0.3, 0.2)$. First row: action-specific temporal error $\mu_{t,i}$ (left), action-specific segmentation granularity ratio $\mu_{g,i}$ (right); Second row: action-specific overlap ratio $\mu_{r,i}$ (left), action-specific missing segments ratio $\mu_{m,i}$ (right); Averages are build over all available trials in the corresponding scenario.

row, right) shows satisfactory results, apart from *put down 1* with <13% undetection rate. Altogether, this experiment shows comparable results for segmentation generated by both methods based on bimodal data.

## 5.7 Summary

This chapter has presented a systematic study and a comparison of the unimodal, bimodal and multimodal methods for interaction decomposition. The employed data pool consisted of multiple trials representing multimodal interaction recorded by four human demonstrators interacting with one object. Based on two types of ground truth, four proposed quality measures have been employed to assess structural and temporal accuracy of the segmentation.

With the goal to explore the modality-specific segmentation semantics based on the

Table 5.17: Overview of the experiment: comparison of two multimodal approaches.

| Parameter | Value |
|---|---|
| segmentation approach | parallel (p), hierarchical (h) |
| simple models employed in hierarchical approach | threshold, AR(1), AR(2), AR(3) |
| simple models employed in parallel approach | threshold, constant |
| $(w_t, w_a, w_j)$ | $(0.5, 0.4, 0)$ |
| $\lambda_p$ | $10^{-6}$ |
| $s_p$ | 8 |
| $\lambda_h$ | $10^{-6}$ |
| $s_{\text{audio}}$ | 5 |
| $s_{\text{tactile}}$ | 20 |
| ground truth type | annotation |
| label collection | semantic |
| number of unconstrained trials | 40 |
| number of constrained trials | 40 |
| number of HDs | 3 |

assigned models, and the influence of the two global parameters $s$ and $\lambda$, we have started our study with a series of unimodal segmentation experiments. A robust finding from all unimodal decomposition experiments was a modality-specific semantic sensitivity to detect segment boundaries with regard to the type of the action primitive. Furthermore, an intermodal comparison of the three modalities has demonstrated that such sensitivity for detecting a particular subset of the complete semantic segmentation resulted in three, to a large extent complementary segmentations.

To exploit the above-mentioned complementary semantic roles, we proceeded to study the bimodal segmentation based on the hierarchical approach. The hierarchical approach, tested with tactile and audio modalities, has yielded a segmentation corresponding to ten action primitives, successfully integrating two very different modalities, and demonstrating robustness w.r.t to three different human demonstrators. Characteristic for this method is the prior knowledge, defining the execution order of modality-specific segmentation steps. Due to substantial differences of the raw data such modality-specific sequential segmentation is highly advantageous. In addition to an optional filter step, the method allows a modality-specific choice of the global parameters $s$ and $\lambda$, which is not possible in the parallel approach.

In the final part of our study, we have explored the decomposition based on all three modalities with the parallel segmentation approach. This method yielded a robust segmentation corresponding to fourteen action primitives, successfully integrating all three modality-specific segmentations. Essential for the parallel approach is the weight vector that determines the influence of different modalities within the joint modality-integrating approach. The empirical evaluations have demonstrated that the suitable weight values, influencing the semantic sensitivities of the corresponding modalities, fully compensate for not choosing modality-specific priors (cf. hierarchical approach).

In all empirical evaluations, both approaches yielded segmentations characterized by

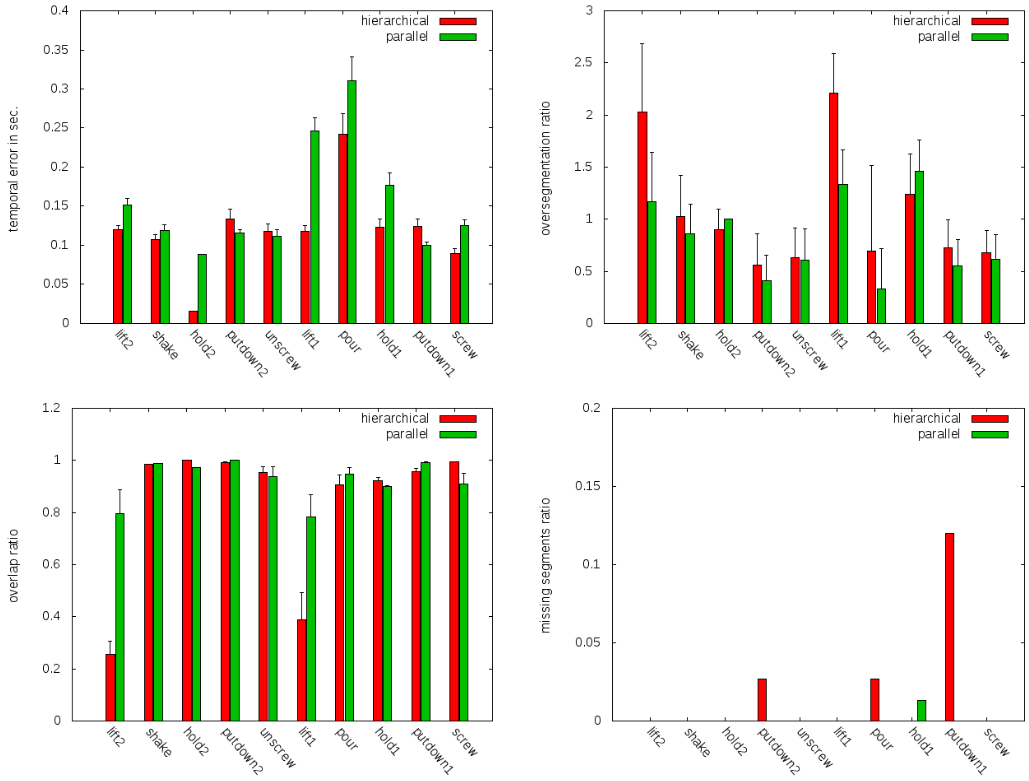


Figure 5.25: Comparison of segmentation quality for both approaches: hierarchical (red) and parallel (green). Top row: action-specific temporal error $\mu_{t,i}$ (left); action-specific segmentation granularity $\mu_{g,i}$. Bottom row: action-specific overlap ratio index $\mu_{r,i}$ (left); action-specific missing segments ratio $\mu_{m,i}$ (right).

a high temporal precision (temporal error ranges from 0.05 to 0.3 second) accompanied by a low undetection rate of less than 5% for thirteen action primitives. In a bimodal comparison of both decomposition approaches, the parallel approach has yielded a slightly better segmentation w.r.t. the detection rate. We believe that the reason for this is the weight vector that has increased the sensitivity of the audio modality to inhomogeneity in the data, yielding a better detection rate.

Evaluation conducted with manual annotation vs. cue-based ground truth has showed, despite a high temporal error, promising results of the cue-based method. However, in future experiments it would be favorable to reduce the temporal error resulting from the imprecise alignment of the HDs. Based on the proposed multimodal segmentation methods, the next chapter deals with the question of high-level modeling and identification of the generated action primitives.

# Chapter 6

# Towards High-Level Modeling

According to our original assumption, representation and identification of action primitives can serve as a building block for higher-level modeling and recognition of actions and activities in interactive scenarios, such as cooperation and assistance (see Chapter 2). Methods presented in Chapters 4 and 5 have been employed to decompose interaction on the lowest level into action primitives by detecting change within multimodal time series. However, in general this segmentation approach does not provide a semantic description of the generated segments. Therefore, this chapter proposes an approach to multimodal representation and classification of action primitives. To this end, we consider the segmentation method presented in the previous sections as a building block of the higher-level modeling and recognition approach.

In order to identify action primitives, segments that contain semantically equivalent data have to be grouped, and models of these groups have to be formed. Our approach of this challenge is motivated by the results of empirical studies, suggesting that the configurational information of the spatiotemporal dynamic form of actions is used by people to group those [38]. Hence, we address both above-mentioned tasks by embedding the concept of ordered means models (OMMs) [36] in a clustering approach. These models have demonstrated to be especially well-suited for incomplete sequential data.

The rest of the chapter is structured as follows:

- Sections 6.1 and 6.2 present the theoretical background of the OMMs and the corresponding clustering approach.
- Section 6.3 presents the evaluation method used to asses the quality of clustering based on the available ground truth.
- Section 6.4 describes the data pool and the clustering experiments, whereby the focus is on the role of multiple modalities in identification of action primitives.
- Section 6.5 summarizes the high-level modeling, before Chapter 7 gives the conclusions for the complete thesis.

## 6.1 Ordered Means Models

OMMs are generative state space models with a hidden state, left-to-right topology, and Gaussian emission densities. The model has been developed by U. Grossekathoefer and T. Lingner [36] and successfully used to model multivariate and multimodal sequential data [88, 33, 35, 34].

In case of the generated action primitives, robust modeling in terms of incomplete data and variable execution speed is required. Even though approaches such as hidden Markov models (HMMs) reach excellent results for complete data, they might not be the optimal choice for scenarios with time series with missing beginnings or endings (see [34]). In particular, HMMs' implicit modeling of segments length distributions in terms of transition probabilities could lead to an inadequate representation for missing data, or execution with different speed. Here, as a major difference in the overall model design, OMMs do not incorporate any transition probabilities. Instead, all paths, i.e. all valid sequences of model states, are equally likely. In the following paragraphs we describe the structure of an OMM in detail.

### 6.1.1 Means Vector and Emission Densities

The central component of the model [36] is an ordered sequence of $K$ model states. The sequence is represented by vectors, corresponding to the expected values of emission densities:

$$\Omega = (\mu_1, \dots, \mu_K),$$

where for $1 \leq k \leq K$, $\mu_k \in \mathbb{R}^d$ and $d$ is the dimensionality of the time series. Each state is characterized by an emission distribution modeled by a Gaussian probability distribution $b_k(\cdot)$:

$$b_k(\mathbf{o}_t) = \mathcal{N}(\mathbf{o}_t; \mu_k, \sigma),$$

where the standard deviation parameter $\sigma$ is identical for all states and is thus a *global hyperparameter*. The sequence of observations emitted by the model is denoted as follows:

$$O = (\mathbf{o}_1, \dots, \mathbf{o}_T), \ \mathbf{o}_i \in \mathbb{R}^d.$$

### 6.1.2 Path Probabilities and Production Likelihood

A path through a model is defined as a valid sequence of states w.r.t. model topology. Unlike HMMs, the main assumption of OMMs is that each path through the model is equally likely. Theoretically OMMs require the definition of an explicit length distribution either by domain knowledge or by estimation from the observed lengths in the training data. This, however, may not be possible due to missing knowledge or non-representative lengths of the observations. To avoid the definition and estimation of the length, the authors assume a flat distribution in terms of an improper prior according to equally probable lengths [34].

For a given length $T$, a path probability for a sequence $\mathbf{q}_T = q_1, \dots, q_T$ is defined as follows:

$$p(\mathbf{q}_T|\Omega) = \begin{cases} \frac{1}{M_T} P(T), & \text{if } q_1 \leq q_2 \dots \leq q_T, \\ 0 & \text{else} \end{cases}$$

where $M_T$ is the number of valid paths for the time series of length $T$ through a $K$-state model:

$$M_T = |\{\mathbf{q}_T : q_1 \leq q_2 \leq \ldots \leq q_T\}| = \binom{K+T-1}{T}.$$

The likelihood to observe a sequence $O$, given path $\mathbf{q}_T$ and model $\Omega$ under the assumption of independent observations is:

$$p(O|\mathbf{q}_T, \Omega) = \prod_{t=1}^{T} p(\mathbf{o_t}|q_t, \Omega) = \prod_{t=1}^{T} b_{q_t}(\mathbf{o}_t).$$

The likelihood of the observed sequence $O$ and a path $\mathbf{q}_T$ for a given model $\Omega$ is:

$$p(O, \mathbf{q}_T|\Omega) = p(O|\mathbf{q}_T, \Omega) \cdot P(\mathbf{q}_T|\Omega).$$

The overall production likelihood for a time series of length $T$ is:

$$p(O|\Omega) = \sum_{\mathbf{q}_T} p(O, \mathbf{q}_T|\Omega).$$

For efficient computation of production likelihoods Grossekathöfer et al. [33] use a dynamic programming solution similar to the forward-backward algorithm used for HMMs, but omitting transition probabilities.

In order to estimate $\Omega = (\mu_1, \ldots, \mu_K)$ from a set of observation sequences $\mathbf{O} = \{O_1, \ldots, O_N\}$ the authors maximize the following log-likelihood [36]:

$$\mathcal{L} = \sum_{i=1}^{N} \log p(O_i|\Omega) \tag{6.1}$$

with respect to the mean vectors $\mu_k$. To solve this optimization problem for an fixed value of $K$, an iterative expectation maximization algorithm is employed [34, 33].

### 6.1.3 State Duration Probabilities

State duration probabilities in an OMM distinguish this model from a HMM. In contrast to the geometric state duration distribution of a standard HMM, the state duration probability in an OMM depends on the sequence length. For a sequence of length $T$ and the number of model states $K$ and it is defined as follows [33]:

$$P(t) = \frac{\binom{T+K-2-t}{K-2}}{\binom{T+K-1}{K-1}}, \tag{6.2}$$

where $t$ denotes the number of time steps the model remains in any given state.

## 6.2 Clustering with OMMs

In our approach OMMs are integrated into a EM-based clustering procedure for identification of action primitives. The output of the proposed segmentation method is a set of

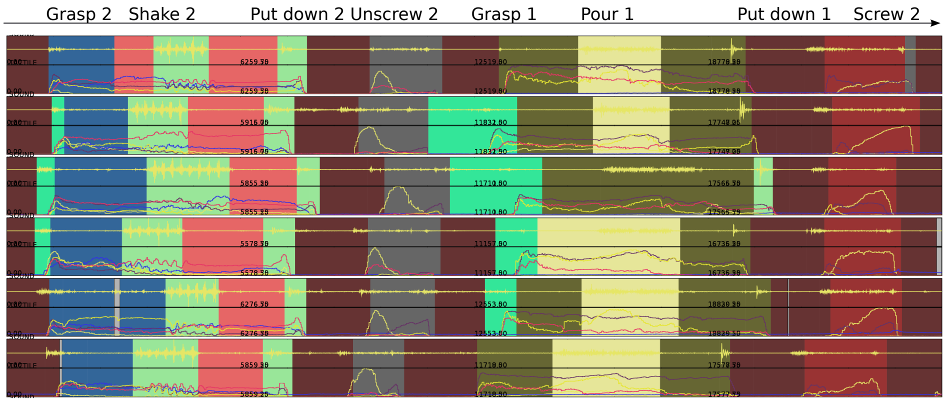


Figure 6.1: Assignment of labels (*designated by random colors*) to segments according to the best matching model in a small subset of trials. In each row, the segmentation, label assignments, audio signal (*top half*) and tactile information (*bottom half*) is showed. Corresponding segments in adjacent trials do not line up because of the randomized timing. $K = 50$, $\sigma = 2.5$, $C = 11$. Note that the trials in the figure are stretched to yield a common predefined figure length.

multimodal data sequences $\{O_n\}_{1 \leq n \leq N}$ that are unlabeled w.r.t. the trials and actions from which they originate. The application of OMMs to partition such a dataset into $C$ groups in an unsupervised manner, can be considered a special case of the well-known $k$-means clustering. OMMs $\Omega_1, \ldots, \Omega_C$ are used as the associated prototypes of $C$ clusters. A suitable distance function then is the negative log-likelihood that a sequence $O_n$ is generated by an OMM $\Omega_j$: $d(O_n, \Omega_j) = -\log P(O_n \mid \Omega_j)$. Given this, a $C$-OMMs clustering algorithm partitions data sequences into $C$ groups by minimizing the objective function [7]:

$$E = -\sum_{n=1}^{N} \sum_{j=1}^{C} w_{n,j} \log P(O_n \mid \Omega_j).$$

subject to $w_{n,j} \in \{0, 1\}$ and $\forall n : \sum_{j=1}^{C} w_{n,j} = 1$.

Fig. 6.1 qualitatively shows the result of applying the sketched clustering procedure in the following way: in a training step, eleven OMMs are formed based on segmentations obtained with the hierarchical segmentation method (see Section 4.3.2). Then, in a test step, segmented action sequences that are not part of the training set are classified to the best-matching OMM model. Identically colored segments are considered semantically equivalent. Note that in all further plots, the trials are stretched to yield a common predefined figure length.

## 6.3 Measures of Clustering Quality

Quality of clustering is evaluated with the help of two entropy-based measures *homogeneity* and *completeness* (described in i.e. [79]). Entropy of a discrete random variable $X$ with possible realizations $\{x_1, \ldots, x_n\}$ is defined as follows:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i), \tag{6.3}$$

where $p$ denotes a probability mass function of X. Uniform distribution yields the highest value of entropy. The more skewed the distribution is, the smaller the entropy value gets.

Let $p_i$, $1 \leq i \leq N$, denote action primitives i.e. *shaking* or *pouring* and $L = \{1, \ldots, C\}$ define the set of labels. The homogeneity of the cluster $l \in L$ grows, the more action primitives corresponding to one type $p_i$ are assigned to it. Let $P_l$ denote the distribution of action primitives in the cluster $l$. The highest homogeneity is achieved in the case, if the probability mass of the corresponding label distribution is concentrated in a single action primitive $p_i$, resulting in $P_l(p_i) = 1$ and $P_l(p_j) = 0$ for $j \neq i$. This distribution has zero entropy. In order to calculate homogeneity of the complete clustering, average entropy over all clusters is calculated as follows:

$$H_l = -1/C \sum_{l=1}^{C} \sum_{i=1}^{N} P_l(p_i) \log P_l(p_i). \tag{6.4}$$

Therefore, the smaller the value of $H_l$, the more homogeneous the clustering. We refer to $H_l$ in the later sections as "label entropy".

The second measure for evaluation of clustering consistency is *completeness*. For an action primitive type $p_i$, $i \in N$, the completeness measure improves the less different labels $l \in L$ have been assigned to it. Let $P_i(l)$ denote the relative frequency of label $l$ being assigned to the observations of the action primitive $p_i$. The highest completeness is achieved, when only one label is assigned to all corresponding points yielding $P_i(l) = 1$ and $P_i(m) = 0$ for all $m \in L$, $m \neq l$. This corresponds to zero entropy for this type of action primitive. The completeness measure for all action primitives is calculated by averaging over all primitives $p_i$, $1 \leq i \leq N$:

$$H_c = -1/N \sum_{i=1}^{N} \sum_{l=1}^{C} P_i(l) \log P_i(l). \tag{6.5}$$

The smaller the value of H$_c$, the better the value of completeness. In the following sections we refer to $H_c$ as "cue entropy". The overall value of entropy used for evaluation of clustering quality is defined as the sum of the completeness value $H_c$ and the homogeneity value $H_l$:

$$H = H_c + H_l. \tag{6.6}$$

## 6.4 Experimental Results

Chapter 5 has presented satisfactory results for multimodal interaction decomposition into action primitives. This section is dedicated to experiments exploring multimodal representation and unsupervised learning of the generated action primitives by OMM-based clustering. The main purpose of the experiments is to investigate the role of multiple modalities for identification of action primitives in our approach.

The following subsections are organized as follows: Subsection 6.4.1 presents the data pool. Subsection 6.4.2 describes the search for optimal clustering parameters: number of model states $K$, number of clusters $C$ and emission variance $\sigma$. Subsection 6.4.3 investigates the impact of multiple modalities on the resulting clustering. Subsection 6.4.4 shows the robustness of OMMs w.r.t. execution velocity of action primitives and analyzes the clustering w.r.t. different human demonstrators. Because the aim of the following subsections is an exploratory study of the clustering semantics, no corroboration of the findings with statistical confidence measures will be provided.

### 6.4.1 Data Pool

The data pool consists of segments generated by a hierarchical method based on audio and tactile modality (see Section 4.3.2). Segmentations generated for both, constrained and unconstrained trials by all human demonstrators have been included, whereby on average the length of segments in the unconstrained trials is half of the length of the segments in the constrained scenario. The corresponding set of primitives consists of:

| description | 1 / 2 (uni-/bimanual) |
|---|---:|
| grasp + lift | 2 |
| hold | 2 |
| shake | 2 |
| put down | 2 |
| pause | idle |
| unscrew | 2 |
| pause | idle |
| grasp + lift | 1 |
| pour | 1 |
| put down | 1 |
| pause | idle |
| screw | 2 |

For evaluation, the ground truth based on annotation (cue label collection, see Appendix B) has been employed. The overview of the data pool is presented in the following table:

| | |
|---|---|
| segmentation method | hierarchical |
| modalities | audio, tactile, joint-angles |
| total number of HD | 4 |
| trial types | constrained, unconstrained |
| total number of trials | 100 |
| total number of segments | c.a. 1700 |
| average length of constrained segment used in clustering | 305 |
| average length of unconstrained segment used in clustering | 161 |
| ground truth type | cue-based, annotation-based |
| number of action primitives | 11 |
| label collection | cues |

For more detailed information regarding the number of trials per human demonstrator, see Table 5.1.

Prior to performing $k$-OMM clustering, two preprocessing steps are applied to the output of the segmentation step. Firstly, the time-domain audio signal is replaced by a coarse characterization in the frequency domain. We apply a sliding-window version of the Discrete Fourier Transform to the audio signal and extract ten coefficients of the lowest frequencies from each result. The time series of these coefficients replaces the audio-signal. This transformation is motivated by the fact that the oscillatory nature of the time-domain audio signal is not compatible with the OMM emission models, which assume piecewise constant data with fixed-variance Gaussian noise. Secondly, we assign constant values to modalities associated with an "inactive" hand for the duration of the inactivity. This step is intended to prevent the representation of patterns that are not related to object manipulation in learned OMMs.

### 6.4.2 Parameter Estimation with Cross-validation

Prior to learning with OMM-clustering, it is necessary to find appropriate values for three parameters: the number of clusters $C$, the number of model states $K$ and the emission variance $\sigma$. For this purpose we have investigated the clustering entropy within a three dimensional grid of triples $(C, K, \sigma)$. An overview of the experiment is presented in the following table:

| | |
|---|---|
| method | 5-fold cross-validation |
| trial type | constrained, unconstrained |
| number of trials | 100 |
| number of HDs | 4 |
| number of segments | c.a. 1700 |
| $K$ | $\{2, 5, 8, \ldots, 50\}$ |
| $C$ | $\{10, 15, \ldots, 30\}$ |
| $\sigma$ | $\{1, 1.5, \ldots, 6\}$ |

In this experiment for each combination of parameters $C$, $K$ and $\sigma$ we calculate the overall entropy $H$ by averaging over the individual entropy values corresponding to the five test sets of the randomly initialized five-fold cross-validation set. The data pool consists of 100 trials corresponding to c.a. 1700 segments.

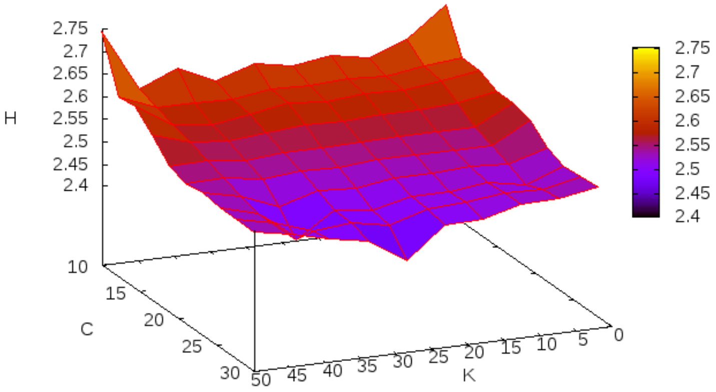


Figure 6.2: Dependency between the number of model states $K$, number of clusters $C$ and the entropy $H$ for a constant value of $\sigma = 5.0$.

Figure 6.2 illustrates the relation between the parameters $K \in \{2, \ldots, 50\}$, $C \in \{10, \ldots, 30\}$ and the entropy $H$ for a constant value of $\sigma = 5.0$. The figure demonstrates a clear dependency between the number of clusters and the entropy: $H$ decreases from 2.6 to 2.4 with increasing number of clusters $C$. Parameter $K$ has a relatively small effect on the entropy. Good results are achieved for large $C$ and $K \in \{25, \ldots, 45\}$.

Figure 6.3 shows the relation between the number of clusters $C \in \{10, \ldots, 30\}$, state emission variance $\sigma \in \{1, \ldots, 6\}$ and entropy $H$ for a constant value of $K = 50$. Again, the number of clusters has a dominating effect on the entropy: the entropy decreases for increasing values of $C$.

Finally, Figure 6.4 shows the dependency between the number of model states $K$, $\sigma$ and entropy $H$ for a constant value of $C = 30$. The figure shows that for a growing $K$ the value of $H$ improves for small values of $K$ up to 20. Further increase of $K$ does not have any influence on the entropy. Increasing the value of $\sigma$ improves consistency of clustering for small values of $K < 20$. For larger values of $K$, increasing $\sigma$ does not effect the entropy.

The positive effect of increasing $C$ on the entropy can be explained by an improvement of the label entropy $H_l$: the growing number of clusters improves the clustering homogeneity until each observation is assigned to its own cluster. Therefore, depending on the targeted number of resulting clusters an appropriate value of $C$ should be chosen. In our case we choose $C = 11$. The experiments also illustrate a constant level of clustering entropy for a large values of $K > 20$ and $\sigma > 1$. For small values of $K$ large $\sigma$ improves the entropy. This implies that it is possible to compensate a possibly insufficient number of model states by choosing a large state emission variance, covering altogether a larger range of values.

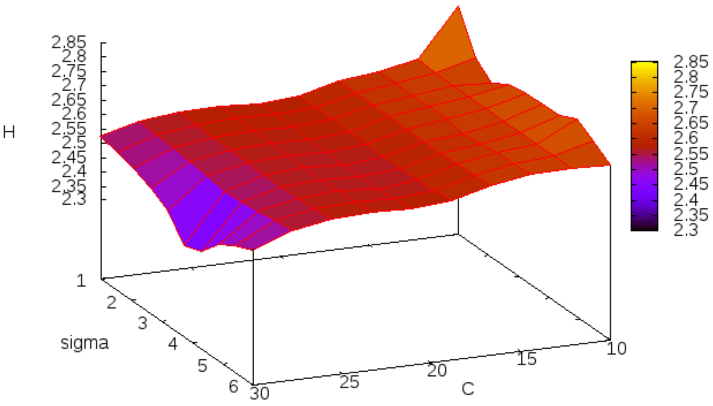


Figure 6.3: Dependency between the number of clusters $C$, the state emission variance $\sigma$ and entropy $H$ for a constant value of $K = 50$.

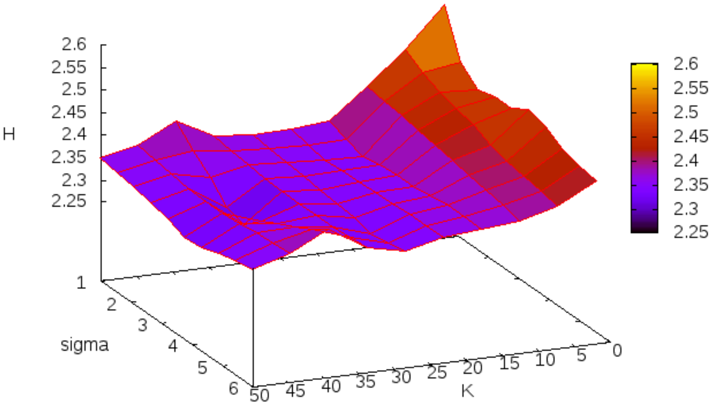


Figure 6.4: Dependency between K, $\sigma$ and $H$ for a constant value of $C = 30$.

Table 6.1: Overview of entropy values for each combination of modalities.

| modality combination | $H$ | $H_c$ | $H_l$ |
|---|---|---|---|
| tactile | 1.97 | 0.96 | 1.01 |
| audio | 1.81 | 0.89 | 0.92 |
| joints | 1.90 | 0.84 | 1.06 |
| tactile+audio | 1.74 | 0.80 | 0.94 |
| audio+joints | 1.77 | 0.89 | 0.88 |
| tactile+joints | 2.26 | 1.09 | 1.17 |
| audio+tactile+joints | 1.91 | 0.96 | 0.95 |

### 6.4.3 Clustering for Different Modality Combinations

In this section we investigate the influence of different modality combinations on the representation of action primitives with OMM-based clustering. The main purpose of this experiment is to explore the main semantic characteristics, advantages and disadvantages of clustering based on each modality combination. In our experiments we are looking for modality-specific characteristics of the classification that would most probably be invariant to the type of features used.

In the following text we present the results of clustering for all seven possible modality combinations. For this experiment we use segmentations generated from 30 trials recorded by one human demonstrator in a constrained scenario; the entropy is calculated w.r.t. the labels generated by clustering of this set during training. A constant parameter set has been used: $K = 40$, $C = 11$, $\sigma = 2.5$.

Figures 6.5-6.11 illustrate clusterings resulting for different modality combinations, Table 6.1 presents an overview of the corresponding entropy values.

Figure 6.5 illustrates clustering based on 10-dimensional tactile data. The resulting labeling clearly differentiates uni- and bimanual actions (e.g. *shake* and *pour*) as well as different levels and patterns of force application. This example shows that after grasping an object only very few labels are used, indicating an approximately constant level of applied force characteristic for this human demonstrator. The cue entropy $H_c$ of this clustering is second lowest among all combinations.

Figure 6.6 depicts clustering based on audio modality represented by 10 Fourier coefficients of the lowest frequencies extracted from the audio input. Compared to the tactile modality, this clustering is characterized by good discriminative ability of different audio-producing action primitives, like shaking, pouring or screwing. At the same time this clustering can not differentiate between uni- and bimanual actions, that are characterized by similar fourier spectrum. Both, bi- and unimanual action primitives (e.g. *grasp2* followed by a bimanual *hold*, and *grasp1* followed by unimanual *hold* assigned to brown and magenta colors resp.) correspond to the same label. Nevertheless both, label entropy $H_l$ and cue entropy $H_c$ are lower for audio- than for tactile-based clustering. This is presumably due to the fact that in some cases uni- and bimanual can be additionally differentiated by different levels of accompanying noise.

Figure 6.7 shows clustering based on joint-angles modality. This labeling clearly illustrates four distinct regions, corresponding to individual joint-angle configurations for four different grasps starting with: *grasp2*, *unscrew2*, *grasp1* and *screw2*. This ambiguity causes the highest label entropy among all three modalities.

Figure 6.8 illustrates labeling resulting from combining tactile and joint-angles modalities, yielding the highest entropy (see Table 6.1) which corresponds to the worst clustering. Combination of both modalities increases both entropy values $H_c$ and $H_l$ in comparison to clustering with individual modalities. As expected, uni- and bimanual actions can be well discriminated in this combination. At the same time the examples demonstrate a low labeling consistency for action primitives within the uni- or bimanual "object contact" regions.

In contrast to the previous example, clustering with a combination of audio and the joint-angles modalities indicates an improvement of label entropy $H_l$ in comparison to both unimodal clusterings, the audio- and the joint-based. (see Table 6.1). The cue entropy $H_c$ stays on the level of joint-based clustering. Figure 6.9 shows an example of label assignment for this modality combination. Presumably due to prevailing influence of audio in this combination, uni- and bimanual action primitives, i.e. *grasping* directly followed by *holding* are assigned to the same cluster (magenta and red resp.).

Figure 6.10 demonstrates clustering based on audio and tactile modalities, yielding the lowest overall entropy among all modality combinations. It is characterized by the lowest value of cue entropy $H_c$ and the second lowest label entropy $H_l$ which is close to the label entropy of the audio-based clustering.

The combination of three modalities used for clustering is illustrated in Figure 6.11. This example shows a prevailing influence of joint-angles and tactile modalities over audio. There is no confusion between the uni- and bimanual action primitives, however, e.g. an audio-specific action primitive like *shaking* is assigned the same label as *grasp2* in this example.

Altogether, the overall entropy ranges from the value 1.74 (tactile+audio) to 2.26 (tactile+joints). Audio-based clustering yields the best performance among individual modalities. The clustering can be improved by adding either the joint-angles or the tactile modality to audio to improve the uni- and bi-manual differentiation.

### 6.4.4 Clustering of Fast and Slow Action Primitives

Because of the execution speed variability between the human demonstrators, in the last experiment we investigate the influence of the execution speed on the results of clustering. To this end, we compare the test entropies of three human demonstrators in two speed categories: fast and slow. In both cases the training is conducted with slow action primitives.

The fast and slow action primitives are generated for each HD from the corresponding unconstrained and constrained trial segmentations respectively. The execution of the fast action primitives is on average two times faster than of the slow action primitives. The training set of this experiment consists of approx. 1000 slow action primitives generated from trials of four human demonstrators. The test set, different from the training set, consists for each human demonstrator of approx. 100 slow and 100 fast action primitives respectively. Table 6.2 presents the test results for a constant parameter set $K = 20$, $C = 10$ and $\sigma = 2$, chosen based on the results of the cross-validation. The table shows only a slight increase of entropy ranging from 0.13 to 0.34 for fast action primitives in comparison to the slow for all HDs. The results in Table 6.2 also illustrate differences in clustering quality among HDs. It amounts to 0.76 in the fast and 0.55 in the slow case. This can be explained

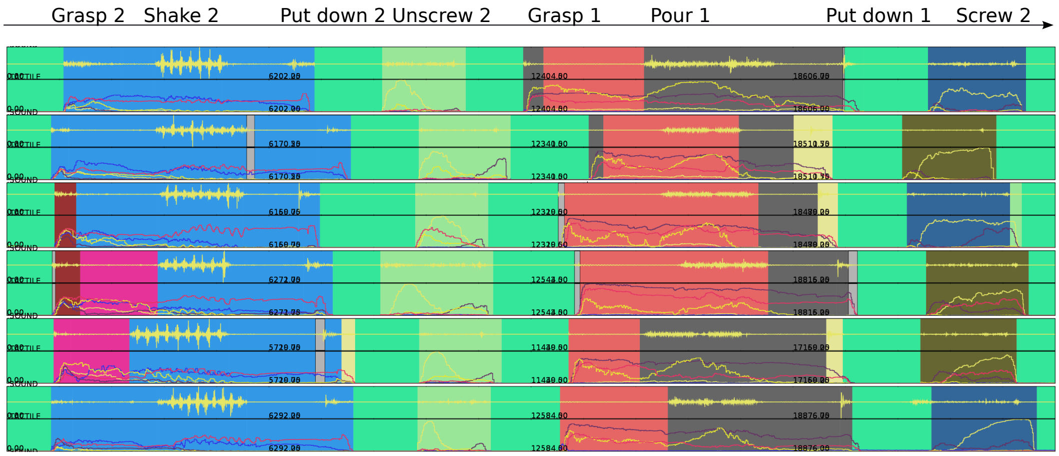


Figure 6.5: An example of label assignment for clustering with tactile modality. Labels are designated by random colors.

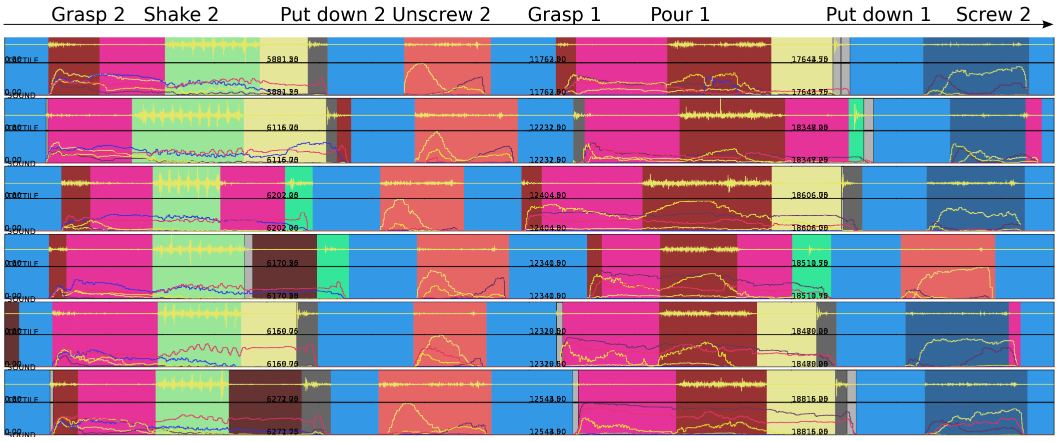


Figure 6.6: An example of label assignment for clustering with audio modality. Labels are designated by random colors.

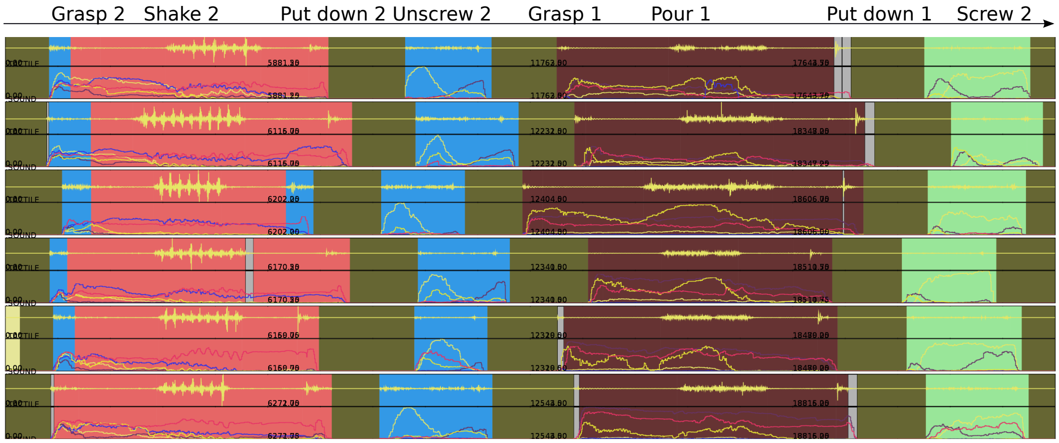


Figure 6.7: An example of label assignment for clustering with joint-angles modality. Labels are designated by random colors.

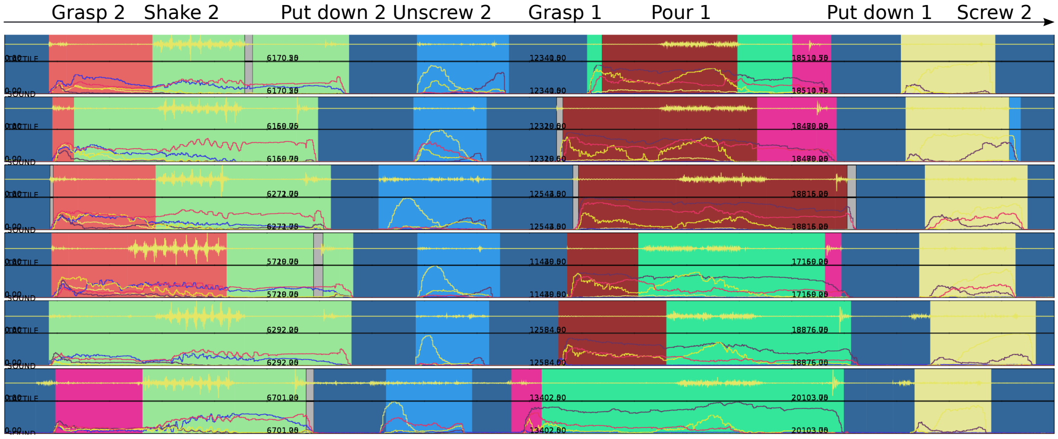


Figure 6.8: An example of label assignment for clustering with a combination of tactile and joint-angles modalities. Labels are designated by random colors.

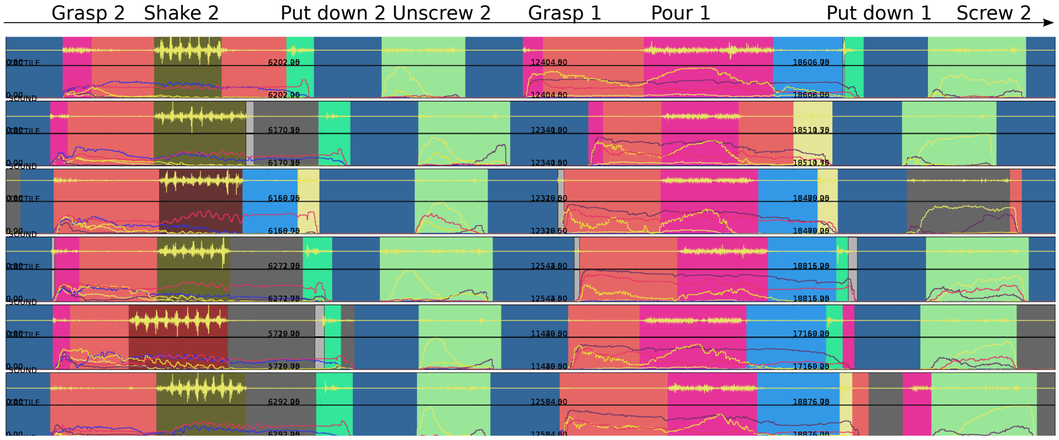


Figure 6.9: An example of label assignment for clustering with a combination of audio and joint-angles modalities. Labels are designated by random colors.

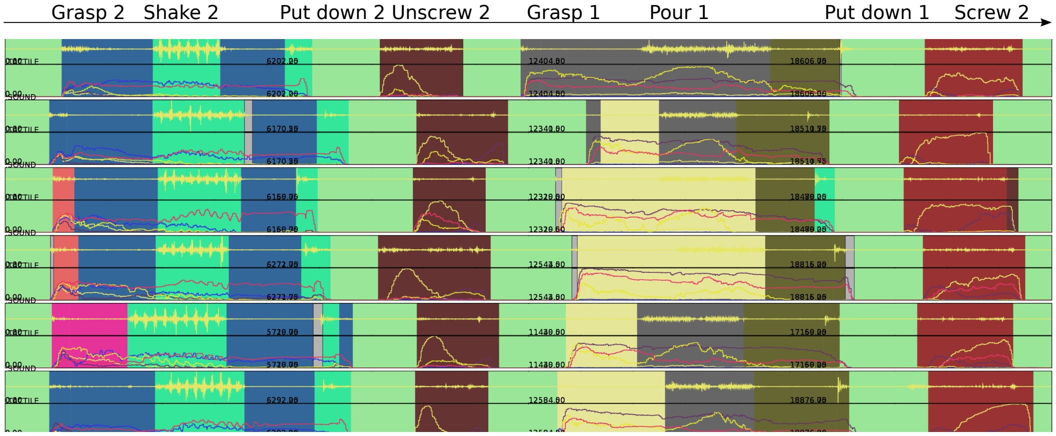


Figure 6.10: An example of label assignment for clustering with a combination of tactile and audio modalities. Labels are designated by random colors.

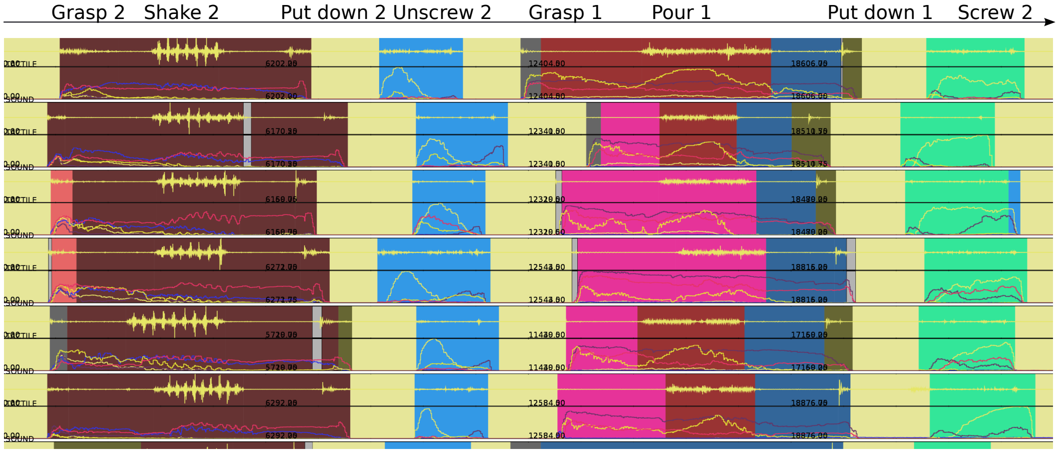


Figure 6.11: An example of label assignment for clustering with all modalities: audio, joint-angles, and tactile modalities. Labels are designated by random colors.

by the differences in the quality of the raw data for different HDs.

Table 6.2: Overview of entropy values resulting from evaluation of clustering for slow and fast action primitives for three human demonstrators.

| human demonstrator | slow | fast | increase |
|---|---|---|---|
| $\mathtt{hd}_1$ | 1.49 | 1.62 | 0.13 |
| $\mathtt{hd}_2$ | 2.04 | 2.38 | 0.34 |
| $\mathtt{hd}_3$ | 1.69 | 1.90 | 0.21 |

Clustering experiments with combinations of slow test sets $\mathtt{hd}_1 + \mathtt{hd}_3$ and $\mathtt{hd}_1 + \mathtt{hd}_2 + \mathtt{hd}_3$ have showed an increase of the entropy in comparison to test sets encompassing only one HD. An increase in cue entropy in each case (see Table 6.3) suggests that the clustering is characterized by HD-specific action primitive clusters.

Table 6.3: Overview of entropy values resulting from evaluation of clustering for slow and fast action primitives for three human demonstrators.

| test set | entropy | $H_c$ | $H_l$ |
|---|---|---|---|
| $\mathtt{hd}_1$ | 1.9 | 0.82 | 1.13 |
| $\mathtt{hd}_2$ | 2.38 | 1.05 | 1.32 |
| $\mathtt{hd}_3$ | 1.62 | 0.84 | 1.05 |
| $\mathtt{hd}_1 + \mathtt{hd}_3$ | 2.38 | 1.14 | 1.23 |
| $\mathtt{hd}_1 + \mathtt{hd}_2 + \mathtt{hd}_3$ | 2.64 | 1.18 | 1.45 |

## 6.5 Summary

This chapter has explored an approach towards unsupervised identification of manual interaction. Based on segments, generated with the help of the hierarchical segmentation, higher-level modeling and classification of action primitives has been realized by an OMM-based clustering. The main goal of the empirical evaluation was to investigate the influence of multiple modalities on the semantic characteristics and the quality of the clustering. In all experiments the evaluation of quality has been conducted with the help of standard entropy-based measures, completeness and homogeneity.

The main part of the experiments has examined the clustering for all seven combinations of the three modalities. The unimodal approaches have demonstrated the expected modality-specific clustering characteristics. On the one hand, the unimodal audio-based clustering performed well for action primitives accompanied by noise, but did not robustly differentiate between uni- and bimanual action primitives. On the other hand, the formed clusters in the unimodal joint-angles and tactile approach, reflected only the hand configuration and the grasp force respectively. The resulting clustering was characterized by an excellent performance differentiating uni- und bimanual action primitives. As expected, the approach did not perform well for action primitives characterized by e.g. a common grasp configuration but different audio features.

The experiments have demonstrated that these results can be improved by clustering multimodal data. The best performance has been achieved for the combinations of the tactile or the joint-angles modality with the audio modality. The final conclusions, including the topic of this chapter, will be presented in the next chapter.

# Chapter 7

# Conclusion and Outlook

In this thesis we have proposed and explored a framework for identification of semantic chunks in a multimodal manual interaction episode. Based on Bayesian statistics, we have first showed how the semantics of the manual interaction data can be modeled for a single modality, and then demonstrated how this modeling can be extended to integrate multiple modalities. We finally showed the suitability of the framework to decompose and encode a representative sequence of actions on an object, and, furthermore, its modularity w.r.t. multimodal and bimanual input.

Activity Theory (see Chapter 2), decomposing interaction on three semantic hierarchical levels into action primitives, actions and activities, has served as a conceptual basis for our approach. Seizing this concept, the initial idea of our approach is recognition through decomposition. Building upon the decomposition, our approach towards recognition of interaction yielded altogether a two-step framework.

In the first step (Chapters 4 and 5) we have proposed to decompose interaction into action primitives and, aiming at scalability to a large number of scenarios, based our approach on detection of change. To this end, we have described how different types of simple unimodal, bimodal and multimodal models can be employed to model the action primitives. Then, by applying Bayesian change detection methods to the above semantic modeling, we have proposed two approaches to carry out multimodal bimanual interaction decomposition. Finally, we have showed, how the above decomposition approach can be applied to decompose a representative multimodal bimanual interaction episode.

In the second step (Chapter 6), we have explored an unsupervised learning approach to encode multimodal action primitives. The underlying modeling is largely motivated by the results of empirical studies suggesting that action concepts contain information about the spatiotemporal dynamic form of actions, whose configurational information is used by humans to group actions [38]. Guided by these results, we have investigated a higher-level representation of action primitives based on state space models.

The following two sections are dedicated to the two steps of the approach, the decomposition and the higher-level modeling. The last section presents final concluding comments.

## 7.1 Decomposition of Interaction

Central for our decomposition approach has been the notion of homogeneity, characterizing action primitives, and, therefore, the detection of change in the homogeneous interaction flow, corresponding to the start of a new action primitive. However, typical change detection algorithms, employed for e.g. fault detection and designed to generate arbitrarily small or large segments, are not suitable for decomposition into action primitives, that have a substantial length. Hence, we have chosen a Bayesian change point detection method, incorporating, besides a model for estimation of homogeneity, a prior distribution on segment lengths. Apart from requiring a suitable prior, this method involves only a minimal amount of prior knowledge, which is an advantage in comparison to other state of the art methods.

In our work, we have proposed the first application of the Bayesian change point detection framework by Fearnhead [26] to multimodal interaction decomposition. A particular advantage of this method in comparison to other change detection methods, is the model uncertainty in each generated segment. This allows application of a wide range of simple as well as joint models. In order to apply the framework, originally developed for scalar time series, to multimodal and bimanual interaction data, our contribution consisted of two main parts: a modeling approach, and an extension of the segmentation procedure to a multimodal method.

A preliminary study of the modality-specific properties of the acquired data has demonstrated the necessity of a modality-specific modeling approach. In order to integrate multiple modalities and, at the same time, to allow multiple model states, we have proposed to employ a mixture of product models. Each product model, corresponding to a mixture model component, combines models representing modality-specific properties, and is, therefore, able to represent multimodal action primitives. This approach, that has been showed to achieve integration of multimodal and bimanual modeling for decomposition of manual interaction, accounts for one of our main contributions.

Based on the above-mentioned modeling approach, we have proposed two extensions of Fearnhead's framework for multimodal data: hierarchical and parallel segmentation approaches. We have showed that the hierarchical approach, characterized by a sequential execution of multiple segmentation steps, is particularly suitable for integrating strongly differing modalities, for which different types of priors need to be applied. An application of this method requires an additional prior knowledge defining the execution order of the modality-specific segmentation steps, which may be considered a disadvantage.

To tackle the above problem, the parallel approach carries out a one-pass segmentation, integrating all input modalities at once. To this end, multimodal integration in this method is additionally controlled by a weight vector, defining the sensitivity of each modality to inhomogeneity in the data. In contrast to the hierarchical segmentation, this method requires the definition of one global prior for all considered modalities, which may not be easy to find with a growing number of modalities.

A large part of this thesis dealt with empirical studies of the above decomposition approaches, and with the issues related to the evaluation of quality for the resulting segmentations. Firstly, as an alternative to the traditional manual annotation, we have proposed an automated cue-based ground truth (Chapter 3). We have demonstrated, how this kind of ground truth can be employed for evaluation and labeling of the generated segments, and discussed its disadvantages. Secondly, to compare the results of different approaches and to estimate the temporal and structural correctness of the generated segmentations within our empirical studies, we have proposed and applied four segmentation quality measures. W.r.t. the ground truth the proposed four measures indicate the temporal precision of the generated segment borders, the missing segments ratio (the undetection rate), the ratio of the segment length, and the granularity of the segmentation. The granularity index shows, whether the interaction episode has been over- or undersegmented. The evaluations have showed that these measures are sufficient for an approximative estimation of the segmentation quality.

With the aim to explore the semantics of different modalities, the first part of our empirical study dealt with unimodal segmentations. A robust result of this study has showed that each modality exhibits a specific semantics leading to a largely complementary set of change points, in comparison to the other modalities. To investigate the semantic range and the robustness of the multimodal decomposition, in our studies we have showed applications of both multimodal approaches, the hierarchical and the parallel segmentation. Integration of all three modalities resulted in robust decomposition of an interaction episode into fourteen action primitives. The empirical evaluations have indicated a very low temporal error (below 0.3 seconds), and a low undetection rate of less than 5% for thirteen action primitives.

## 7.2 Higher-level Modeling

Similar to the decomposition approach, the main goal of the higher-level modeling has been to explore the semantic role of multiple modalities for recognition of action primitives. Supported by the previous work that demonstrated particular suitability of OMMs for action representation, to achieve the above goal we have conducted an OMM-based clustering of the generated segments.

Based on clustering consistency measures, completeness and homogeneity, we have compared clusterings generated for all seven combinations of the three modalities. The unimodal clustering study has showed that the audio modality, yielding the highest overall consistency in comparison with the tactile and the joint-angles, is particularly well-suited. Furthermore, we have demonstrated that audio has to be combined with another modality (tactile or joint-angles) in order to distinguish uni- vs. bimanual action primitives. The best results have been achieved for the combination of the audio and the tactile modalities.

With the goal to investigate the influence of the execution speed variability on the clustering, we have discussed clusterings for two speed categories: the slow and the fast (on average executed two times faster than slow action primitives). The experimental results have demonstrated a slight decrease of consistency in the case when action primitives from different speed categories, i.e. slow and fast, have been used for training and testing respectively.

## 7.3 Final Comments

In this thesis we have contributed to the field of manual interaction recognition by proposing a generic framework for decomposition of multimodal bimanual interaction, based on a Bayesian offline change detection method developed by Fearnhead. Furthermore, we have showed that our approach can serve as a building block for a higher-level modeling of interaction. Finally, empirical studies have showed the significance of different modalities for both, the decomposition of interaction into action primitives, and their identification.

Following the above results, our attention is now directed towards two main issues. Firstly, we would like to address the challenge of online decomposition, by extending the presented offline decomposition approach to applications in online scenarios. Such an extension can be based on the procedure introduced by Fearnhead and Liu [28], and can be helpful for assisting interaction with robots and virtual agents. Another challenge is an extension of our approach to identification of actions and activities. Further on, the Activity Theory along with the overview provided by e.g. Bobick [13] (see Chapter 2) can serve as a conceptual basis for this task. However, to improve the robustness of the trimodal approach for a larger range of scenarios, including modeling of actions and activities, as well as to provide us with the higher-level insights into the semantics of interaction, we plan to pursue integration of further modalities in our approach, primarily the speech modality.

# Appendix A

# Instructions for Human Demonstrators

Here we present the instructions given to the human demonstrators, before a recording session in an unconstrained and in a constrained scenario.

## A.1 Action Execution: Unconstrained Scenario

The sequence is inspired by taking a bottle of juice, shaking it, opening it, pouring juice in a glass, and closing the bottle. Please carry out the following actions:

- **idle position**: The execution starts with the idle position for both hands: the palms are turned up, both hands are apart resting on the table;
- **pick up and lift** the bottle: grasp the bottle with both hands and lift it;
- **shake** the bottle (3 times);
- **put down** the bottle and release both hands;
- **unscrew the lid**: grasp the bottle with one hand and unscrew the lid of the bottle with the other. Conduct one continuous turning movement without releasing the hand in-between. Release both hands as soon as ready;
- **pick up and lift**: grasp the bottle with one hand and lift it;
- **pour**: tilt the bottle c.a. 100 degrees and bring it back to the original position;
- **put down** the bottle and release both hands;
- **screw the lid**: grasp the bottle with one hand, and screw the lid of the bottle with the other hand. Conduct one continuous turning movement without releasing the hand in-between. Release both hands as soon as ready.
- **idle position**.

## A.2 Action Execution with Cues: Constrained Scenario

The sequence is inspired by taking a bottle of juice, shaking it, opening it, pouring juice in a glass, and closing the bottle. The beginnings and the ends of action execution are signaled by start and end cues respectively. Each cue consists of four beep tones: three preparatory and one main. Please align the execution of the following actions as precise as possible with the beginning of the main cue:

- **idle position**: The execution starts with the idle position for both hands: the palms are turned up, both hands are apart resting on the table;
- **pick up and lift** the bottle (start cue): grasp the bottle with both hands and lift it;
- **hold** the bottle;
- **shake** the bottle (start and end cue): shake the bottle as long as the end cue is given;
- **hold**;
- **put down** (start cue): release the hands from the bottle and bring them in idle position;
- **idle**;
- **unscrew the lid** (start and end cue): grasp the bottle with one hand and unscrew the lid of the bottle with the other. Conduct one continuous turning movement without releasing the hand in-between. Release both hands on the end cue;
- **idle**;
- **pick up and lift** (start cue): grasp the bottle with one hand and lift it;
- **pour** (start and end cue): tilt the bottle c.a. 100 degrees and bring it back to the original position on the end cue;
- **hold**;
- **put down** (start cue);
- **idle**;
- **screw the lid** (start and end cue): grasp the bottle with one hand and screw the lid of the bottle with the other hand. Conduct one continuous turning movement without releasing the hand in-between. Release both hands on the end cue;
- **idle position**.

# Appendix B

# Annotation Rules

Here we present a detailed description of the rules used for video- and audio-based annotation of the acquired data. Five different annotation types will be described in the following subsections: tactile, audio, joint-angles, semantic and cue.

## B.1 Tactile Tier

The labeling of the tactile tier depends on the contact state of the hands. Dependent on whether the human demonstrator is holding the object in one or both hands, one of the following four labels is assigned:

- contact with both hands (LR),
- contact with left hand (Lr),
- contact with right hand (lR),
- no contact (lr).

The following table presents an overview of the tactile tier labeling:

| tactile tier label | description | uni-/bimanual (1/2) |
|---|---|---:|
| lr | no contact with both hands | 2 |
| LR | contact with both hands | 2 |
| lR | contact with right hand | 1 |
| Lr | contact with left hand | 1 |
| noise | irrelevant data | |

## B.2 Audio Tier

The audio tier annotation contains one name for each type of an audio event. The regions without audio are labeled by "off". The following labels are given to the regions corresponding to the different audio events:

- grasping of the object with one or two hands: grasp1 and grasp2s respectively,
- putting the object down with one or two hands: putdown1 and putdown2 respectively.
- object manipulations: shake, pour, pourup, pourdown.

The following table presents an overview of the audio tier labeling:

| audio tier label | description | 1/2 |
|---|---|---|
| grasp2s | symmetric grasp with two hands | 2 |
| shake | shaking | 2 |
| putdown2 | put object down with both hands | 2 |
| grasp1 | grasping with one hand | 1 |
| pour | pouring (down and up, if done together in one go) | 1 |
| pourdown | pouring (tilting the bottle down) | 1 |
| pourup | recovering the orientation of the bottle after pouring | 1 |
| putdown1 | put object down with one hand | 1 |
| off | no audio accompanying the manipulation | |
| noise | irrelevant data | |

## B.3 Joint-angles Tier

The joint-angles tier label collection contains different types of finger dynamics. Dependent on whether actions are conducted symmetrically or asymmetrically with both hands the movements are called "symmetric" or "asymmetric" (denoted by "s" and "as" resp.). Here we differentiate between reaching and grasping, releasing, turning and pausing hand-movements:

- grasping: close2s, close2as, close1
- releasing: open2s, open2as, open1
- screwing/unscrewing: turn2ccw, turn2cc
- pausing: hold2, hold1, idle

An overview of labeling of the hand posture tier is described in the following table:

| joint-angles tier label | description | 1/2 |
|---|---|---|
| idle | no dynamics | 2 |
| close2s | symmetrical closing of the hands before grasping | 2 |
| hold2 | hold the bottle | 2 |
| open2s | symmetrical opening of both hands, releasing of the bottle | 2 |
| close2as | asymmetrical closing of both hands before grasping | 2 |
| turn2ccw | turn the lid of the bottle counter clockwise | 2 |
| turn2cw | turn the lid of the bottle clockwise | 2 |
| open2as | asymmetrical opening of both hands, releasing of the bottle | 2 |
| close1 | closing of one hand before grasping | 1 |
| hold1 | hold the bottle | 1 |
| open1 | opening of one hand, releasing of the bottle | 1 |
| noise | irrelevant data | |

## B.4 Semantic Tier

The labeling of the semantic tier combines the joint, tactile and audio labeling. The following table presents an overview of the semantic label collection:

| semantic tier label | description | uni-/bimanual (1/2) |
|---|---|---|
| close2s | s. joint-angles tier | 2 |
| grasp2s | squeezing of the bottle during grasping | 2 |
| lift2 | lift the bottle | 2 |
| shake | shake the bottle | 2 |
| hold2s | hold the bottle | 2 |
| putdown2 | put the bottle down | 2 |
| open2s | s. joint-angles tier | 2 |
| close2as | s. joint-angles tier | 2 |
| turn2ccw | s. joint-angles tier | 2 |
| turn2cw | s. joint-angles tier | 2 |
| open2as | s. joint-angles tier | 2 |
| close1 | s. joint-angles tier | 1 |
| grasp1 | squeezing of the bottle during grasping | 1 |
| lift1 | lift the bottle | 1 |
| hold1 | hold the bottle | 1 |
| pour | tilt the bottle and bring it in the original pos. | 1 |
| putdown1 | put the bottle down | 1 |
| open1 | s. joint-angles tier | 1 |
| idle | no of hand posture, no audio, no tactile | |
| noise | irrelevant data | |

## B.5 Cues Tier

Cues tier is a label collection corresponding to the structure of audio cues. The following table presents the overview of the cue label collection:

| cue tier label | description | uni-/bimanual (1/2) |
|---|---|---|
| idle | idle position | 2 |
| grasp2 | grasp with both hands | 2 |
| shake | shake | 2 |
| hold2 | hold with both hands | 2 |
| putdown2 | put down with both hands | 2 |
| turn2ccw | turn counter clock-wise | 2 |
| grasp1 | grasp with one hand | 1 |
| pour | pour | 1 |
| hold1 | hold with one hand | 1 |
| putdown1 | putdown with one hand | 1 |
| turn2cw | turn clock-wise | 2 |
| noise | irrelevant data | |

# Appendix C

# Unimodal Segmentation

Here we describe unimodal segmentation experiments, conducted in order to explore the influence of global parameters, the prior segment length parameter $\lambda$ and the subsampling rate $s$. The following two subsections are dedicated to the audio and the joint-angles modalities.

## C.1 Audio Modality

The segmentations of audio modality has been conducted with two types of models: AR and constant models (see Sections 5.3.3.1 and 5.3.3.2 respectively).

### C.1.1 Autoregressive Model

In this subsection we describe the segmentation of audio signal based on a mixture of AR models of order 1,2 and 3. Due to the sensitivity of the contact microphone, not only audio-related action primitives, e.g. *shaking*, *pouring* or *putting down*, but also tactile-related action primitives, e.g. *grasping* or *screwing* can be detected in the audio signal. Therefore for evaluation of segmentation we use the combination of audio and tactile label collections generated from manual annotation (see Appendix B). For calculation of averages $\mu_t, \mu_g, \mu_r$ and $\mu_m$ we use segmentations generated for 20 trials recorded by one human demonstrator in a constrained scenario. The values of the clamping parameter $c = 5$ and scale range $\rho = 12$ remain constant in all experiments.

#### C.1.1.1 Subsampling Rate $s$

Representation of high frequency structures becomes less accurate due to growing subsampling rate $s$. This experiment examines the effect of growing subsampling rate $s \in \{5, 7, 10, 12, 20\}$ on the segmentation quality. Parameter $\lambda = 10^{-5}$ remained constant. The subsampling rate $s$ is applied after the default subsampling of 50 of the original raw signal of frequency of 44100 Hz.

The following table presents an overview of the experiment estimating the segmentation quality of audio modality with AR models for different values of the subsampling rate $s$:

| $s$ | $\mu_t$ | $\mu_g$ | $\mu_r$ | $\mu_m$ |
|---|---|---|---|---|
| 5 | 0.17 | 0.91 | 0.83 | 0.14 |
| 7 | 0.18 | 0.75 | 0.81 | 0.21 |
| 10 | 0.21 | 0.62 | 0.75 | 0.29 |
| 12 | 0.23 | 0.52 | 0.73 | 0.34 |
| 20 | 0.29 | 0.22 | 0.33 | 0.64 |

The table indicates that the increasing value of $s$ influences all four segmentation indices:

$$s \uparrow \;\; \rightarrow \;\; \mu_t \uparrow, \mu_g \downarrow, \mu_m \uparrow, \mu_r \downarrow . \tag{C.1}$$

For a larger subsampling rate the temporal error grows from 0.17 to 0.29 seconds and the segmentation granularity falls from 0.91 to 0.22. Therefore only a fifth of the ground truth change points are generated. The very small rate of generated segments $\mu_g$ corresponds to a large rate of missing segments with $\mu_m = 0.64$ for $s = 20$. The higher rate of subsampling eradicates some segmentation-relevant substructure, causing the signal of the complete trial to become more homogeneous. Figures C.1-C.3 show examples of the generated segmentation along with the ground truth segmentation for three different values: $s = 5, 12, 20$. The figures demonstrate an example of decreasing segmentation granularity for increasing values of $s$.

#### C.1.1.2 Prior Parameter $\lambda$

In this experiment the influence of the value of $\lambda \in \{10^{-4}, \ldots, 10^{-9}\}$ on segmentation with AR models has been investigated. The other parameters stayed constant at $s = 10$, $\rho = 12$. The following table presents an overview of the quality of segmentation of audio modality with AR models for different values of parameter $\lambda$:

| $\lambda$ | $\mu_t$ | $\mu_g$ | $\mu_r$ | $\mu_m$ |
|---|---|---|---|---|
| $10^{-4}$ | 0.21 | 0.72 | 0.86 | 0.19 |
| $10^{-5}$ | 0.23 | 0.61 | 0.87 | 0.22 |
| $10^{-6}$ | 0.22 | 0.55 | 0.85 | 0.24 |
| $10^{-7}$ | 0.22 | 0.59 | 0.84 | 0.28 |
| $10^{-8}$ | 0.22 | 0.59 | 0.84 | 0.30 |
| $10^{-9}$ | 0.23 | 0.51 | 0.83 | 0.34 |

Similar to the previous experiments, the table indicates a large impact of $\lambda$ on $\mu_g$ and on $\mu_m$:

$$\lambda \uparrow \;\; \rightarrow \;\; \mu_g \uparrow, \mu_m \downarrow . \tag{C.2}$$

At the same time, for the detected segments there is only a negligible change of the overlap ratio as well as the corresponding temporal error, implying that $\lambda$ does not have an impact on the position of the generated segments. Note, the influence of $\lambda$ differs from the one of the parameter $s$, which has demonstrated to have a strong effect on all segmentation indices.

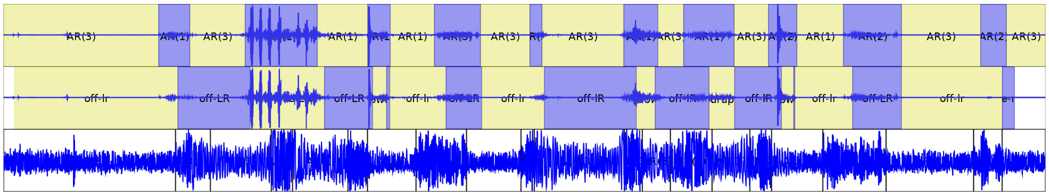


Figure C.1: An example of an audio signal segmentation with AR models of order 1,2 and 3; three subfigures show: generated segmentation (top), ground truth segmentation (middle) and input signal (bottom). For $s = 5$ the resulting segmentation is close to the ground truth structure.

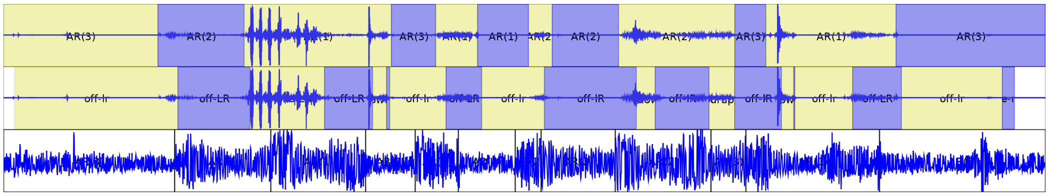


Figure C.2: An example of an audio signal segmentation with AR models of order 1,2 and 3; three subfigures show: generated segmentation (top), ground truth segmentation (middle) and input signal (bottom). Less segments are generated for $s = 12$ in comparison to $s = 5$.

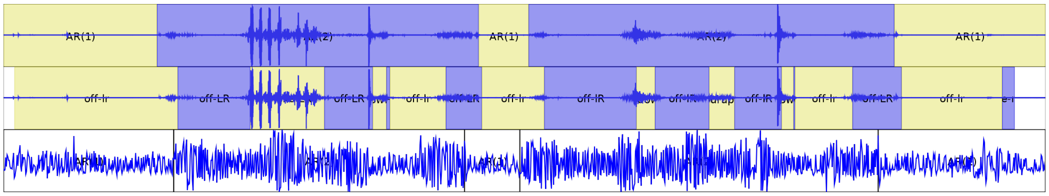


Figure C.3: An example of an audio signal segmentation with AR models of order 1,2 and 3; three subfigures show: generated segmentation (top), ground truth segmentation (middle) and input signal (bottom). For $s = 20$ the procedure selects only very rough structure within the signal.

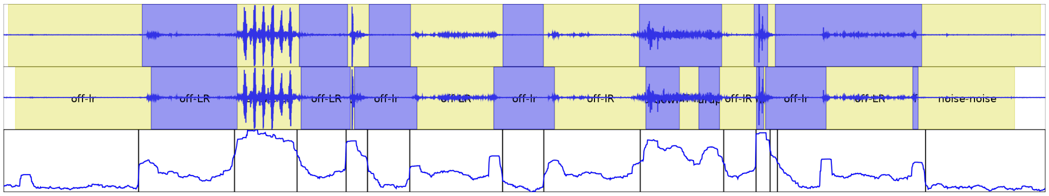


Figure C.4: An example of segmentation of audio data with constant models; $s = 16$ and $\lambda = 10^{-7}$. First row: generated segmentation; second row - ground truth segmentation, third row - the audio signal after preprocessing. Generated segmentation for this parameters is close to the ground truth.

### C.1.2 Constant Model

Following paragraphs are dedicated to segmentation experiments demonstrating the influence of the parameters $\lambda$ and $s$ on segmentation of audio with constant models. Parameter $w$ remains constant with $w = 20$. For building of averages $\mu_t, \mu_g, \mu_r$ and $\mu_m$ we have used 20 trails recorded by one human demonstrator in a constrained scenario. The ground truth is based on tactile and audio labels of the manual annotation, like in the previously described experiments with AR models.

#### C.1.2.1 Subsampling Rate $s$

This experiment examines the effect of growing subsampling rate, involving less accurate representation of high frequency structure, on the structure of change points, generated by segmentation of audio signal with constant models. For $\lambda = 10^{-6}$ we have investigated the segmentation for $s \in \{10, 12, 14, 16\}$. The following table presents the experimental results of segmentation of audio signal with constant model for different rates of subsampling:

| $s$ | $\mu_t$ | $\mu_g$ | $\mu_r$ | $\mu_m$ |
|---|---|---|---|---|
| 16 | 0.18 | 0.80 | 0.79 | 0.10 |
| 14 | 0.17 | 0.94 | 0.80 | 0.09 |
| 12 | 0.15 | 1.04 | 0.79 | 0.08 |
| 10 | 0.14 | 1.13 | 0.82 | 0.05 |

It indicates the following influence of the increasing value of $s$:

$$s \uparrow \quad \rightarrow \quad \mu_t \uparrow, \mu_g \downarrow, \mu_m \uparrow . \tag{C.3}$$

For a constant value of $\lambda$ smaller values of $s$ correspond to a higher segmentation granularity $\mu_g$ and a smaller missing segments index $\mu_m$. An example trial segmentation whose structure is close to ground truth is illustrated in Figure C.4.

#### C.1.2.2 Prior Parameter $\lambda$

As previously described, parameter $\lambda$ is designed to influence the granularity of the generated segmentation. In this section we examine the influence of $\lambda$ on the segmentation of audio signal with constant models.

Table C.1: Segmentation quality of the joint-angle modality with constant models for different values of $s$.

| $s$ | $\mu_t$ | $\mu_g$ | $\mu_r$ | $\mu_m$ |
|---|---|---|---|---|
| 5 | 0.12 | 1.57 | 0.65 | 0.04 |
| 10 | 0.15 | 0.94 | 0.82 | 0.05 |
| 15 | 0.19 | 0.82 | 0.87 | 0.11 |
| 20 | 0.25 | 0.71 | 0.88 | 0.23 |

For a constant value of $s = 10$ the experimental results for $\lambda \in \{10^{-9}, \ldots, 10^{-6}\}$ are presented in the following table:

| $\lambda$ | $\mu_t$ | $\mu_g$ | $\mu_r$ | $\mu_m$ |
|---|---|---|---|---|
| $10^{-9}$ | 0.15 | 0.99 | 0.81 | 0.09 |
| $10^{-8}$ | 0.15 | 1.02 | 0.80 | 0.08 |
| $10^{-7}$ | 0.15 | 1.05 | 0.80 | 0.06 |
| $10^{-6}$ | 0.14 | 1.13 | 0.82 | 0.05 |

The table indicates:

$$\lambda \uparrow \;\; \rightarrow \;\; \mu_g \uparrow, \mu_m \downarrow . \tag{C.4}$$

Like in the previous experiment, the values of temporal error $\mu_t$ and $\mu_r$ remain approximately constant for varying values of $\lambda$. This implies that the positions of the generated segments are invariant to the changes of $\lambda$. The granularity of segmentation $\mu_g$ increases along with the increase of $\lambda$: for higher values of $\lambda$ more segments are generated. Corresponding to this, missing segments index $\mu_m$ decreases for larger values of $\lambda$.

## C.2 Joint-angles Modality

We investigate the influence of global parameters $\lambda$ and $s$ on the segmentation of joint-angles with constant models. We present the results of the evaluation $\mu_t, \mu_g, \mu_r$ and $\mu_m$. For calculation of averages 20 trials recorded by one human demonstrator in a constrained scenario have been used. For ground truth we have used the joint-angles label collection of the manual annotation, consisting of different types of *grasping*, *releasing* and *screwing* (see Appendix B).

### C.2.1 Subsampling Rate $s$

In this experiment we show on several examples the influence of the subsampling rate $s$ on the segmentation of joint-angles modality with constant models. The following table demonstrates the experimental results for $s \in \{5, 10, 15, 20\}$ and $\lambda = 10^{-15}$:

For increasing $s$ the dependencies are as follows:

$$s \uparrow \;\; \rightarrow \;\; \mu_t \uparrow, \mu_g \downarrow, \mu_m \uparrow, \mu_r \uparrow . \tag{C.5}$$

The strong influence of $s$ on all segmentation indices is similar to the previous experiments. The increase of $s$ has a negative effect on the average temporal error $\mu_t$: the value increases

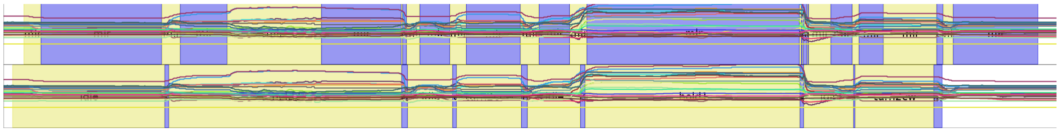


Figure C.5: Joint-angle segmentation for subsampling rate $s = 5$ results in oversegmentation; generated segmentation (first row); ground truth segmentation (second row).

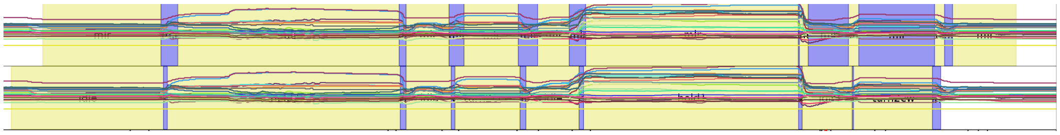


Figure C.6: Joint-angle segmentation for subsampling rate $s = 10$ results in segmentation structure that is close to ground truth; generated segmentation (first row); ground truth segmentation (second row).

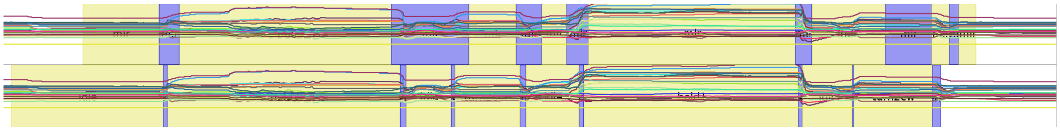


Figure C.7: Joint-angle segmentation for subsampling rate $s = 20$ results in too low segmentation granularity; generated segmentation (first row); ground truth segmentation (second row).

from 0.12 to 0.25 seconds; he decrease in segmentation granularity $\mu_g$ goes along with the larger segments, thus resulting in a larger overlap ratio $\mu_r$. With decreasing $s$ the missing segments index falls.

Figures C.5 - C.7 show examples of segmentation with three different subsampling rates: $s = 5, 10, 20$. These examples illustrate the effect of the increasing sampling rate on the segmentation granularity. Figure C.5 shows that for a small sampling rate $s = 5$ even *shaking* (third yellow region) has been segmented based on the subtle dynamics of the joint-angle modality. This segment disappears with larger subsampling rates. Figure C.5 demonstrates strong oversegmentations, resulting from constant model fitting in the regions of steep rise of the overall hand activity. Figure C.6 presents an example of a segmentation that matches well with the annotation. The generated segmentation mainly corresponds to *grasping* and *releasing* before and after an object contact. Figure C.7 shows that a further increase of the sampling rate results in a further decrease of the number of generated segments and in generation of insufficient number of segments.

### C.2.2 Prior parameter $\lambda$

This experiment aims at examining of the influence of varying value of the parameter $\lambda$ on the generated segmentation. Parameter $\lambda$ is designed to control the prior distribution on the segment length, therefore smaller values of $\lambda$ should result in increase of segmentation granularity. The following table shows the dependency between the value of $\lambda \in \{10^{-15}, \ldots, 10^{-5}\}$ and values of four segmentation indices. Subsampling rate is constant with $s = 10$:

| $\lambda$ | $\mu_t$ | $\mu_g$ | $\mu_r$ | $\mu_m$ |
|---|---|---|---|---|
| $10^{-15}$ | 0.15 | 0.94 | 0.82 | 0.05 |
| $10^{-13}$ | 0.15 | 1.01 | 0.80 | 0.05 |
| $10^{-11}$ | 0.14 | 1.08 | 0.78 | 0.05 |
| $10^{-9}$ | 0.15 | 1.41 | 0.82 | 0.03 |
| $10^{-7}$ | 0.14 | 1.40 | 0.72 | 0.04 |
| $10^{-5}$ | 0.12 | 1.76 | 0.64 | 0.04 |

The dependency is as follows:

$$\lambda \uparrow \quad \rightarrow \quad \mu_g \uparrow, \mu_r \downarrow . \tag{C.6}$$

For increasing values of $\lambda$ along with the slight decrease of temporal error $\mu_t$, the segmentation granularity increases from 0.94 to 1.76, the generated segments become smaller, the overlap index $\mu_r$ decreases.

# Bibliography

[1] J. K. Aggarwal and Sangho Park. Human motion: Modeling and recognition of actions and interactions. In *Proceedings of the 3D Data Processing, Visualization, and Transmission, 2nd International Symposium*, Washington, DC, USA, 2004. IEEE Computer Society.

[2] Eren Erdal Aksoy, Alexey Abramov, Florentin Wörgötter, and Babette Dellen. Categorizing object-action relations from semantic scene graphs. In *ICRA*. IEEE, 2010.

[3] R. Andre-Obrecht. A new statistical approach for the automatic segmentation of continuous speech signals. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 36(1), 1988.

[4] Pedram Azad, Tamim Asfour, and Rudiger Dillmann. Toward an Unified Representation for Imitation of Human Motion on Humanoids. In *Robotics and Automation, 2007 IEEE International Conference on*, 2007.

[5] A. Barchunova, R. Haschke, M. Franzius, and H. Ritter. Multimodal segmentation of object manipulation sequences with product models. In *ICMI*, 2011.

[6] A. Barchunova, J. Moringen, R. Haschke, and H. Ritter. Hierarchical bayesian modeling of manipulation sequences from bimodal input. In *International Conference on Cognitive Modeling*, 2012.

[7] Alexandra Barchunova, Robert Haschke, Ulf Grossekathoefer, Sven Wachsmuth, Herbert Janssen, and Helge Ritter. Unsupervised segmentation of object manipulation operations from multimodal input. In Barbara Hammer and Thomas Villmann, editors, *New Challenges in Neural Computation*, Machine Learning Reports. 2011.

[8] D. Barry and J.A. Hartigan. A bayesian analysis for change point problems. *Journal of the American Statistical Society*, 1993.

[9] Daniel Barry and J.A. Hartigan. Product partition models for change point problems. *The Annals of Statistics*, 1992.

[10] Michèle Basseville and Igor V. Nikiforov. *Detection of Abrupt Changes: Theory and Application.* 1993.

[11] K. Bernardin, K. Ogawara, K. Ikeuchi, and R. Dillmann. A hidden Markov model based sensor fusion approach for recognizing continuous human grasping sequences. In *Humanoid Robots*, 2003.

[12] Gerald Bieber, Jörg Voskamp, and Bodo Urban. Activity recognition for everyday life on mobile phones. In *HCI (6)*, 2009.

[13] A. F. Bobick. Movement, activity and action: The role of knowledge in the perception of motion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 352(1358), 1998.

[14] A. F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23, 2001.

[15] S. Bødker. *Through the Interface.* Lawrence Erlbaum Publishers, 1990.

[16] Sylvain Calinon. *Continuous Extraction of Task Constraints in a Robot Programming by Demonstration Framework.* PhD thesis, Ecole Polytechnique Federale de Lausanne (EPFL), 2007.

[17] G. Calvert, Ch. Spence, and B. Stein, editors. *The Handbook of Multisensory Processes.* MIT Press, 2004.

[18] C Castellini and R. Kõiva. Using surface electromyography to predict single finger forces. In *IEEE International Conference on Biomedical Robotics and Biomechatronics*, June 2012.

[19] F. Chersi, A. Mukovskiy, L. Fogassi, P. F. Ferrari, , and W. Erlhagen. A model of intention understanding based on learned chains of motor acts in the parietal lobe. *Computational Neuroscience*, 2006.

[20] Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu.

[21] Philip R. Cohen and David R. Mcgee. Tangible multimodal interfaces for safety-critical applications. *Communications of the ACM*, 47, 2004.

[22] Naresh P. Cuntoor, B. Yegnanarayana, and Rama Chellappa. Activity modeling using event probability sequences. *IEEE Transactions on Image Processing*, 17(4), 2008.

[23] Immersion CyberGlove II. http://www.cyberglovesystems.com/products/cyberglove-ii/overview.

[24] Dominik Endres, Andrea Christensen, Omlor L., and Martin A. Giese. Segmentation of Action Streams: Human Observers vs. Bayesian Binning. *Advances in Artificial Intelligence*, 75:8, 2011.

[25] P. Fearnhead. http://www.maths.lancs.ac.uk/ fearnhea/software/ARPS.html.

[26] Paul Fearnhead. Exact bayesian curve fitting and signal segmentation. *Signal Processing*, 53, 2005.

[27] Paul Fearnhead. Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, 2006.

[28] Paul Fearnhead and Zhen Liu. On-line inference for multiple change points problems. *Journal of the Royal Statistical Society B*, 69, 2007.

[29] J.R. Flanagan and R.S. Johansson. Action plans used in action observation. *Nature*, 2003.

[30] M. A. Girshick and Herman Rubin. A bayes approach to a quality control model. *Ann. Math. Statist.*, 1952.

[31] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12), December 2007.

[32] Peter J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82, 1995.

[33] Ulf Großekathöfer, Amir Sadeghipour, Thomas Lingner, Peter Meinicke, Thomas Hermann, and Stefan Kopp. Low Latency Recognition and Reproduction of Natural Gesture Trajectories. In *ICPRAM (Int.Conf. on Pattern Recognition Applications and Methods)*, 2012.

[34] U. Grossekathöfer, A. Barchunova, R. Haschke, T. Hermann, M. Franzius, and H. Ritter. Learning of object manipulation operations from continuous multimodal input. In *Humanoids*. IEEE, 2011.

[35] Tobias Großhauser, Ulf Großekathöfer, and Thomas Hermann. New Sensors and Pattern Recognition Techniques for String Instruments. In *International Conference on New Interfaces for Musical Expression*, 2010.

[36] Ulf Großekathöfer and Thomas Lingner. Neue ansätze zum maschinellen lernen von alignments. Master's thesis, Bielefeld University, 2004.

[37] Gutemberg Guerra-Filho and Arnab Biswas. The human motion database: A cognitive and parametric sampling of human motion. *Image and Vision Computing*, 2011.

[38] P. E. Hemeren. *Mind in Action: Action Representation and the Perception of Biological Motion.* Lund University, 2008.

[39] P.E. Hemeren and S. Thill. Deriving motor primitives through action segmentation. *Front. Psychology*, 2011.

[40] Jesse Hoey, Thomas Plötz, Dan Jackson, Andrew Monk, Cuong Pham, and Patrick Olivier. Rapid specification and automated generation of prompting systems to assist people with dementia. *Pervasive and Mobile Computing*, 7(3), 2011.

[41] L. Hoste, B. Dumas, and B. Signer. SpeeG: A Multimodal Speech- and Gesture-based Text Input Solution. In *Proceedings of AVI 2012, 11th International Working Conference on Advanced Visual Interfaces*, Naples, Italy, May 2012.

[42] Aitor Ibarguren, Iñaki Maurtua, and Basilio Sierra. Layered architecture for real-time sign recognition. *Comput. J.*, 53, October 2010.

[43] Julie A. Jacko and Andrew Sears, editors. *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 2003.

[44] O Ruanaidh J. J. K. and Fitzgerald. *Numerical Bayesion Methods Applied to Signal Processing*. Springer, 1996.

[45] Victor Kaptelinin. *Activity Theory*. The Interaction-Design.org Foundation, Aarhus, Denmark, 2012.

[46] H. Kawasaki, K. Nakayama, and G. Parker. Teaching for multi-fingered robots based on motion intention in virtual reality. In *IECON*, 2000.

[47] Andrew J. King. Multisensory integration. *Science*, 1993.

[48] Nathan Koenig. Behavior-based segmentation of demonstrated task. In *In International Conference on Development and Learning*, 2006.

[49] Jens Kohlmorgen and Steven Lemm. A dynamic hmm for on-line segmentation of sequential data. In *Advances in Neural Information Processing Systems 14 (NIPS 2001*. MIT Press, 2002.

[50] Charles T. Krebs. Multi-sensory neurons. a new paradigm in sensory processing. In *Brain Gym Conference*, 2010.

[51] V. Krüger, D. Kragic, A. Ude, and C. Geib. The Meaning of Action: A Review on action recognition and mapping. *Advanced Robotics*, 21(13), 2007.

[52] Volker Krüger and Daniel Grest. Using hidden markov models for recognizing action primitives in complex actions. In *Proceedings of the 15th Scandinavian conference on Image analysis*, SCIA'07, Berlin, Heidelberg, 2007. Springer-Verlag.

[53] Dana Kulic, Wataru Takano, and Yoshihiko Nakamura. Online segmentation and clustering from continuous observation of whole body motions. *Trans. Rob.*, 25, October 2009.

[54] A. Leontiev. *Activity, consciousness, and personality*. Prentice-Hall, 1978.

[55] Vaia Lestou, Frank E. Pollick, and Zoe Kourtzi. Neural substrates for action understanding at different description levels in the human brain. *J. Cognitive Neuroscience*, 20, February 2008.

[56] C. Li, P.R. Kulkarni, and B. Prabhakaran. Motion Stream Segmentation and Recognition by Classification. In *ICASSP*. IEEE, 2006.

[57] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2011.

[58] Ana Paula Brandão Lopes, Eduardo Alves do Valle Jr., Jussara Marques de Almeida, and Arnaldo de Albuquerque Araújo. Action recognition in videos: from motion capture labs to the web. *CoRR*, abs/1006.3506, 2010.

[59] Kazuya Matsuo, Kouji Murakami, Tsutomu Hasegawa, Kenji Tahara, and Kurazu Ryo. Segmentation method of human manipulation task based on measurement of force imposed by a human hand on a grasped object. In *IROS*. IEEE, 2009.

[60] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.*, 104(2), November 2006.

[61] V Moskvina. Application of the singular-spectrum analysis to change-point detection in time series. *Communication in Statistics Simulation Computation*, 44(0), 2003.

[62] Hans-Hellmut Nagel. Analysing sequences of tv-frames. In *IJCAI*, 1977.

[63] Hans-Hellmut Nagel. From image sequences towards conceptual descriptions. *Image and Vision Computing*, 6, 1988.

[64] A.S. Ogale, A. Karapurkar, and Y. Aloimonos. View-invariant modeling and recognition of human actions using grammars. In *Workshop on Dynamical Vision at ICCV*, volume 5. Springer, 2005.

[65] Georg Ogris, Thomas Stiefmeier, Paul Lukowicz, and Gerhard Troster. Using a complex multi-modal on-body sensor system for activity spotting. *Wearable Computers, IEEE International Symposium*, 2008.

[66] Antonios Oikonomopoulos, Ioannis Patras, Maja Pantic, and Nikos Paragios. Trajectory-based representation of human actions. In *Proceedings of the ICMI 2006 and IJCAI 2007 international conference on Artifical intelligence for human computing*, Berlin, Heidelberg, 2007. Springer-Verlag.

[67] Sharon Oviatt and Philip Cohen. Perceptual user interfaces: multimodal interfaces that process what comes naturally. *Commun. ACM*, 43, March 2000.

[68] E.S. Page. Continuous inspection schemes. *Biometrika*, 1954.

[69] J. Park, S. Park, and JK Aggarwal. Model-based human motion tracking and behavior recognition using hierarchical finite state automata. *Computational Science and Its Applications-ICCSA 2004*, 2004.

[70] D.J. Poirier. *The Econometrics of Structural Change*. 1976.

[71] Ronald Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6), June 2010.

[72] T. Pozzo, C. Papaxanthis, J. L. Petit, N. Schweighofer, and N. Stucchi. Kinematic features of movement tunes perception and action coupling. *Behavioural Brain Research*, 2006.

[73] http://opportunity-project.eu/.

[74] Elena Punskaya, Christophe Andrieu, Arnaud Doucet, and William J. Fitzgerald. Bayesian curve fitting using mcmc with applications to signal segmentation. *IEEE Transactions on Signal Processing*, 50, 2002.

[75] L.R. Rabiner and B.H. Juang. An introduction to hidden markov models. *IEEE Acoust., Speech, Signal Process. Mag.*, 1986.

[76] Nimrod Raiman, Hayley Hung, and Gwenn Englebienne. Move, and i will tell you who you are: detecting deceptive roles in low-quality data. In *Proceedings of the 13th international conference on multimodal interfaces*, New York, NY, USA, 2011. ACM.

[77] S. Richardson and P. Green. On Bayesian analysis of mixtures with unknown number of components. *J. Roy. Stat. Soc. B*, 1997.

[78] Giacomo Rizzolatti and Laila Craighero. The mirror-neuron system. *Annual Review of Neuroscience*, 27, 2004.

[79] Andrew Rosenberg and Julia Hirschberg. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007.

[80] Jan Steffen, Michael Pardowitz, and Helge Ritter. A manifold representation as common basis for action production and recognition. In *32nd German Conference on Artificial Intelligence*, Paderborn, Germany, 2009. Springer Berlin Heidelberg, Springer Berlin Heidelberg.

[81] Barry E. Stein and M. Alex Meredith. *The merging of the senses.* 1993.

[82] Thomas Stiefmeier, Georg Ogris, Holger Junker, Paul Lukowicz, and Gerhard Troster. Combining motion sensors and ultrasonic hands tracking for continuous activity recognition in a maintenance scenario. *Wearable Computers, IEEE International Symposium*, 0, 2006.

[83] Dag Svanaes. *Understanding Interactivity: Steps to a Phenomology of Human-Computer Interaction.* PhD thesis, Norwegian University of Science and Technology, 2000.

[84] W. Takano. *Stochastic Segmentation, Proto-Symbol Coding and Clustering of Motion Patterns and Their Application to Signifiant Communication between Man and Humanoid Robot.* PhD thesis, University of Tokyo, 2006.

[85] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11), 2008.

[86] M. Wallace, R. Ramachandran, and B. Stein. A revised view of sensory cortical parcellation. *Proceedings of the National Academy of Science USA*, 2004.

[87] Jamie A. Ward, Paul Lukowicz, Gerhard Troster, and Thad E. Starner. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *TPAMI*, 2006.

[88] Nils-Christian Wöhler, Ulf Großekathöfer, Angelika Dierker, Marc Hanheide, Stefan Kopp, and Thomas Hermann. A calibration-free head gesture recognition system with online capability. In *Pattern Recognition*, Istanbul, Turkey, 2010.

[89] Xiang Xuan and Kevin Murphy. Modeling changing dependency structure in multivariate time series. In *ICML*. ACM, 2007.

[90] Zahoor Zafrulla, Helene Brashear, Thad Starner, Harley Hamilton, and Peter Presti. American sign language recognition with the kinect. In *Proceedings of the 13th international conference on multimodal interfaces*. ACM, 2011.

[91] R. Zöllner and R. Dillmann. Using multiple probabilistic hypothesis for programming one and two hand manipulation by demonstration. In *IROS*. IEEE, 2004.