Hierarchical Bayesian Modeling of Manipulation Sequences from Bimodal Input

Alexandra Barchunova (abarch@cor-lab.uni-bielefeld.de)
Bielefeld University, Cor-Lab, Universitaetsstrasse 21-23
Bielefeld Germany

Jan Moringen, Robert Haschke, Helge Ritter
Bielefeld University, Universitaetsstrasse 21-23
Bielefeld Germany

Abstract

We propose a hierarchical approach for Bayesian modeling and segmentation of continuous sequences of bimanual object manipulations. Based on bimodal (audio and tactile) low-level time series, the presented approach identifies semantically differing subsequences. It consists of two hierarchically executed stages, each of which employs a Bayesian method for unsupervised change point detection (Fearnhead, 2005). In the first step we propose to use a mixture of model pairs for bimanual tactile data. To this end, we select "object interaction" and "no object interaction" regions for the left and the right hand synchronously. In the second step we apply a set of Autoregressive (AR) models to the audio data. This allows us to select regions within "object interaction" segments according to qualitative changes in the audio signal. Two simple model types that allow the calculation of modality-specific segment likelihoods serve as a foundation for this modeling approach. Based on the acquired ground truth, empirical evaluation has showed that the generated segments correctly capture the semantic structure of the test time series. The developed procedure can serve as a building block for higher-level action and activity modeling frameworks.

Introduction

An important objective of today's interdisciplinary research of human-machine interaction is machine perception of human action and activity (Krüger, Kragic, Ude, & Geib, 2007), (Bobick & Davis, 2001), (Turaga, Chellappa, Subrahmanian, & Udrea, 2008), (Aggarwal & Park, 2004). Areas like cognitive and social robotics, artificial intelligence, ambient intelligence, sports science and neurobiology collaborate on understanding of the mechanisms of human movements. Research in cognitive robotics is aimed towards enabling robots to interact with humans in everyday scenarios. Within this area, we focus on the topic of autonomous identification of bimanual object manipulations from low-level bimodal observation sequences. In order to participate in a simple interaction scenario or learn from a human, a robot needs the ability to autonomously single out relevant parts of the movement executed by a human.

Analysis of various sensor readings describing the human hand dynamics during manual interaction have been conducted recently by different researchers (Bernardin, Ogawara, Ikeuchi, & Dillmann, 2005; Dillmann, Rogalla, Ehrenmann, Zöllner, & Bordegoni, 2000; Kawasaki, Nakayama, & Parker, 2000). In general, one is interested in autonomous identification of action primitives in the context of imitation learning and human-machine interaction (Sanmohan, Krüger, & Kragic, 2010; Takano & Nakamura, 2006). Within this domain, Matsuo et al. focused on force feedback (Matsuo, Murakami, Hasegawa, Tahara, & Ryo, 2009) while a combination of different sensors like CyberGlove, Vicon or magnetic markers and tactile sensors has been used by (Pardowitz, Knoop, Dillmann, & Zöllner, 2007), (Kawasaki et al., 2000) and (Li, Kulkarni, & Prabhakaran, 2006). In (Zöllner, Asfour, & Dillmann, 2004) a bimanual approach is described. Audio and ultra-wide band tags have been successfully used in (Ogris, Stiefmeier, Lukowicz, & Troster, 2008) and (Ward, Lukowicz, Troster, & Starner, 2006).

Identification and learning of manual action primitives from continuous sequences is still an open question. We address it by proposing a novel hierarchical approach for uni- and bimanual time series. The bimodal approach is inspired by the fact, that humans employ different perception channels like hearing, proprioception, haptics and vision. Furthermore, our recent work has showed that audio and tactile data can generate a symbolic sequence of action primitives of sufficient granularity and semantic content (Barchunova, Haschke, Franzius, & Ritter, 2011). The automatically extracted action primitives have been successfully used in a classification application with HMM-based models (Grossekathöfer et al., 2011). During a considerable number of simple object manipulations (e.g. grasping, shifting, shaking, pouring, stirring or rolling) application of force is naturally accompanied by specific types of sound. We exploit this fact by performing modeling and segmentations based on the analysis of the audio signal structure and contact forces recorded on the fingertips. Our method handles three main challenges arising in automated modeling of action sequences: (i) inter- and intrapersonal variance of sensor data, (ii) absence of prior knowledge about the structure of the action sequence (i.e. location, type and number of action primitives) (iii) modality fusion. The data recorded for different human demonstrators exhibits a high degree of interpersonal and intrapersonal variance. However, our method solely depends on the temporal structure of the data and is invariant to absolute data values, the speed of action execution, way of grasping or the manipulation object. Furthermore, the output is person-invariant. Our method does not employ any specific knowledge about the components of the action sequence. Based on two simple models, the modeling does not require a large set of domain-specific heuristics describing each action primitive as is commonly the case in similar approaches (Pardowitz et al., 2007; Kawasaki et al., 2000; Zöllner & Dillmann, 2004). Due to the simplicity of these two fundamental

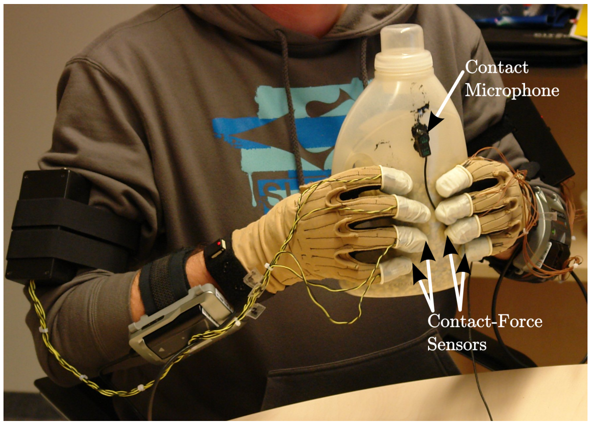

Figure 1: Experimental setup: a human demonstrator wearing contact sensors performs manipulation operations with a plastic bottle equipped with a structure-borne microphone. The CyberGlove trajectories also recorded in this scenario are not evaluated in this work.

models and the modeling concepts used within our approach, the developed procedure can be easily used in a wide range of scenarios, like imitation learning, cooperation and assistance. Because the segmentation steps for individual modalities are executed hierarchically, no additional multimodal fusion (e.g. (Ogris et al., 2008)) is necessary.

We evaluate our method in an everyday scenario in which a human demonstrator performs several object manipulation operations with a large non-rigid plastic bottle with a handle. In this evaluation, we assess the performance of the segmentation method w.r.t. the accuracy of the generated segment borders. The rest of this paper is organized as follows: Sec. ”Experimental Setup” explains the acquisition of action sequences within the scenario. Sec. ”Segmentation Method” introduces the two steps of the proposed method. In Sec. ”Evaluation” we discuss our evaluation method and experimental results of the procedure, Sec. ”Conclusion and Outlook” concludes the paper with a brief discussion and outlook.

Experimental Setup

In our scenario, a human demonstrator performs sequences of simple uni- and bimanual object manipulations with a gravel-filled plastic bottle[1], as can be seen in Fig. 1.

We use two types of sensors to record the time series of the performed action sequences (corresponding modality names used in formulas appear in parentheses):

- A structure-borne microphone AKG C411 L attached to the bottle records an audio signal (a), which is focused on in-object generated sound, ignoring most environmental noise.
- 2×5 FSR pressure sensors attached to the fingertips of each CyberGlove (t: both hands, tl: left hand, tr: right hand) record the contact forces.

The human demonstrator was instructed to perform a sequence of basic manipulation actions in the fixed order showed in the enumeration below. To obtain ground truth for later evaluation of computed segment borders we have used two methods of ground truth acquisition: manually annotated (*unconstrained*) and automated cue-driven (*constrained*). In the *unconstrained* scenario the human demonstrator was asked to conduct the sequence at her/his natural speed. The annotation of the action sequences has been conducted based on a video recording of the interaction scene. Within the cue-driven *constrained* scenario the aspired beginning or end of an action within a sequence was signalled to the human demonstrator via headphones as explained in (Barchunova, Haschke, Franzius, & Ritter, 2011). To achieve a rich variance for individual action primitives between different trials in the *constrained* scenario, we added Gaussian noise to the nominal time of the action primitives as specified in parentheses:

1. pick up and hold the bottle with both hands (2 s + η_1)
2. shake the bottle with both hands (.7 s + η_2)
3. hold the bottle with both hands (.3 s + η_3)
4. put down the bottle and pause (1 s + η_4)
5. unscrew the cap with both hands (1.2 s + η_5)
6. release cap and pause (1 s + η_6)
7. grasp and lift the bottle with right hand (2 s + η_7)
8. pour with right hand (1 s + η_8 + 1 s + η_9)
9. hold the bottle (.3 s + η_{10})
10. put down the bottle and pause (1 s + η_{11})
11. screw the cap with both hands (1.2 s + η_{12})

The random variables $\eta_i \sim \mathcal{N}(0, .5\ s)$ denote the randomized timing of subsequences. The overall length of the time series of a trial accumulates to approximately 30 seconds. Both annotation methods have specific advantages and disadvantages. The cue-driven annotation is a completely automated way of ground truth acquisition, avoiding time-consuming manual annotation but putting constraints on the execution speed and sequence. This method is suitable for acquiring ground truth for large number of trials. Manual annotation is more precise and more time-consuming, but it does not rely on perfect adherence to audio cues by the human demonstrator.

[1] The use of gravel instead of liquid is due to safety concerns. We have used liquids in a similar scenario restricted to the audio modality.

Segmentation Method

The recorded time series of multiple sensors capture complex descriptions of action sequences. This section describes how such time series data is segmented and modeled. Our segmentation approach applies Fearnhead's method (Fearnhead, 2005) previously used for unsupervised detection of multiple change points in one-dimensional time series. In his work Fearnhead describes a deterministic method that maximizes the posterior distribution of the number and location of change points w.r.t given observations. The estimated change points are optimal in the sense that a combination of a prior distribution on segmentations and segment-wise likelihoods is maximized. The segment likelihood is computed with respect to a single model chosen from a fixed set of models. Our approach combines a preprocessing step with a set of simple, modality-specific models to enable a probabilistic description of the recorded time series as the basis for applying Fearnhead's algorithm (introduced in Sec. "Bayesian Segmentation"). In the preprocessing step each sensory channel is reduced to a compact scalar description capturing its temporal structure. Two basic data models, an autoregressive and a threshold model, are employed within the procedure for the channel-specific modeling (see Sec. "Basic Data Models"). Their association to the tactile and audio modalities is explained in Sec. "Signal Preprocessing for Basic Models". Finally, in the Sec. "Two-Stage Segmentation" the two stages of the procedure – the segmentation based on the tactile modality and the subsegmentation based on audio – is described.

Bayesian Segmentation

In general, Fearnheads' algorithm segments an arbitrary time series $y_{1:T}$ by determining a set of change points $1 < \tau_1 < \cdots < \tau_N < T$ at which qualitative changes occur in the data. Within the probabilistic framework of Fearnhead's algorithm, the optimal segmentation is obtained by maximizing the Bayesian posterior[2] $P(y_{1:T} \mid \tau_{1:N})P(\tau_{1:N})$ which consists of a likelihood term and a prior distribution over segmentations $P(\tau_{1:N})$. In a common choice of this prior, the probability $P(\tau_{1:N})$ is composed of probabilities of individual segment lengths which are computed according to the geometric distribution $P(l) = \lambda(1-\lambda)^{l-1}$. Consequently, the prior is characterized by a single parameter λ that is reciprocal to the expected segment length under a geometric distribution, i.e. $\lambda \propto 1/u$ where u is the expected length of subsequences. Once λ has been chosen, neither the number of change points N nor any information regarding their positions have to be specified in advance.

The likelihood term $P(y_{1:T} \mid \tau_{1:N})$ is the probability, that the observed time series originates from a set of given models, which are fixed over the period of an individual segment. To this end, a finite set of models $\mathcal{M}$ is employed. Given a particular model $m \in \mathcal{M}$, the marginal likelihood $P(y_{s:t} \mid m)$ is the probability, that the entire subsequence $y_{s:t}$ can be explained by this model. Prior probabilities $P(m)$ can be associated with all models to reflect their relative frequency.

Basic Data Models

In order to locally represent the preprocessed sensor data we employ two simple kinds of probabilistic models: a *threshold model* and a set of autoregressive models AR(1), AR(2), AR(3).

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

$$P(y_{s:t} \mid m_{<\gamma}) = \prod_{k=s}^{t} p(y_k | m_{<\gamma}), \tag{1}$$

$$\text{where} \quad p(y_k \mid m_{<\gamma}) = \begin{cases} 1, \text{ if } y_k < \gamma \\ p_o \text{ otherwise} \end{cases} \tag{2}$$

where $p(y_k | m_{<\gamma})$ is the probability, that a single sample y_k fits the model assumption. The parameter p_0 is the probability of a simple data point y_k being an outlier w.r.t. the model. Denoting the segment length by $u = t - s$ and the number of not fitting samples by $n = |\{y_k > \gamma \mid s \leq k < t\}|$, and ignoring the constant normalization factor, we can derive the following, more compact formulas for both models:

$$P(y_{s:t} \mid m_{<\gamma}) = {p_o}^n \quad \text{and} \quad P(y_{s:t} \mid m_{>\gamma}) = {p_o}^{u-n} \tag{3}$$

As can be seen from Eq. 3, the marginal likelihood becomes smaller, the more data points are on the wrong side of the threshold.

The Autoregressive model is a special case of a general linear model $y_{s:t} = G^{(p)}_{s:t}\beta + \varepsilon$, where β and ε denote the parameter vector and white noise respectively. The matrix of the basis vectors for the autoregressive model of order $p = 3$ is defined as follows:

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

Please refer to (Fearnhead, 2005), Section II and III.B for the method of likelihood calculation for this model.

Signal Preprocessing for Basic Models

The preprocessing steps are modality-specific and facilitate subsequent likelihood calculations.

Tactile signal. The tactile feedback is susceptible to strong noise and large variations within action primitives (e.g. during shaking). Thus, tactile values for each hand are summed up to yield a cumulative tactile force for each time spot. The *threshold models* are applied to this scalar time series to discriminate "object contact" from "no object contact" for each

[2] We suppress the constant normalization factor $P(y_{1:T})^{-1}$.

Table 1: Overview of channel-specific models.

Sensor channel	Model	Notation
tactile sum left	threshold	m_L, m_l
tactile sum right	threshold	m_R, m_r
audio signal	AR	$m_{AR(1)}$, $m_{AR(2)}$, $m_{AR(3)}$

hand. The parameter γ specifies the threshold for recognizing hand-object contact. We denote the "object contact"-models with capital-letter subscripts: m_L and m_R for the left and right hand respectively. The corresponding notations for the "no object contact"-models are m_l and m_r. Note that the assignment of a "contact" or "no-contact" model by the segmentation method automatically yields an identification of the contact status during the segment.

Audio signal. Often, actions are accompanied by a typical sound, whose structure and volume remains approximately constant during the whole action primitive. Consider for example shaking an object or pouring water into a glass. Also segment boundaries are sometimes accompanied by a short, but strong change of the audio signal, e.g. placing or dropping an object.

Hence, we consider the local oscillating structure of the recorded audio signal. The signal is also subsampled and recording artifacts are removed by discarding samples whose amplitude exceeds a specified threshold. The resulting time series is logarithmized and whitened to normalize it to a given variance range w.r.t. amplitudes of individual samples. To the preprocessed data we apply the autoregressive models denoted by $m_{AR(1)}$, $m_{AR(2)}$ and $m_{AR(3)}$.

The Table 1 summarizes the association of data models to sensor channels.

Two-stage Segmentation

In our two-stage segmentation approach, we use tactile information to obtain a rough split of the sequence into subsequences of "object interaction" and "no object interaction". Subsequences that have been recognized as "object interaction" are analyzed in detail w.r.t. qualitative changes of the audio signal in order to refine the rough segmentation.

In the following two subsections, we describe the application of Fearnhead's algorithm to bimanual tactile data (first segmentation step) and to audio modality (second subsegmentation step). This is based on two respective sets of models $\mathcal{M}$ and $\mathcal{M}_{\text{sub}}$. Hereby $\mathcal{M}$ consists of a mixture of product models based on the threshold model applied in the first step; $\mathcal{M}_{\text{sub}}$ is a set of simple AR models, which is applied in the second step. The two-stage application of the segmentation procedure and the modality-specific local and bimanual models constitute the main contributions of this paper.

Segmentation Based on Tactile Modality The first step performs a rough joint analysis of the tactile signals of both hands. The analysis of bimanual data is based a mixture of four pairs of threshold models $\mathcal{M}$, combined in a multiplicative way. Each pair corresponds to a particular contact state of the left and the right hand at once. All possible combinations of pairs define the following set $\mathcal{M} := \{m_{lr}, m_{Lr}, m_{lR}, m_{LR}\}$, where "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 likelihoods of these joint models are computed as products of the individual likelihoods, e.g.:

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

An overview of the notation can be found in the Table 2. Assignments of the four joint contact-state models to segments in a computed segmentation are illustrated in the first row of Fig. 2. Contact assignments identify parts of the time series that are directly associated with object interactions. With this approach no additional fusion is necessary for modeling of the bimanual tactile data. The contact state for both hands is determined in one pass.

Table 2: Overview of notation used for product models.

Notation	Description
m_{lr}	no contact for both hands
m_{lR}	contact for right hand
m_{Lr}	contact for left hand
m_{LR}	contact for both hands

In contrast to a pointwise application of threshold methods, Fearnhead's method – even when used with threshold models – is not sensitive to noise which could otherwise lead to severe oversegmentation with many extremely small segments.

Sub-segmentation of Object-Contact Segments Based on Audio Modality In this subordinate segmentation step, all segments related to object interaction are further subsegmented employing the audio modality and using Fearnhead's method once again. This time, the signal is assumed to be produced by Auto-Regressive (AR) models of order 1, 2 or 3: $\mathcal{M}_{\text{sub}} = \{m_{AR(1)}, m_{AR(2)}, m_{AR(3)}\}$. Thus the subsegmentation is formed by selecting segments that exhibit homogeneous oscillatory properties within the audio modality. The sequential application of segmentation and selection steps yields a set of segments that are characterized by constant contact topology in respect to overall hand activity as well as homogeneous characteristics of the audio signal.

Evaluation

We recorded 60 trials of the action sequence described in Sec. "Experimental Setup" with three subjects. This corresponds to ca. 10^5 data points and $60 \times 11 = 660$ expected change points in total. For each subject, 10 trials have been recorded with automated cue-based scheduling for ground truth and 10 trials have been manually annotated. Cue-based ground truth has been described in our previous work (Barchunova, Haschke, Franzius, & Ritter, 2011). In order

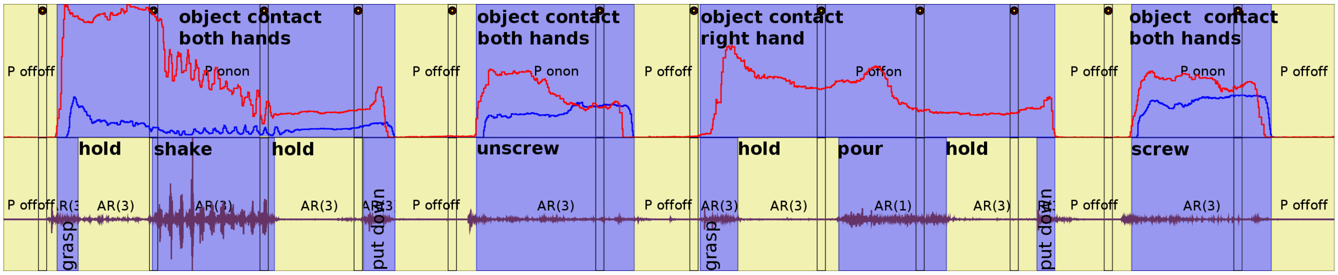

Figure 2: Hierarchical segmentation of a multimodal time series of action primitives: The first row shows the summed tactile signals of both hands and the resulting segmentation (cf. Sec. "Segmentation Based on Tactile Modality"). The segments indicated by an alternating coloring, perfectly match the underlying object-interaction structure. The second row displays the raw audio signal and the more detailed segmentation structure gained from refining object-interaction segments obtained from tactile-based segmentation (cf. Sec. "Segmentation Based on Audio Modality"). This subordinate segmentation step can correctly identify the subsequences: grasp, hold, shake, put down and pour.

to obtain the annotated ground truth for the beginning and the end of actions within the sequences, the data has been hand-labelled based on the video and audio recorded by an additional camera[3]. The corresponding labels have been set to match exactly the action primitives described in Sec. : *grasp+lift2, shake, hold2, putdown2, unscrew, grasp+lift1, pour, hold1, putdown1, screw*. In the following section we analyze the results of applying the two-stage segmentation to the constrained and unconstrained scenario.

Segmentation Quality

We assess the obtained segmentations w.r.t. the timing accuracy of the generated segmentation in both, the constrained and the unconstrained scenario. In order to assess the average error μ, the temporally closest generated change point is searched within a temporal window around the ground truth change point. The average distance between the ground truth and the generated change point measures the accuracy of the segmentation. The value μ is calculated by averaging over the trials.

The Fig. 3 (left) shows an overview of subject-specific timing deviations for all three human demonstrators `hd1`, `hd2`, `hd3` in the constrained scenario. As can be seen from the figure the average action-specific temporal error lies in the range from 0.05 seconds to 0.3 seconds for all subjects. In most cases it lies below 0.2 seconds. Furthermore, the action-specific error between different subjects varies in the range of 0.1 seconds. The variance of the error is negligible.

The Fig. 3 (right) compares the constrained and unconstrained execution scenarios. The red bars illustrate the temporal error averaged over all subjects for the constrained scenario (see left plot). The green bars present the temporal error averaged over all subjects in the unconstrained scenario. The plot does not show strong difference between the action-specific errors for both scenarios. The largest differences occur for “pour”, “hold” and “putdown1”.

Tests conducted with different test objects, fluids, different sequences of object manipulations have showed comparable results. The assignment of product models to the bimanual tactile data in the first step has yielded 100 percent correct results. A more detailed evaluation of the segmentation procedure can be found in (Barchunova, Haschke, Grossekathoefer, et al., 2011).

Conclusions and Outlook

In this paper, we presented a novel method for unsupervised bimodal segmentation and modeling of object manipulation operations in the context of a bimanual interaction scenario. We carried out experiments with human subjects and applied the proposed method to the collected data in two different scenarios: constrained and unconstrained. The experimental evaluation has showed satisfactory results in both scenarios. In particular, the results showed invariance of the segmentation quality w.r.t. different human demonstrators and speed of execution.

The robustness and generalization ability make the method suitable for use as a building block in higher-level modeling procedures. Due to the simplicity of the two fundamental models and the modeling concepts used within our approach, the developed procedure can be easily used in a wide range of scenarios. Furthermore, the hierarchical approach to segmentation makes traditionally applied fusion unnecessary.

In our future work we seek to apply our method online within a higher-level human-machine interaction scenario.

Acknowledgments

Alexandra Barchunova gratefully acknowledges the financial support from Honda Research Institute Europe.

References

Aggarwal, J. K., & Park, S. (2004). 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: IEEE Computer Society.

[3] Here we have used a QuickCam Pro 9000 Analog Mono

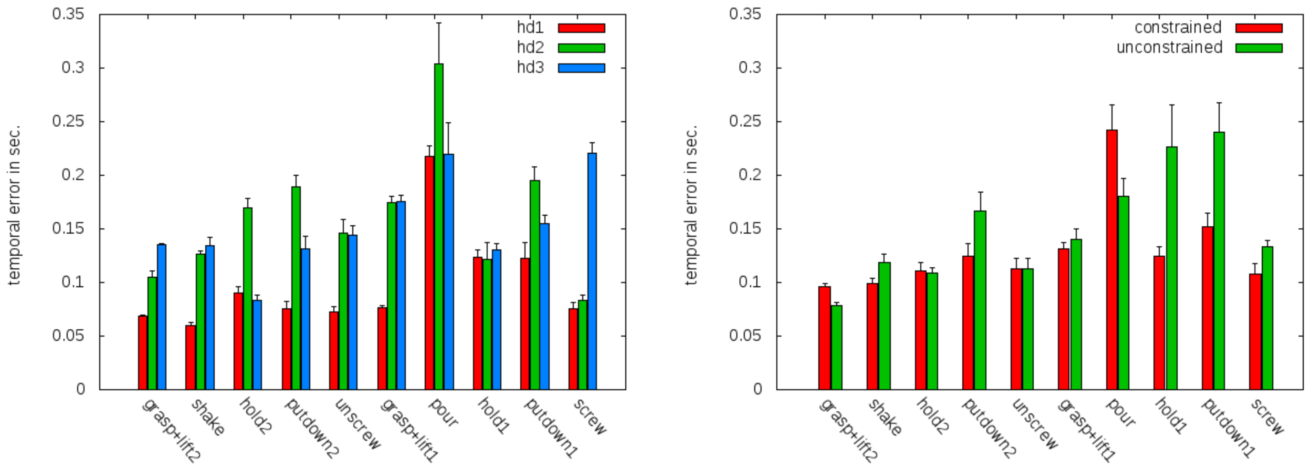

Figure 3: Action-specific timing error μ and the corresponding variance σ plotted for three human demonstrators (left subfigure); comparison between temporal error for constrained vs. unconstrained trials (right subfigure) Prior segment length parameter $\lambda = 10^{-5}$. Index 1 or 2 distinguishes between uni- and bi-manual versions action primitives respectively.

Barchunova, A., Haschke, R., Franzius, M., & Ritter, H. (2011). Multimodal segmentation of object manipulation sequences with product models. In *Icmi.*

Barchunova, A., Haschke, R., Grossekathoefer, U., Wachsmuth, S., Janssen, H., & Ritter, H. (2011). Unsupervised segmentation of object manipulation operations from multimodal input. In B. Hammer & T. Villmann (Eds.), *New challenges in neural computation.*

Bernardin, K., Ogawara, K., Ikeuchi, K., & Dillmann, R. (2005). A sensor fusion approach for recognizing continuous human grasping sequences using hidden markov models. *Robotics, IEEE Transactions on*.

Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*.

Dillmann, R., Rogalla, O., Ehrenmann, M., Zöllner, R., & Bordegoni, M. (2000). Learning robot behaviour and skills based on human demonstration and advice: the machine learning paradigm. In *Robotics research.*

Fearnhead, P. (2005). Exact bayesian curve fitting and signal segmentation. *Signal Processing*, *53*.

Grossekathöfer, U., Barchunova, A., Haschke, R., Hermann, T., Franzius, M., & Ritter, H. (2011). Learning of object manipulation operations from continuous multimodal input. In *Humanoids.*

Kawasaki, H., Nakayama, K., & Parker, G. (2000). Teaching for multi-fingered robots based on motion intention in virtual reality. In *Iecon.*

Krüger, V., Kragic, D., Ude, A., & Geib, C. (2007). The Meaning of Action: A Review on action recognition and mapping. *Advanced Robotics*, *21*(13), 1473–1501.

Li, C., Kulkarni, P., & Prabhakaran, B. (2006). Motion Stream Segmentation and Recognition by Classification. In *Icassp.*

Matsuo, K., Murakami, K., Hasegawa, T., Tahara, K., & Ryo, K. (2009). Segmentation method of human manipulation task based on measurement of force imposed by a human hand on a grasped object. In *Iros.* IEEE.

Ogris, G., Stiefmeier, T., Lukowicz, P., & Troster, G. (2008). Using a complex multi-modal on-body sensor system for activity spotting. *Wearable Computers, IEEE International Symposium*.

Pardowitz, M., Knoop, S., Dillmann, R., & Zöllner, R. (2007). Incremental learning of tasks from user demonstrations, past experiences, and vocal comments. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, *37*(2), 322–332.

Sanmohan, B., Krüger, V., & Kragic, D. (2010). Unsupervised learning of action primitives. In *Humanoid robots.*

Takano, W., & Nakamura, Y. (2006). Humanoid robot's autonomous acquisition of proto-symbols through motion segmentation. In *Humanoid robots.*

Turaga, P., Chellappa, R., Subrahmanian, V., & Udrea, O. (2008). Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, *18*(11), 1473–1488.

Ward, J. A., Lukowicz, P., Troster, G., & Starner, T. E. (2006). Activity recognition of assembly tasks using body-worn microphones and accelerometers. *TPAMI*.

Zöllner, R., Asfour, T., & Dillmann, R. (2004). Programming by demonstration: Dual-arm manipulation tasks for humanoid robots. In *Iros.*

Zöllner, R., & Dillmann, R. (2004). Using multiple probabilistic hypothesis for programming one and two hand manipulation by demonstration. In *Iros.*