# Unsupervised Segmentation of Object Manipulation Operations from Multimodal Input

Alexandra Barchunova Jan Moringen Ulf Grossekathoefer Robert Haschke
Sven Wachsmuth Herbert Janssen Helge Ritter

**Abstract**

We propose a procedure for unsupervised identification of bimanual high-level object manipulation operations in multimodal data. The presented procedure applies a two-stage segmentation and a selection step to observation sequences. We employ an unsupervised Bayesian segmentation method to identify homogeneous segments which correspond to primitive object manipulation operations. The data is recorded using a contact microphone, a pair of Immersion CyberGloves and ten pressure sensors positioned on the fingertips.

The assessment of the temporal correctness and structural accuracy of the segmentation procedure has showed satisfactory results. We have achieved an average error of 0.25 seconds in comparison to the actual segment borders. The examination of the structural accuracy for a given parameter combination has showed only insignificant deviation of the generated segmentation structure from the corresponding test data.

Finally, we sketch an application of our method to unsupervised learning and representation of object manipulations.

## 1 Introduction

An important objective of today's robotics research is to enable robots to interact with humans in everyday scenarios. Within this area, we focus on the topic of autonomous learning and identification of bimanual object manipulations from 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. It also needs a mechanism to identify and organize these parts. In order to address this requirement, we propose a novel approach for unsupervised identification of high-level bimanual object manipulation operations within action sequences. Inspired by the fact, that humans employ different information sources – like hearing, proprioception, haptics and vision – to fulfill this task, we propose a multi-modal approach to segment and identify action sequences. To this end we consider an audio signal, tactile sensor readings from all finger tips, and hand postures acquired by CyberGloves [1].

Analysis of various sensor readings describing the human hand dynamics during manual interaction have been conducted recently by different researchers [2, 3, 4]. In general, one is interested in autonomous identification of action primitives in the context of imitation learning and human-machine interaction [5, 6]. Within this domain, Matsuo et al. focused on force feedback [7] while a combination of different sensors like CyberGlove, Vicon or magnetic markers and tactile sensors has been used by [8], [4] and [9]. In [10] a bimanual approach is described.

Despite the variety of sensors and approaches used in action segmentation and identification, one modality, namely the audio signal, has been mostly ignored in this domain. However, in the area of speech recognition it is well known, that the audio signal not only transmits the mere verbal content, but also conveys temporal structure of interactions and actions [11].

Our past work has been concerned with unsupervised segmentation and classification of raw motion data and its linear projection into a low-dimensional space [12]. The experiments within this preliminary study have showed that the absence of structural analysis of object manipulation sequences restricts the scenario to a small set of distinct and unambiguous manipulations. To tackle more complex and ambiguous action sequences, we employ a Bayesian segmentation method to analyze the sequential structure.

In our scenario, during a considerable number of simple high-level object manipulations (e.g. grasping, shifting, shaking, stirring or rolling) application of force is naturally accompanied by a sound. We exploit this fact by performing segmentations based on the analysis of the audio signal structure and of contact forces recorded on the fingertips. The resulting segmentation solely depends on the temporal structure of the data and is invariant to absolute data values, way of grasping or the manipulation object. Our method does not employ any specific knowledge about the parts of the action sequence. Furthermore, it does not require a large set of domain-specific heuristics describing each action primitive as is commonly the case in similar approaches [8, 4, 13].

We evaluate our method in an everyday scenario in which a human subject performs several object manipulation operations with a large non-rigid plastic bottle with a handle. In this evaluation, we assess the performance of the

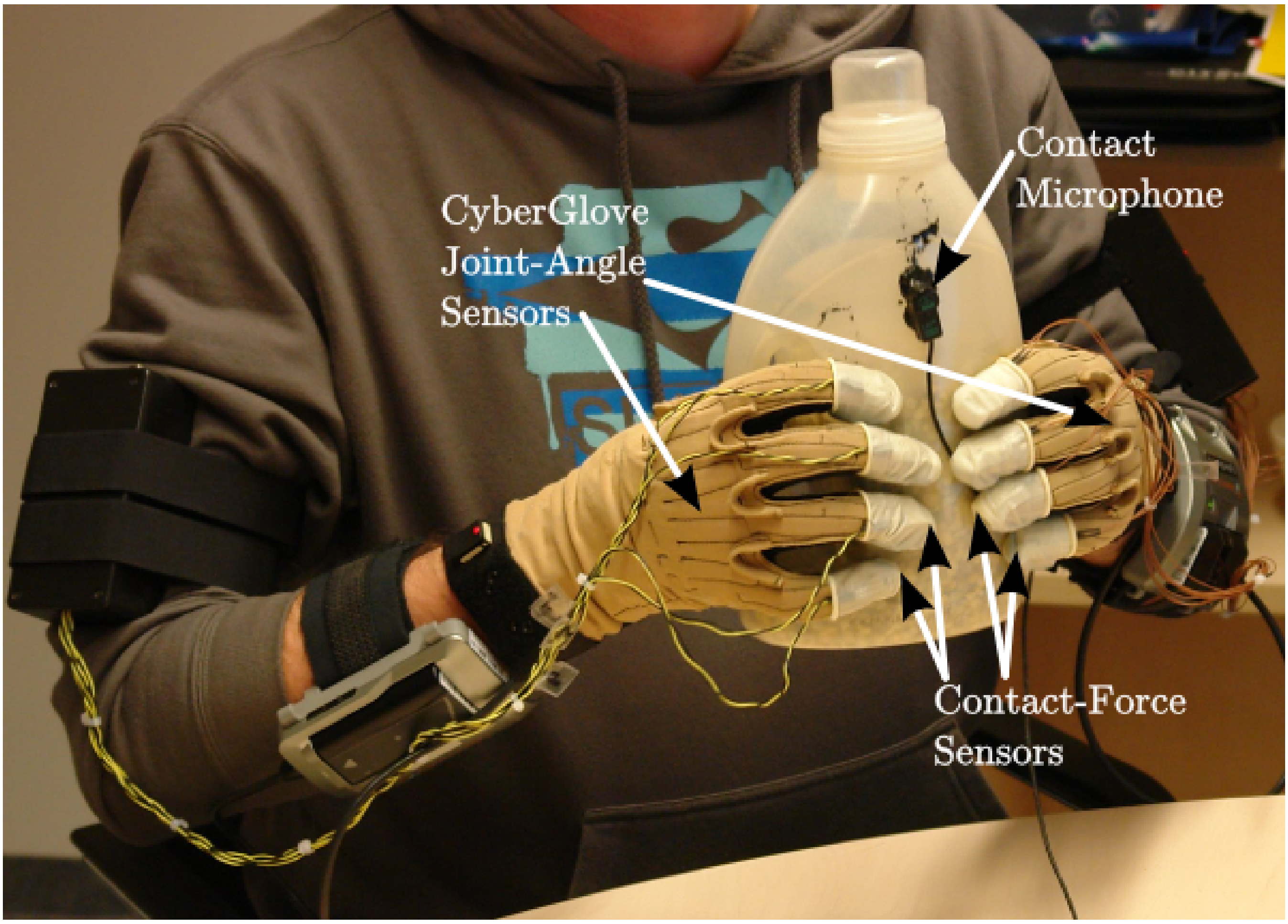

Figure 1: Experimental setup: a subject wearing contact and joint angle sensors performs manipulation operations with a (instrumented) plastic bottle provided with a contact microphone.

segmentation method w.r.t. the accuracy of the generated segment borders and the overall structure of the produced segmentation. Additionally we briefly outline the results of applying an unsupervised learning procedure, which has been used in similar tasks ([14, 15]), to cluster the identified action segments. The developed method is applicable to interactive scenarios such as imitation learning, cooperation and assistance.

The rest of this paper is organized as follows: Sec. 2 explains the acquisition of action sequences within the scenario. Sec. 3 introduces the two steps of the proposed method: preprocessing (Sec. 3.1) and segmentation (Sec. 3.2). In Sec. 4, we discuss our evaluation method and experimental results of the segmentation procedure, and report on an application of the proposed method as a preprocessing stage of an action recognition module (Sec. 5). Sec. 6 concludes the paper with a brief discussion and outlook.

## 2 Scenario and Experimental Setup

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

We use the following sensors to record multimodal time series of the performed action sequences (corresponding modality names used in formulas appear in parenthesis):

- one contact microphone attached to the bottle (a). The contact microphone focuses on in-object generated sound, ignoring most environmental noise.
- 2 × 24 joint-angles calculated from the measurements of two Immersion CyberGlove devices (j: both hands, jl: left hand, jr: right hand). The Immersion CyberGlove II devices output sensors values describing the configurations of finger- and palm-joints.

[1] The use of gravel instead of liquid is due the necessity of a distinct audio signal and also safety concerns.

- $2 \times 5$ FSR pressure sensors attached to the fingertips of each CyberGlove (`t`: both hands, `tl`: left hand, `tr`: right hand) record the contact forces.

This collection of sensors yields a 29-dimensional (24 + 5) representation for each hand in addition to a scalar audio signal. The subject was told to perform a sequence of basic manipulation actions in fixed order as listed in the following enumeration. To obtain ground truth for later evaluation, the beginning or end of an action within a sequence was signalled to the subject as explained in Sec. 4. To achieve a rich variance of timing between trials, the desired duration of most elements was sampled from a Gaussian distribution with standard deviation of 0.5s as specified in parentheses:

1. pick up the bottle with both hands (2 s + $\eta_1$ )
2. shake the bottle with both hands (.7 s + $\eta_2$ )
3. put down the bottle (1 s)
4. pause (1 s)
5. unscrew the cap with both hands (1.2 s + $\eta_3$ )
6. pause (1 s)
7. pick up the bottle with right hand (2 s + $\eta_4$ )
8. pour with right hand (1 s + $\eta_5$ + 1 s + $\eta_5$ )
9. put down the bottle (1 s)
10. fasten the cap with both hands (1.2 s + $\eta_6$ )

The random variables $\eta_i \sim \mathcal{N}(0, .5\ s)$ denote the randomized timing of subsequences. The overall length of the time series accumulates to approximately 30 seconds.

# 3 Method

The recorded time series of multiple sensor modalities capture complex and high-dimensional descriptions of action sequences. The focus of this paper is on segmentation and selection of relevant data. Furthermore, we briefly outline a subsequent clustering step to demonstrate that the proposed method can serve as a preprocessing stage for an unsupervised learning procedure to recognize action primitives. In the following paragraphs we describe the segmentation process based on the tactile and audio modalities.

## 3.1 Preprocessing

In a preprocessing step, the original audio-signal is normalized to a given variance range with respect to the amplitudes of individual samples. The signal is also subsampled and recording artifacts are removed by discarding samples whose amplitude exceeds a specified threshold. We use the resulting processed audio signal in the segmentation step described in Sec. 3.2. This preprocessing is necessary for successful segmentation due to the characteristics of Auto-Regressive models used in the segmentation process.

## 3.2 Segmentation

In our two-stage segmentation approach, we use tactile information to obtain a preliminary rough split of the sequence into subsequences of “object interaction” and “no object interaction”. This analysis of hand-object contacts uses force data from both hands. 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 both stages, the segmentation is performed by applying Fearnhead’s method [16] for unsupervised detection of multiple change-points in time series. The input to Fearnhead’s algorithm is a time series $y_{1:T}$ [2] and a set of models $\mathcal{M}$ for homogeneous subsequences. The output is a set of integer change-points $1 < \tau_1 < \cdots < \tau_N < T$ at which qualitative changes in the data $y_{1:T}$ occur. A set of such change-points is dual to segmentation of the form $(y_{s_i:t_i})_{1 \leq i \leq N+1}, s_1 = 1, t_i = \tau_i = s_{i+1}, t_{N+1} = T$ which partitions the data into $N + 1$ subsequences. Within the probabilistic framework of Fearnhead’s algorithm, the optimal segmentation is obtained by maximizing the Bayesian

[2] We use the notation $x_{a:b} \equiv (x_a, \ldots, x_b)$. We use $x_{\_|\texttt{mod}}$ to indicate the restriction to modality `mod`.

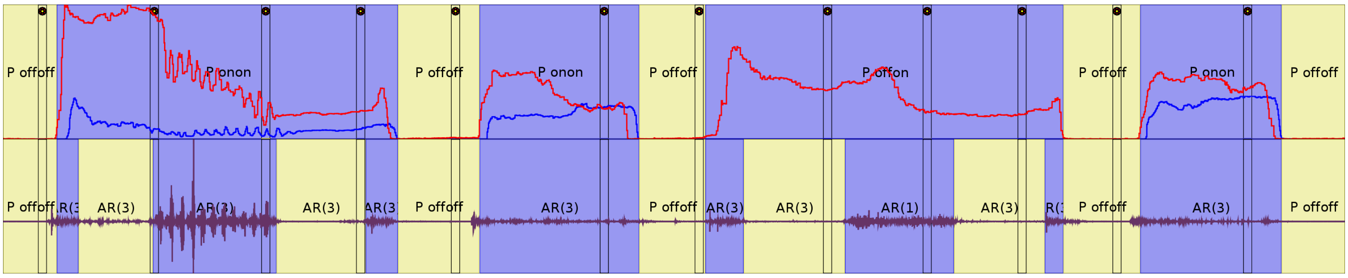

Figure 2: Initial segmentation and "subordinate" sub-segmentation for one multimodal time series. The first row shows the result of applying Fearnhead's method with joint threshold models of the tactile data of both hands (see Sec. 3.2.1 for details). The segmentation is overlaid with the tactile signals of both hands. The second row shows the refinement of the segmentation in the first row that is computed by applying Fearnhead's method to the audio signal within each "contact" segment (see Sec. 3.2.2 for details). In the second row the segmentation is overlaid with the audio signal.

posterior[3] $P(y_{1:T} \mid \tau_{1:N})P(\tau_{1:N})$ 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 $\lambda$ 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 $\lambda$ has been chosen, neither the number of change-points $N$ nor any information regarding their positions have to be specified in advance. Due to the difference in the input content of the time series, both segmentation steps of our procedure specify their own method for $\lambda$ calculation. We use the notation $\lambda^{\alpha}$ in the first stage and $\lambda_{\text{sub}}$ in the second, subordinate segmentation stage. These values will be discussed in the respective subsections.

In addition to the prior distribution of segment lengths, the algorithm employs a finite set of models $\mathcal{M}$ to represent different regimes in segments of the time series. Each model $m \in \mathcal{M}$ assigns marginal likelihoods $P(y_{s:t} \mid m)$ to segments $y_{s:t}$, $1 \le s < t \le T$, of the time series. Prior probabilities $P(m)$ are associated with all models. In this paper, we only consider sets of up to four models with uniform prior distributions.

In the following two subsections, we describe the application of Fearnhead's algorithm to two different subsets of the available modalities in combination with two suitable sets of models $\mathcal{M}$ and $\mathcal{M}_{\text{sub}}$. The two-stage application of the segmentation procedure and the modality-specific local models constitute the main contributions of this paper.

### 3.2.1 Segmentation based on Tactile Modalities

The first step performs a rough joint analysis of the tactile signals of both hands. For the application of Fearnhead's method in this stage, we set the value of the prior parameter $\lambda^{\alpha} = 1/T^{\alpha}$ for each trial $\alpha$ of length $T^{\alpha}$. 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. The analysis uses four pairs of threshold models. Each model of a pair describes the tactile state, i.e. "object contact" vs. "no object contact", for one hand. We denote the "object contact"-models with capital-letter subscripts: $m_L$ and $m_R$ for the left and the right hand respectively. The corresponding notation for the "no object contact"-models is $m_l$ and $m_r$.

The marginal likelihood, that a model fits to a time series segment $y_{s:t}$ is of the form:

$$P(y_{s:t} \mid m_l) = {p_o}^{n} \quad \text{and} \quad P(y_{s:t} \mid m_L) = {p_o}^{u-n}$$

where $p_o$ is the fixed probability that a sample does not fit the model (in this case $m_l$), $u = t - s$ is the segment length, and $n$ is the number of such samples within the time series segment, e.g. $n = \left|\{y_{k|\mathtt{t1}} > \gamma \mid s \le k < t\}\right|$. The parameter $\gamma$ specifies the threshold for recognizing contact.

Combining these individual models, $\mathcal{M}$ consists of the following four joint models: "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 likelihoods of these joint models are computed as products of the individual likelihoods, e.g.:

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

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. Accidental movements of one or both hands between manipulations are separated from manipulation

[3] We suppress $P(y_{1:T})^{-1}$, which is irrelevant for the maximization.

operations in this step. Such movements occur for instance during an approach phase prior to grasping. The assignment of models to segments can be exploited to exclude joint and tactile modalities (`jl`, `tl` for left hand; `jr`, `tr` for right hand) of "inactive" hands from subsequent processing steps (e.g. clustering, see Sec. 5). For example, the assignment of $m_{l,R}$ to a segment $y_{s:t}$ leads to the corresponding data fragment $y_{s:t|\mathtt{jl},\mathtt{tl}}$ being excluded. When the model $m_{l,r}$ is assigned, the segment in question can be ignored entirely.

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. On the downside, Fearnhead's method requires the specification of a prior distribution on segment lengths, i.e. the $\lambda$ parameter.

### 3.2.2 Sub-segmentation of Object Contact Segments Based on Audio Signal

In this subordinate segmentation step, all segments produced and not discarded in the previous step are sub-segmented using Fearnhead's method. This time, the audio signal in the constructed sub-segments is assumed to be produced by Auto-Regressive (AR) models of order 1, 2 or 3: $\mathcal{M}_{\text{sub}} = \{AR(1), AR(2), AR(3)\}$ [16]. Thus the sub-segmentation is formed by selecting segments that exhibit homogeneous oscillatory properties within the audio modality. In contrast to the procedure outlined in the previous paragraph, the value of the segment length distribution parameter $\lambda_{\text{sub}}$ is fixed. In our evaluation (Sec. 4), a suitable value for $\lambda_{\text{sub}}$ is estimated by means of a grid-search.

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. The assignment of "object contact" threshold models from the first segmentation step is discarded in this final segmentation result since it is not exploited in further steps.

# 4 Experimental Results

## 4.1 Data Pool

We recorded 50 trials of the action sequence described in Sec. 2 with a single subject in one session. In principle, the structure of all these trials should be identical except the timing. However, it turned out to be rather difficult for the subject to perform such a high number of trials without structural variations. As a result, some trials exhibit structural differences like missing or additional tactile contacts or repeated actions. However, we made no attempt to correct these irregularities.

In the domain of unsupervised recognition of human actions, there is no established methodology for quantitative evaluation. To avoid time-consuming hand-labelling of our data, we generate and use randomized action time schedules for all trials in the following way: for a particular trial, audio cues are emitted according to the corresponding schedule to mark the start or end times of actions. We rely on the subject to react to these cues and align their executed actions as closely to them as possible. The audio cues (similar to dial tones) are provided via head phones to prevent their presence in the recorded audio modality.

Each cue consists of a sequence of four beep sounds[4] : the first three are preparatory and allow the subject to anticipate the fourth signal which notifies the associated event (beginning or end of action execution) to the subject. The timing of cues is derived from the structure described in Sec. 2 by randomizing the duration of individual actions. We record timestamps of generated cues, as an indication of the timing of scheduled actions. In our evaluation, we use these recorded cue timestamps as ground-truth. This enables us to assess the correctness of timing and the number of generated segments. Note that this ground-truth is an approximation due to differences between cues and the actual timing of action execution. 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 $\alpha$ [5].

## 4.2 Segmentation Quality

In this section, we analyze the results of applying the two-stage segmentation described in Sec. 3.2 to the data discussed above. We assess the obtained segmentations w.r.t. the following three aspects: the number of calculated segments, the number of undetected segment borders and the timing accuracy of the generated segmentation. We perform this assessment of our procedure for a large set of combinations of the adjustable parameters. These are: the contact threshold value $\gamma$, the segment length distribution parameter $\lambda_{\text{sub}}$ and the range parameter for the normalization of the audio signal $\rho$. In our experiments, we have used all possible combinations of $\lambda_{\text{sub}} \in 10^{\{-4,-5,-6,-7,-8\}}$, $\rho \in \{6, 8, 10, 12\}$ and $\gamma \in \{15, 20, 30, 40, 50, 60, 70, 80\}$. The goal of these experiments is to assess the respective

[4] The preparatory cue signals are .1 s long, the pause between signals is .2 s long and the main signal lasts .2 s.

[5] When the trial is clear from context or not important, we drop the superscript and write cue times as just $c_{i,j}$.

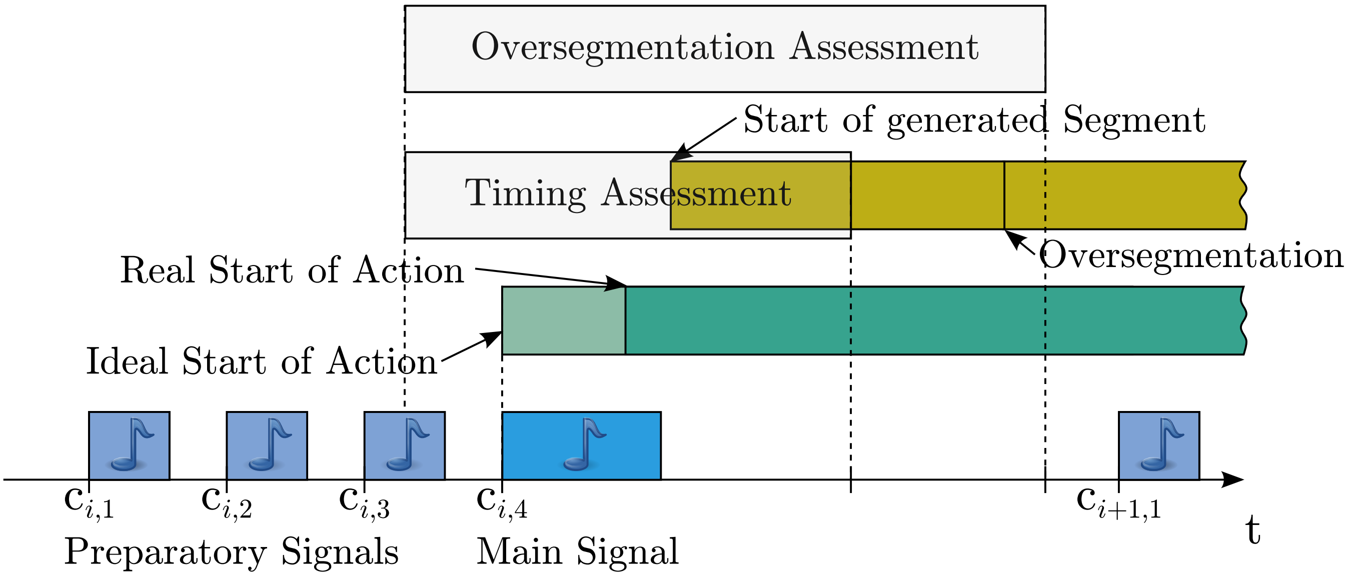

Figure 3: Temporal relations between cues, actions and generated segments. The execution of an action by the subject is expected to start ( light green bar ) at the beginning of the cue signal $c_{i,4}$, but the actual beginning of the execution usually deviates ( dark green bar ). In our evaluation, we try to find automatically generated segments ( dark yellow bar ) that correspond to these actions in different areas ( light gray boxes ) around the cues (See Sec 4.2 for details).

influences of the parameters and to find a parameter combination that yields segmentations most close to the ground-truth in all three abovementioned aspects.

To obtain quantitative results, the cue-based ground truth data is exploited as follows: First, for each main cue signal $c^{\alpha}_{i,4}$ within each trial $\alpha$, the temporally closest generated change-point is searched within a temporal window around the start time of the cue signal. Depending on whether timing or oversegmentation is assessed, we use a smaller or larger window (See Fig. 3 for an illustration of the procedure). If such a change-point can be identified, the temporal distance to the cue signal serves as an indicator of the accuracy of the segmentation. Otherwise the manipulation performed in response to the cue signal is considered as not having been detected.

Fig. 4 (left) shows the timing deviation of the estimated segments from the ground-truth. In order to determine the timing error for a given parameter combination, we first average over all trials resulting in twelve values, one for each cue. We calculate the displayed values of mean and variance by additionally aggregating all twelve cues. The window size around each cue used in the evaluation was set to $[c_{i,3}, (c_{i,4}+c_{i+1,1})/2]$ (see Fig. 3). The resulting average time-intervals between the cues and the closest estimated segment border are sorted in the order of ascending error. From Fig. 4 (left) it can be easily seen, that segment borders generated by the proposed method are extremely robust w.r.t. all parameters. The average error lies in the range 0.25 to 0.29 seconds. We observe lower error values in conjunction with higher values of $\lambda_{\text{sub}}$ and lower values of $\gamma$. The higher error values co-occur with smaller values of $\lambda_{\text{sub}}$ and larger values of the $\gamma$ parameter. We note that the remaining minimal error of approx. 0.25 seconds might originate from the subject's need to adapt the hands before executing the scheduled movement. The parameters that yielded the best results in this experiment were $\lambda_{\text{sub}} = 10^{-5}$ and $\gamma = 15$.

The goal of the following experiment is to evaluate the number of segments generated for each cue. Fig. 4 (right) illustrates the dependency of the average number of estimated segment borders on the parameters. Average and variance values are calculated analogously to Fig. 4 (left). However, the environment used to estimate the number of candidates for one cue is set to be $[c_{i,3}, (c_{i,4} + c_{i+1,3})/2]$. Thus the whole sequence is covered by the calculation (see Fig. 3). This experiment shows a strong dependency between the amount of oversegmentation and the parameter $\lambda_{\text{sub}}$. Smaller values of $\lambda_{\text{sub}}$, yield fewer candidates within a cue environment. We observed the best results for $\lambda_{\text{sub}} = 10^{-8}$. This parameter has a clear and a considerable influence on the structure of the resulting segmentation. Despite the increase in deviation of timing from the ground-truth of about 0.01 seconds, we choose the parameter set $\lambda_{\text{sub}} = 10^{-8}, \rho = 12, \gamma = 40$ for further calculations.

Fig. 5 (left) shows the cue-specific average number of generated segments for the abovementioned parameter set. In this figure one can clearly differentiate between two groups of events: double events and single events. The first group contains for example *pick up and lift* and *put down*. The start of these actions is marked by a cue, but the duration is so short that no end-cue can be issued correctly to signal the end of the action to the subject. Thus the

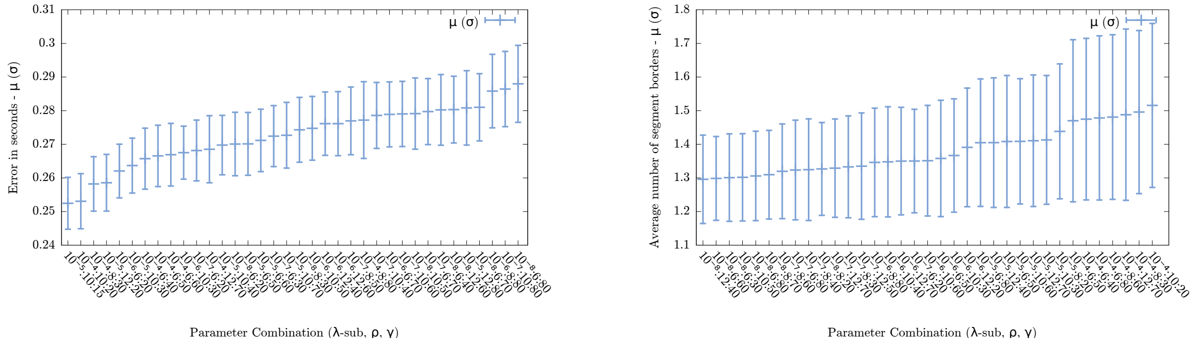

Figure 4: Left: Average distances between cues and corresponding estimated segment borders. Distances (*y-axis*) are in seconds and sorted by increasing average error for different combinations of the parameters $\lambda_{\text{sub}}$, $\rho$ and $\gamma$ (*x-axis*). Right: Sorted average number of estimated segment borders for actions (*y-axis*) w.r.t. parameter combinations (*x-axis*). In both figures, averages are over all trials and all actions for each parameter combination.

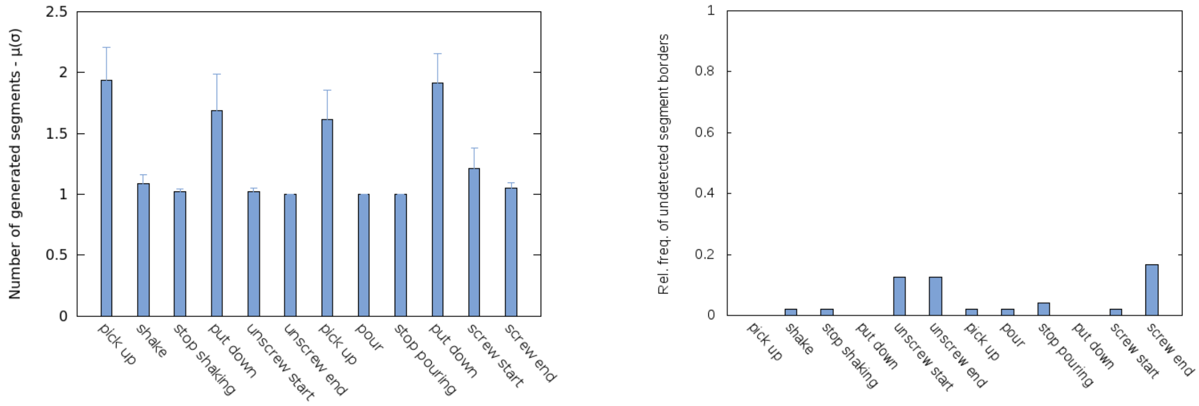

Figure 5: Left: Average number of segment borders (*y-axis*) for each action (*x-axis*) for the parameter combination $\lambda_{\text{sub}} = 10^{-8}, \rho = 12, \gamma = 40$. Right: Relative frequencies of undetected segment borders (*y-axis*) for each action (*x-axis*) for the aforementioned combination of parameters. Note that some actions consist of multiple sub-actions for which no ground-truth information is available (See Sec. 4.2 for details).

average number of generated change-points close to two is almost optimal. The second group contains single events like *start shaking*, *end shaking*, *start unscrewing* or *end unscrewing*. In this group, the beginning or the end of the action is marked by the cue. Thus the average number of generated change-points, approximately one, is close to optimal as well. The average and the variance values are computed over all trials. Fig. 5 (right) shows the cue-specific average relative frequency of undetected segments. The high likelihood of detection failures for the *screwing* event is possibly due to the incorrect execution timing of the subject.

## 5 Application Example for Unsupervised Learning with OMMs

We consider the segmentation method we presented and evaluated in the previous sections as a building block for more sophisticated unsupervised methods. To support this claim, we briefly outline an unsupervised procedure for identification and representation of action primitives based on the proposed method.

To perform identification and representation of action primitives, segments which contain semantically similar actions have to be grouped and models of these groups have to be formed. We address both tasks by embedding the concept of Hidden Markov Models (HMM), which yields good results in representation and modeling of sequential data, in a clustering approach. In the procedure sketched here, we use Ordered Means Models [17], an efficient variant of HMMs with flexible left-to-right topology and Gaussian emission densities.

From the perspective of unsupervised clustering and representation, the output of the proposed segmentation method is a set of multimodal data sequences $\{y^\beta\}_{1\leq\beta\leq B}$ that are unlabeled w.r.t. the trials and actions from which they originate. The application of OMMs to partition such a dataset into $k$ groups in an unsupervised manner, can be considered a special case of the well-known $k$-means clustering. OMMs $\lambda_1, \ldots, \lambda_k$ are used as the associated prototypes

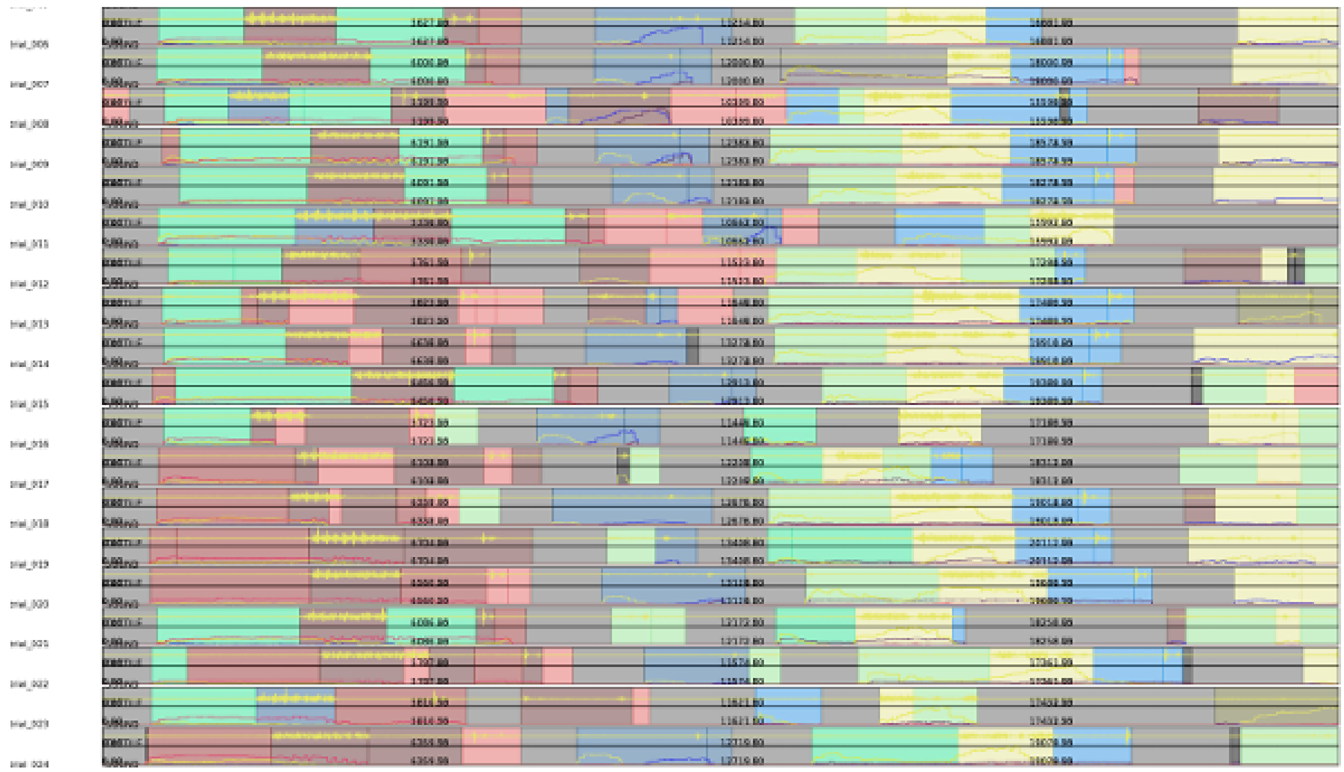

Figure 6: Assignment of labels (*designated by random colors*) to segments according to the best matching model in a small subset of trials. In each row, the segmentation, label assignments, audio signal (*top half*) and tactile information (*bottom half*) is shown. Corresponding segments in adjacent trials do not line up because of the randomized timing.

of $k$ clusters. A suitable distance function then is the negative log-likelihood that a sequence $y^\beta$ is generated by an OMM $\lambda_j$: $d(y^\beta, \lambda_j) = -\log P(y^\beta \mid \lambda_j)$. Given this, a $k$-OMMs clustering algorithm partitions data sequences into $k$ groups by minimizing the objective function

$$E = -\sum_{\beta=1}^{B} \sum_{j=1}^{k} w_{\beta,j} \log P(y^\beta \mid \lambda_j).$$

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

Prior to performing $k$-OMMs 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.

Fig. 6 qualitatively shows the result of applying the sketched clustering and learning procedure in the following way: in a training step, twelve OMMs are formed based on segmentations obtained with the presented segmentation method. Then, in a test step, segmented action sequences are classified to the best-matching OMM model. Identically colored segments are considered semantically equivalent.

## 6 Conclusions and Outlook

In this paper, we presented a novel method for unsupervised identification of object manipulation operations in the context of a bimanual interaction scenario. We carried out experiments with a human subject and applied the proposed method to the collected data. The experimental evaluation has showed satisfactory results for both: the segmentation timing and the structural accuracy. These results and an application in an OMM-clustering has showed that the method is able to select primitive object manipulation operations. Future research will be concerned with learning higher level representations of sequences of object manipulation operations. Within this context, the problem of semantically equivalent clusters will be addressed as well. It is also desirable to reduce the number of tunable parameters.

## 7 Acknowledgments

Alexandra Barchunova and Jan Moringen gratefully acknowledge the financial support from Honda Research Institute Europe.

## References

[1] Http://www.cyberglovesystems.com/products/cyberglove-ii/overview.

[2] K. Bernardin, K. Ogawara, K. Ikeuchi, and R. Dillmann, “A sensor fusion approach for recognizing continuous human grasping sequences using hidden markov models,” *IEEE Trans. on Robotics*, 2005.

[3] R. Dillmann, O. Rogalla, M. Ehrenmann, R. Zollner, and M. Bordegoni, “Learning robot behaviour and skills based on human demonstration and advice: the machine learning paradigm,” in *Proc. ISRR*, 2000.

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

[5] B. Sanmohan, V. Krüger, and D. Kragic, “Unsupervised learning of action primitives,” in *Proc. Humanoid Robots*, 2010.

[6] W. Takano and Y. Nakamura, “Humanoid robot’s autonomous acquisition of proto-symbols through motion segmentation,” in *Proc. Humanoid Robots*, 2006.

[7] K. Matsuo, K. Murakami, T. Hasegawa, K. Tahara, and K. Ryo, “Segmentation method of human manipulation task based on measurement of force imposed by a human hand on a grasped object,” in *Proc. IROS*, 2009.

[8] M. Pardowitz, S. Knoop, R. Dillmann, and R. Zollner, “Incremental learning of tasks from user demonstrations, past experiences, and vocal comments,” *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics,*, 2007.

[9] C. Li, P. Kulkarni, and B. Prabhakaran, “Motion Stream Segmentation and Recognition by Classification,” in *Proc. ICASSP*, 2006.

[10] R. Zollner, T. Asfour, and R. Dillmann, “Programming by demonstration: Dual-arm manipulation tasks for humanoid robots,” in *Proc. IROS*, 2005.

[11] L. Schillingmann, B. Wrede, and K. Rohlfing, “A computational model of acoustic packaging,” *Trans. on Autonomous Mental Development*, vol. 1, 2009.

[12] A. Barchunova, M. Franzius, M. Pardowitz, and H. Ritter, “Identification of high-level object manipulation operations from multimodal input,” in *Conf. ACIT*, 2010.

[13] R. Zollner and R. Dillmann, “Using multiple probabilistic hypothesis for programming one and two hand manipulation by demonstration,” in *Proc. IROS*, 2004.

[14] T. Grosshauser, U. Großekathöfer, and T. Hermann, “New sensors and pattern recognition techniques for string instruments,” in *NIME*, 2010.

[15] N.-C. Wöhler, U. Großekathöfer, A. Dierker, M. Hanheide, S. Kopp, and T. Hermann, “A calibration-free head gesture recognition system with online capability,” in *Pattern Recognition*, Istanbul, Turkey, 2010.

[16] P. Fearnhead, “Exact and efficient bayesian inference for multiple changepoint problems,” *Statistics and Computing*, 2006.

[17] U. Großekathofer, T. Lingner, H. Ritter, and P. Meinicke, “What is a hidden markov model without transition probabilities?” *submitted*, 2010.