# Towards Semantic Segmentation of Human Motion Sequences

Dirk Gehrig[1], Thorsten Stein[2], Andreas Fischer[2], Hermann Schwameder[2], and Tanja Schultz[1]

[1] Institute for Anthropomatics,
[2] Institute for Sport and Sport Science,
Karlsruhe Institute of Technology, Germany
dirk.gehrig@kit.edu

**Abstract.** In robotics research is an increasing need for knowledge about human motions. However humans tend to perceive motion in terms of discrete motion primitives. Most systems use data-driven motion segmentation to retrieve motion primitives. Besides that the actual intention and context of the motion is not taken into account. In our work we propose a procedure for segmenting motions according to their functional goals, which allows a structuring and modeling of functional motion primitives[3]. The manual procedure is the first step towards an automatic functional motion representation. This procedure is useful for applications such as imitation learning and human motion recognition. We applied the proposed procedure on several motion sequences and built a motion recognition system based on manually segmented motion capture data. We got a motion primitive error rate of 0.9 % for the marker-based recognition. Consequently the proposed procedure yields motion primitives that are suitable for human motion recognition.

## 1 Introduction

In the field of robotics exists an increasing need for knowledge about human motions, as a humanoid robot has to be empowered with knowledge about motion sequences [1]. Given the continuous nature of motion, there is an unlimited number of motion sequences that can be performed. Therefore, it is impossible to enumerate a complete set of motion primitives. Boundaries of motion primitives are often arbitrarily defined, making it difficult to automate the motion segmentation process [2]. However, humans tend to perceive motion in terms of discrete motion primitives [3, 4] and thus motion segmentation is still considered useful for some applications, including imitation learning [5] and human motion recognition [6].

Various different approaches can be found in the literature as to what should be seen as a motion primitive and to how these motion primitives can be modeled

[3] This work has been supported by the Deutsche Forschungsgemeinschaft (DFG) within Collaborative Research Center 588 “Humanoid Robots - Learning and Cooperating Multimodal Robots”

[5]. Various types of motion primitives are used ranging from low-level motions, e. g. moving the hand forward, up to complex motions such as setting the table. Currently, most approaches are data driven [8, 9] and exhibit a gap between kinematic and functional motion representations. In many scenarios it is not sufficient to know the kinematic or dynamic parameters of a motion, since the goal of the motion might not be reached although the execution of the motion is correct. For example, if a robot wants to grasp a glas, it has to make sure that the glas is properly grasped. It is not sufficient, if the robot only performs a motion trajectory based on kinematic and dynamic parameters, which does not result in grasping the glas.

In this paper we start to bridge this gap by looking at the problem from top-down. We think that a system for decomposing motions into motion primitives should take the goals of a motion into account. It is not suitable to decompose a motion in an arbitrary way, since the goals of the motions have to be fulfilled to perform the motion properly. To the best of our knowledge there is no system for decomposing arbitrary daily-life motions into motion primitives based on functional information. In our work we propose a system which allows us to retrieve a motion decomposition into motion primitives based on functional knowledge. Relatively few papers have so far dealt with higher abstraction levels of human motions which touch the border of semantics. Some papers try to segment the data based on object relationships [10, 11]. We do not use these appoaches since we also want to segment communicative gestures which are not object related. Another step in the direction of functional motion representation has been done by Guerra-Filho and Aloimonos [12]. They started to close the semantic gap between a WordNet and sensorimotor information by grounding a set of primitive words. Similar ideas have been presented by Ivanov et al. [13], whereas they assume a natural decomposition of motions into low-level primitives and higher-level semantic information. In our work we propose a systematic approach for a manual segmentation of motions into motion primitives based on the motion goals. The manual segmentation is a first step towards an automatic functional motion segmentation and will act as a baseline for the automatic segmentation.

## 2 A functional procedure to identify motion primitives

In this section a heuristic procedure is introduced that enables a decomposition of voluntary motions into motion primitives. Before analyzing the motion sequence in detail, the motion context should be considered (e. g. which objects are in the environment and where are the objects located). The motion context needs to be defined since this information is necessary to constitute the solution space. After the solution space is issued, an analysis of the motion sequence itself has to be carried out. The individual elements of the motion are primitives, which carry a specific function according to the overall goal of the motion sequence. These primitives therefore are called *functional primitives*. We distinguish between *main functional primitives* and *supporting functional primitives*. Main functional primitives appear at least once during the motion and deter-

mine the goal of the motion. In contrast, supporting functional primitives are not directly related to the goal of the motion and functionally dependent on other functional primitives. Besides this functional relation there is a temporal relation (Fig. 1). Thereby, *preparatory supporting functional primitives* improve the situation for subsequent functional primitives. In contrast, *assistant supporting function primitives* improve the execution of concurrent functional primitives. Finally, *transitional supporting functional primitives* transform the present motion situation into a new situation [14]. The two axes of the diagram in Fig. 1 represent the functional and temporal relationships of the motion primitive. For a more precise representation of the dependencies lines and arrows are used. An arrow hereby specifies a functional relationship between two motion primitives while a line specifies only a temporal relationship.

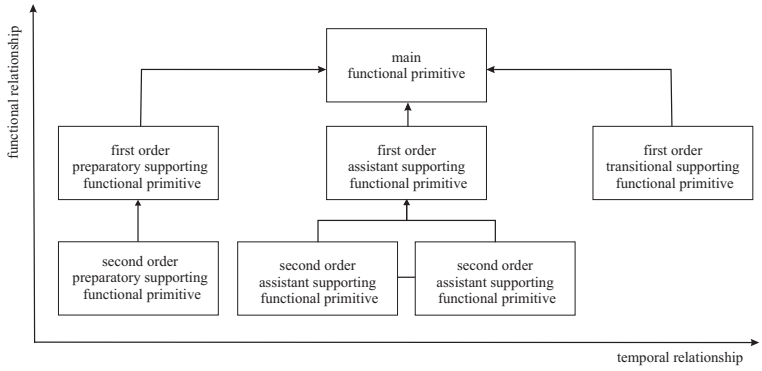


**Fig. 1.** Types and relationships of different functional primitives.

There are three possibilities for the decomposition of a motion sequence into functional primitives. In the case of the *inductive functional structuring* perceivable motions of the performer are the starting point for the decomposition. These motions are performed because they fulfill certain functions in the context of the motion goal and therefore these motions lead to functional primitives. The origin of the *deductive functional structuring* are not motions but the motion goal and the motion context. Thereby the motion goal has to be decomposed into sub-goals and according actions of the performer have to be defined. Due to the phenomenon of motor equivalence [15] in biological motor control different motions can be defined that fulfill the same goal. A third possibility is a *combined functional structuring* which corresponds to a synthesis of the inductive and deductive structuring [14]. For each of the identified functional primitives *temporal* and *positional* constraints have to be examined. This has to be done at the beginning, during and at the end of each functional primitive, e. g. where the hands of the person must be, at the beginning, during and at the end of each functional primitive. It also has to be specified, whether the motion primitive

has to be performed in a certain period of time. Finally, the segmentation of the motion into different functional primitives can be applied. The procedure does not guarantee that the decomposition always results in the same motion primitives and the same structure. The procedure is mainly a possibility to retrieve functionally plausible motion primitives, whereas the plausibility may depend on the desired application. The decomposition of a motion sequence results in different possible structures of motion primitives. Also the actual performed motion might be different for the same motion primitive. In other words the object positions and the used limb (e. g. left arm versus right arm) have to be taken into account for the performed motion. If for example an object is already at the desired position it can happen that no motion has to be performed for a motion primitive.

## 3 Application of the procedure

In this section we applied the above introduced procedure to a daily-life motion in a kitchen, cutting an apple (see Fig. 2). The motion sequence is part of the scenario of the Collaborative Research Center (CRC) 588 - Humanoid Robots. The goal of the CRC 588 is the construction of humanoid service robots that share their activity space with human partners. For the application of our procedure we assume that a person is facing a table. The result of the decomposition strongly depends on the environmental conditions, e. g. present object. The objects involved in the task "cutting an apple" are an apple, a knife and a cutting board. At the beginning and the end of the motion sequence the objects are placed on the table.

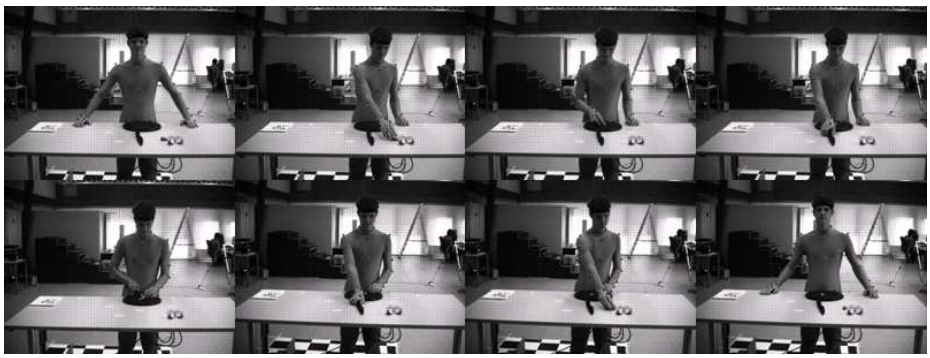

**Fig. 2.** Complex human motion sequence in a kitchen scenario: cutting an apple.

We used a *deductive functional structuring* to decompose the motion sequence into motion primitives. Fig. 3 shows two structures of the task "cutting an apple". For simplicity reasons we assume that only one motion primitive is performed at a time. Multiple motion primitives performed at the same time will be addressed

in future work. The dotted line represents an arbitrary number of repetitions of the motion primitive "cut apple". In our case the motion primitives have no temporal constraints besides the temporal structure shown in Fig. 3. At the beginning of the primitive we have no constraints induced by the motion context. The positional constraints during the motion primitives are: FP4: cut at apple, other hand at apple. Constraints at the end of motion primitives are: FP1: apple on top of cutting board, FP2: knife on top of apple, FP3: hand at apple, FP6: knife at original position, FP7: apple leftover at original position.

## 4 Evaluation of the procedure

For the evaluation we applied the procedure to five tasks in total: *cutting an apple, pouring water, grating an apple, stiring, mashing.* The decomposition resulted in 24 motion primitives including the ones described in Sec. 3. For training purposes of a recognizer, a subject performed each task 20 times in a single session. For data acquisition each motion primitive was always done with the same hand. We manually segmented the 100 motion sequences based on the retrieved motion structure and build a motion recognition system. We evaluated, what recognition performance can be reached when using the extracted motion primitives. We tested the recognition system without using a motion grammar to guide the recognition process, when using a statistical bigram model and when using a motion grammar deduced from the motion structure. For the segmentation and the grammar of the task cutting an apple we used the upper motion structure in Fig. 3.

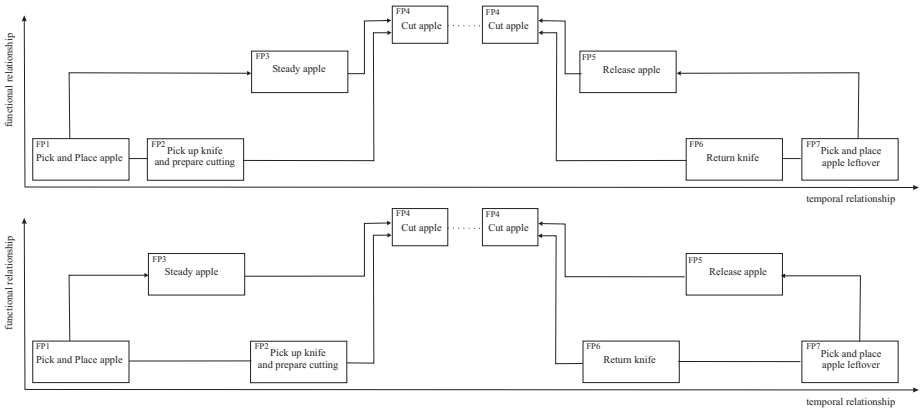


**Fig. 3.** Two functional and temporal structures of a complex human motion sequence in a kitchen scenario: cutting an apple.

Due to the specific task, we only collected data from the persons upper body. The motion sequences were simultaneously recorded with a Vicon motion capture

system and the camera system of a humanoid robot head (see Fig. 2). Only the marker-based data was used to build a recognition system and to evaluate our procedure. For the marker based motion capture 10 Vicon cameras were used to capture the motion of the subject with 20 fps. To capture the human motions 35 reflecting markers were attached to the subjects upper body, head and arms. The Vicon system output 3-dimensional positions and labels of the markers. Based on these marker information the related joint angle trajectories were calculated using a kinematic model [16]. As a result, the kinematic model outputs per time step one feature vector consisting of the 24 joint angles.

Our human motion recognition system features the one pass IBIS decoder [17], which is part of the Janus Recognition Toolkit JRTk [18]. We used this toolkit to recognize human motions based on joint angle velocities. The following paragraphs describe the components of our system, i. e. the input features, the model topology, the model initialization, training, and optimization, as well as the decoding strategy.

**Feature vectors:** The marker-based recognition system uses a 24-dimensional feature vector as input, consisting of 24 joint angle velocities from the upper body, which were calculated based on the joint angles resulting from the kinematic model. The input feature vectors are normalized by mean subtraction and normalizing the standard deviation to 1.

**HMM models:** Each motion primitive is statistically modeled with a left-to-right Hidden Markov Model (HMM). The number of states was optimized in cross-validation experiments as described below. Each state of the left-to-right HMM has two equally likely transitions, one to the current state, and one to the next state. The emission probabilities of the HMM states are modeled by Gaussian mixtures. The number of Gaussians per mixture was also optimized in the cross-validation experiments. A motion sequence was modeled as a sequential concatenation of these motion primitive models. In total, we discriminated 5 types of human motion sequences as mentioned above, consisting of the 24 different motion primitives.

**Model initialization:** To initialize the HMM models of the motion primitives, we manually segmented the data into the motion primitives. The manually segmented data were equally divided into one section per state, and a Neural Gas algorithm was applied to initialize the HMM-state emission probabilities.

**Model Training:** For HMM model training and development we used 10-fold cross-validation on the 100 motion sequences. For the experiments we varied the number of Gaussians between 1 and 64 and the number of states between 1 and 12. HMM training was performed featuring the Viterbi EM algorithm based on forced alignment on the unsegmented motion sequences.

**Decoding:** Decoding of the systems was carried out as a time-synchronous beam search. Large beams were applied to avoid pruning errors. We did three different types of decodings. First we did not use a motion grammar, whereat all transitions from one motion primitive to another are equally likely (1/24). In a second experiment we used a statistical bigram language model, where the probability of a motion primitive depends on the primitive before. As a third

experiment we used a motion grammar deduced from the motion structure. Recognition performance is reported in terms of motion primitive error rate.

**Results:** When using 9 states for each motion primitive and 4 Gaussians for each state, we got a motion primitive error rate of 4.2 % without using a motion grammar. When using a simple automatically generated statistical bigram model we got an error rate of 2.3 % whereas when using the deduced grammar, we only got an error rate of 0.9 %. Figure 4 shows that the increase in recognition rate can be found in every of the five kitchen tasks.

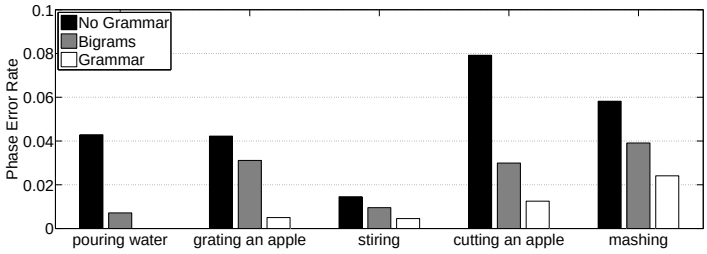


**Fig. 4.** Recognition error rates for 5 motion sequences.

## 5 Conclusion and Future Work

We showed a way how to approach the task of functional motion decomposition. The proposed procedure is not a step by step instructions manual for segmenting human motions into motion primitives since there is no general way how to do that. It is still necessary to approach the segmentation of human motions systematically to achieve motion primitives, which are essential in the sense that the goals of the motions are represented in the motion primitives. If these motion primitives are used for motion generation or motion recognition, by fulfilling the goals of each motion primitive the achievement of the overall goal of the motion sequence is guaranteed to be reached. We applied the procedure on our motion capture data and performed human motion recognition based on the motion primitives and their temporal and functional structure. We got a motion primitive error rate of 0.9 % for marker-based recognition when using the motion grammar deduced from the motion structure. This shows, that the proposed procedure yields promising motion primitives and grammars. Nevertheless, this approach should be combined with automatic human motion segmentation approaches to automatically learn new motion primitives. The first step in the direction of automization will be the automatic deduction of motion grammars based on the proposed motion structuring. Besides, the effort to build such a motion structure will be reduced step by step through automization of the retrieval of the motion structure. In addition to the automization the possibility of multiple motion primitives performed at the same time will be addressed.

## References

1. Schaal, S.: The New Robotics – towards human-centered machines. HFSP J. Volume 1, Issue 2, pp. 115-126 (July 2007)
2. Kahol, K.: Gesture Segmentation in Complex Motion Sequences. Proceedings IEEE International Conference on Image Processing, 105-108 (2003)
3. Giese, M., Poggio, T.: Neural Mechanisms for the Recognition of Biological Movements. Nature Reviews, 4:179192, (2003)
4. Rizzolatti, G., Fogassi, L., Gallese, V.: Neurophysiological Mechanisms Underlying the Understanding and Imitation of Action. Nature Reviews, 2:661670, (2001)
5. Pastor, P., Hoffmann, H., Asfour, T., Schaal, S.: Learning and generalization of motor skills by learning from demonstration. ICRA'09: Proceedings of the 2009 IEEE international conference on Robotics and Automation, 1293-1298 (2009)
6. Gehrig, D., Kuehne, H., Woerner, A., Schultz, T.: HMM-based Human Motion Recognition with Optical Flow Data. 9th IEEE-RAS International Conference on Humanoid Robots, Humanoids (2009)
7. Krueger, N., Piater, J., Woergoetter,F., Geib, Ch., Petrick, R., Steedman, M.; Ude, A., Asfour, T., Kraft, D., Omrcen, D., Hommel, B., Agostino, A., Kragic, D., Eklundh, J., Kruger, V., Dillmann, R.: A Formal Definition of Object Action Complexes and Examples at different Levels of the Process Hierarchy (2009)
8. Barbic, J., Safonova, A., Pan, J.-Y., Faloutsos, C., Hodgins, J.K., Pollard, N.S.: Segmenting Motion Capture Data into Distinct Behaviors. In Graphics Interface, 185-194 (2004)
9. Reng, L., Moeslund, T.B., Granum, E.: Finding Motion Primitives in Human Body Gestures. GW 2005, number 3881 in LNAI, Springer, 133-144 (2006)
10. Aksoy, E.E., Abramov, A., Woergoetter, F., Dellen, B.: Categorizing Object-Action Relations from Semantic Scene Graphs, IEEE International Conference on Robotics and Automation, 398-405 (2010)
11. Sridhar, M., Cohn, G.A., Hogg, D.: Learning functional object categories from a relational spatio-temporal representation, 18th European Conference on Artificial Intelligence, (2008)
12. Guerra-filho, G., Aloimonos, Y.: Towards a sensorimotor WordNet SM : Closing the semantic gap. In Proc. of the International WordNet Conference (GWC) (2006)
13. Ivanov, Y.A., Bobick, A.F.: Recognition of Visual Activities and Interactions by Stochastic Parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 852-872 (2000)
14. Goehner, U.: Einfuehrung in die Bewegungslehre des Sports, Teil 1: Die sportlichen Bewegungen (Introduction to human movement science, part 1: sports movements). Hofmann, Schorndorf (1992)
15. Kelso, J.A.S., Fuchs, A., Lancaster, R., Holroyd, T., Cheyne, D., Weinberg, H.: Dynamic cortical activity in the human brain reveals motor equivalence. Nature 392, 814-818 (1998)
16. Simonidis, C., Seemann, W.: MkdTools - human models with Matlab. In Wassink, R. (Ed.), The 10th International symposium on 3D Analysis of Human Movement - Fusion Works (2008)
17. Soltau, H., Metze, F., Fügen, C., Waibel, A.: A one-pass decoder based on polymorphic linguistic context assignment. ASRU, 214-217 (2001)
18. Finke, M., Geutner, P., Hild, H., Kemp, T., Ries, K., Westphal, M.: The Karlsruhe-Verbmobil speech recognition engine. ICASSP, 1:83-86 (1997)