Representation and Recognition of Complex Actions: A Grammatical-Semantic Approach

Fatemeh Ziaeetabar, Minija Tamosiunaite and Florentin Wörgötter

***Abstract*— Human-Robot Interaction (HRI) is a multidisciplinary field with contributions from human–computer interaction, artificial intelligence, robotics, natural language understanding, humanities and social sciences. It has many potential applications in industry, service, education and medicine. To have an efficient interaction, a robot must be able to properly understand human actions and respond appropriately in a short time. To achieve this goal, it is not only necessary for the robot to represent and recognize compound actions, but also to be equipped with a correct understanding of time in order to regulate the temporal relationship between its actions relative to human actions. To this end, we have presented a hybrid description-based method with a bottom-up approach in which we first define the actions of the constituent unit under the heading of atomic actions in the form of a quadruple. These actions can be represented and recognized by our previously defined ESEC framework. Then by defining the possible temporal relations between the atomic actions and their inclusion in a Context-Free Grammar (CFG) in a semantic way, each complex action or interaction is represented as a composition of the atomic actions.**

This approach was used to classify the manipulation actions in the MANIAC dataset and achieved a remarkable accuracy of 91%.

I. Introduction

Manipulation actions are an important category of human actions. Robots equipped with the ability to represent, recognize and execute these actions, cause a significant progress in industry as well as services. Many of the interactions that take place between humans and robots involve manipulation tasks. Therefore, manipulation actions are an important part of the "Human Robot Interaction (HRI)" area. Consider this simple scenario as an example: a human and a robot plan to work together to make a sandwich. First, the human takes a piece of bread and places it on a plate, then while the human heads towards the second piece of bread, the robot puts a spoon in the jar of jam and carries it towards the plate, then spreads some marmalade on the bread in the plate. The spoon is next placed away, coconut powder container is grasped and sprinkled on the spread marmalade. Next, the human places the second piece on the bread with marmalade and thus the sandwich is prepared. This interaction is a combination of some simpler manipulation actions such as "pick and place", "put on top", "take down", "shake", "rub" and etc.

To perform this interaction efficiently, the following conditions are necessary:

- The robot at every stage understands what the human is doing.
- The robot performs an action to complement the human's action in order to advance towards the determined goal.

In addition, if the robot is also equipped with the ability to predict actions (and not only recognizing them), while the human has not yet completed his/her task, the robot will then be able to plan and execute the next task that should be done in continuation (or along with) of the human task, thus speeding up the interaction.

In order to interact efficiently, we need to design a framework that can represent simple and complex manipulation actions, and then use this representation framework to recognize and predict actions. So far, however, several methods have been proposed for this purpose, but most of them are only useful for representing and recognizing simple actions, while their use in complex actions and interactions is challenging.

Therefore, we need a concise and abstract way for representing the semantics of manipulation actions, avoiding the tedious and inefficient details that make it difficult to generalize. To achieve this, we developed the so-called Enriched Semantic Event Chain (ESEC) framework [1], which is a much extended version of the Semantic Event Chain (SEC) [2]. ESEC uses different static spatial (''around, above, below, inside…'') and dynamic spatial (''getting close, moving apart''…) relations between each pair of manipulated objects involved in a manipulation in its semantic description, while SEC only uses "touching" and "not-touching" information.

Previously, we applied the ESEC framework to represent, recognize, and predict simpler (push, put…) and slightly more complex (cut, stir…) manipulation actions. Now, by combining this framework with a Context-Free Grammar (CFG) structure, we intend to create a high-level action descriptor with the ability to generalize to complex actions as well as interactions. To this end, we first define a number of atomic actions as the cornerstone of our new framework. We then define all possible temporal relationships between these atomic actions and next, by combining the two together with the grammatical rules, we create a high level semantic descriptor for the definition and representation of complex actions and interactions.

II. Related Works

F.Z., M.T and F.W. are with the III Physikalisches Institut-Biophysik, Department for Computational Neuroscience, Bernstein Center for Computational Neuroscience, Georg-August-Universität Göttingen (e-mail: fziaeetabar@gwdg.de).

There are two distinct approaches in manipulation actions representation and execution: one at the trajectory level [3] and the other at the symbolic level [4]. The former gives more flexibility for the definition of actions, while the latter defines actions at a semantic level which allows for generalization and planning actions at a higher level. For trajectory level representation, there are several well-established techniques such as splines [5], Hidden Markov Models (HMMs) [6], Gaussian Mixture Models (GMMs) [7] and Dynamic Movement Primitives (DMPs) [3, 8]. On the other hand, high level symbolic representations usually use graph structures and relational representations [9, 2]. Sridhar et al. [9] represented a whole video sequence by an activity graph with discrete levels, each of which represent qualitative spatial and temporal relations between objects involved in activities, however, large activity graphs and the difficulty of finding the exact graph isomorphism makes this framework expensive and sensitive to noise. Along the same line, Aksoy et al. [2] used semantic event chains (SECs) as a high-level action descriptor. SECs are generic action descriptors that capture the underlying spatio-temporal structure of continuous actions by sampling only decisive key temporal points derived from the spatial interactions between hands and objects in the scene.

On the other hand, various methodologies have recently been developed toward the recognition of high-level activities. The approaches are classified into three categories: statistical approaches, syntactic approaches, and description-based approaches. In the case of statistical approaches, one statistical model is generally constructed for each activity and then the likelihood between the corresponding activity model and a given input image sequence is computed [10, 11]. Syntactic approaches model human activities as multiple production rules generating a string of symbols, and adopt parsing techniques from the field of programming languages to recognize the activities from a given string [12, 13]. While, the description-based approaches recognize human activities by maintaining their description (or representation) on the temporal and spatial structure of the activities which they are designed to recognize [14, 15].

Here, we introduce a description-based method for representation and recognition of complex manipulation actions -which is a combination of atomic actions- as well as interactions. We plan to represent each atomic action by the ESEC framework and then combine them through involving temporal relation rules in a Context-Free Grammar (CFG) format.

III. Our Approach

A. Atomic Actions

To define atomic action as the smallest unit of an action, we need to clarify the components of this question: "WHO did WHAT, on WHICH (object), and WHERE?". In fact, atomic actions must provide a complete definition of the performer of the actions, the type of action, the object on which the action is performed, and the place where the action has occurred.

If each atomic action is explained as a sentence, "who", "what", "which" and "where" are placed in the position of "subject", "verb", "object" and "adverb", respectively. Therefore, we represent each atomic action in the form of the following quadruple:

(Subject, Verb, Object, Adverb)

For example, if a hand touches a book on a table, its corresponding quadruple will be: (Hand, touch, book, table).

In this way, we have defined a list of fourteen atomic actions (Figure 1). They can describe more complex actions in combination with each other.

Atomic Actions	Explanation	Quadruple
A1	Hand touches an object on the ground	(H,T,O,G)
A2	Hand touches an object on another object	(H,T,O1,O2)
A3	Hand untouches an object on the ground	(H,U,O,G)
A4	Hand untouches an object on the other object	(H,U,O1,O2)
A5	Merged entity touches an object on the ground	(ME,T,O,G)
A6	Merged entity touches an object on another object	(ME,T,O2,O3)
A7	Merged entity untouches an object on the ground	(ME,U,O,G)
A8	Merged entity untouches an object on another object	(ME,U,O2,O3)
A9	Merged entity touches the ground	(ME,T,G,null)
A10	Merged entity untouches the ground	(ME,U,G,null)
A11	Merged entity moves together on an object on the ground	(ME,MT,O2,G)
A12	Merged entity moves together on an object on another object	(ME,MT,O2,O3)
A13	Merged entity moves together on the ground	(ME,MT,G,null)
A14	Merged entity moves together on the air	(ME,MT,air,null)

Figure 1: List of our fourteen defined atomic actions. "H", "ME", "O","T", "U", "MT", "G" represent "Hand", "Merged Entity =hand + a touched object", "an object", "touch", "untouch", "Move together" and "ground" respectively. Moreover, "null" is used as the fourth item of our quadruple when the ground is an object on which the action is performed on and "air" is used when an action is performed in the air without a support surface.

Each of the above atomic actions are easily representable and distinguishable by the ESEC framework. Figure 2 shows an example of the sequence of three atomic actions (A1, A2 and A3 according to the Figure 1's list) in construction of the "Pushing" action. This action is a combination of "touching an object on the ground by hand", "moving that object on the ground by hand" and "removing the hand from the object".

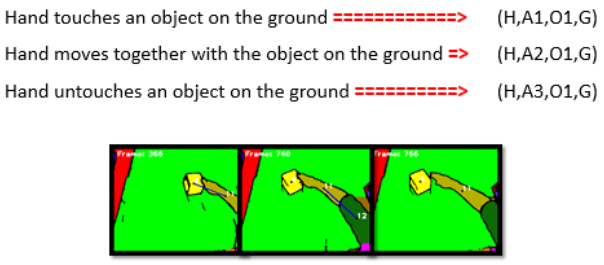

Figure 2: "Pushing" action is a combination of three atomic actions.

B. Temporal Relations

Two atomic actions can have some temporal relations to each other. For example, if we call them A1 and A2, they can have the same start point or end point, A1 can happen during A2, A2 can take place after A1 and etc. Figure 3 includes all possible temporal relations between them.

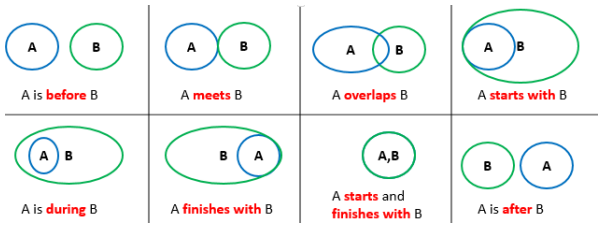

Figure 3: The possible temporal relations between two atomic actions.

C. Complex Actions

In complex actions, there are a number of constructive atomic actions that follow each other and the beginning of one coincides with the end of the other. Figure 4 shows the sequence of atomic actions in the construction of "put on top" action.

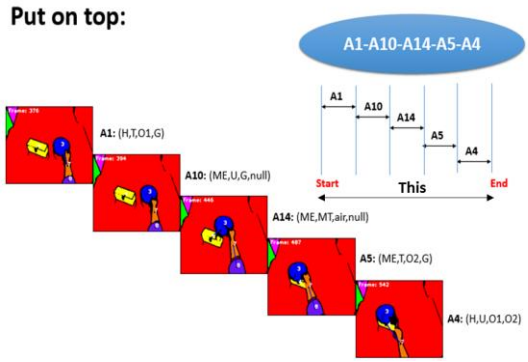

Figure 4: Atomic actions sequence in "put on top" action.

Thus, according to Figure 3 the temporal relations between each pair of the atomic actions is "meets". If the whole operation time of "put on top" action is shown with "This" parameter, the following temporal relations are obtained:

A1 **starts with** "This".
A1 **meets** A10.
A10 **meets** A14.
A14 **meets** A5.
A5 **meets** A4.
A4 **finishes with** "This"

D. Interactions

An interaction involves several actions performed by several hands in parallel or sequentially to advance a goal in a scene. An interaction example between two persons (two hands) is shown in Figure 5.

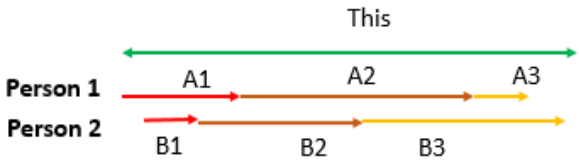

Figure 4: An example of an interaction between two persons.

The temporal relations between the atomic actions of the two participants in the interactions shown above are as follows:

A1 **starts with** "This".
A2 **meets** A1.
A3 **meets** A2.
B2 **meets** B1.
B3 **meets** B2.
B1 is **during** A1.
B2 **overlaps** A1 and A2.
B3 **overlaps** A2.
A3 **during** B3.
B3 **finishes** with "This".

E. Context- Free Grammar (CFG)

So far, we have defined the atomic actions and the possible temporal relations between them and introduced complex actions as well as interactions as a chain of atomic actions considering their temporal relations. We now intend to express the method of representing these actions using Context-Free Grammars (CFG).

A CFG is defined as ***G* = <*S*, *N*, *T*, *P*>,** where ***S*** is the start point, ***N*** is list of non-terminals and ***T*** contains list of terminals. Moreover, ***P*** includes the list of production rules. In a CFG, every production rule is of the form: $A \rightarrow \alpha$, where A is a single nonterminal symbol, and α is a string of terminals and/or non-terminals. A formal grammar is considered "context free" when its production rules can be applied regardless of the context of a nonterminal. No matter which symbols surround it, the single nonterminal on the left-hand side can always be replaced by the right-hand side.

We borrow the structure of these grammars in our purpose of action representation. In this scope, "hand(s)", "atomic actions type", "temporal relations", "object" and "places" are defined as terminals. Moreover, "AS" or "Action Sentence" and "HS" or "Hand Sentence" are our non-terminals.

Currently, we have defined the following grammar to parse complex actions to their components and we further extend this grammar to parse and describe interactions by adding the possibility of having more than one hand and also including the temporal relations.

Production rules:

1) AS -> AOP|AHS
2) HS -> HAS|HSAS
3) H -> hand
4) A -> atomic actions
5) O -> objects
6) P -> place

The architecture of this grammar is motivated by the following observations: 1) the main and only driving force in manipulation actions are the hands. Thus, a specialized non-terminal symbol "H" is used for their representation; 2) An "Action" (A) can be applied to an "Object" (O) directly on a "Place" (P), or to a "Hand Sentence" (HS), which in turn contains an "Object" (O). This is encoded in Rule (1), which builds up an "Action Sentence" (AS); 3) An "Action Sentence" (AS) can be combined either with the "Hand" (H), or a "Hand Sentence" (HS). This is encoded in rule (2), which recursively builds up the "Hand Sentence". The rules above form the syntactic rules of the grammar used in the parsing algorithms.

This definition is similar to the one proposed earlier by Yang et al in [16], although they did not consider the concept of time and place.

Figure 6 shows how the "Pushing" action –which was shown in Figure 2- is parsed and decomposed into its three atomic sub-actions (A1, A2 and A3 according to the Figure 1's list) according to this grammar.

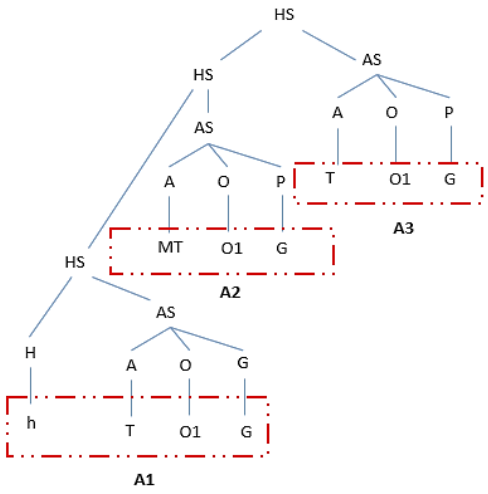

Figure 5: Decomposition of "Pushing" actions to its components.

IV. Results

In this section, we examine the proposed grammatical method -based on the decomposition of complex actions into their constructive atomic operations and the recognition of atomic operations by the ESEC framework- on the MANIAC data set [17]. This dataset consists of the following eight manipulation actions: "push", "put", "take", "stir", "cut", "chop", "hide" and "uncover". Each action type is performed in fifteen different versions by five human actors. Each version has a differently configured scene with different objects and poses.

We first decompose the compound actions based on the grammar presented in Section III-E into their constructor atomic actions and simply recognize these atomic actions by the ESEC framework. Then, by putting them together, we make descriptive chains like what can be seen in Figure 4. Finally, the action chains are grouped by applying a classification algorithm that considers ten samples for training and five samples for testing (among fifteen samples for each action). Figure 7 includes the classification results on MANIAC dataset.

	Put on top	Take down	Push	Cut	Chop	Stir	Put over	Uncover
Put on top	100%	0%	0%	0%	0%	0%	0%	0%
Take down	0%	100%	0%	0%	0%	0%	0%	0%
Push	0%	0%	100%	0%	0%	0%	0%	0%
Cut	4%	0%	0%	74%	22%	0%	0%	0%
Chop	2%	0%	0%	16%	82%	0%	0%	0%
Stir	0%	0%	0%	4%	0%	96%	0%	0%
Put over	0%	0%	0%	0%	0%	0%	92%	4%
Uncover	0%	0%	0%	0%	0%	0%	12%	88%

Figure 6: Classification accuracy of 8 manipulation actions in MANIAC dataset.

Our description-based hybrid method, which is a mixture of a grammatical approach and the ESEC framework, eliminates many unpractical details such as the shape of objects, their static and dynamic spatial relations, arrangement of the scene, hand trajectories - that vary from person to person – and classifies the manipulations with 91.5% accuracy.

V. Conclusion

In this paper, we presented a hybrid semantic- grammatical method to represent and recognize complex manipulation actions. Here, the same cognitive approach that humans use to recognize actions -i.e., breaking a compound action into its smaller constructive actions- has been applied. From a bottom-up perspective, first the basic components of actions were described as atomic actions. Each atomic action was defined as a quadruple which includes the basic information (performer, object, action type, occurrence place) about that action. Then we defined the possible temporal relations between the atomic actions. In complex actions, each atomic action appears after the previous one, while the atomic actions can have more variant temporal relations in interactions.

Next, a CFG was used to represent complex actions. This grammar decomposed complex actions into their components (top-down approach) and, conversely, allows compound actions to be produced by combining simpler actions (bottom-up approach). This method resulted in a successful classification on the MANIAC dataset, which outperformed the classification achieved by using only ESECs [1].

In future, we intend to strengthen our CFG by involving the temporal relations as well as enabling it to accept several hands as the new terminals. This will allow for the recognition and prediction of interactions.

References

[1] F. Ziaeetabar, E. E. Aksoy, F. Wörgötter, and M. Tamosiunaite, "Semantic analysis of manipulation actions using spatial relations," in

2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2017, pp. 4612–4619.

[2] E. E. Aksoy, A. Abramov, J. Dörr, K. Ning, B. Dellen, and F. Wörgötter, "Learning the semantics of object–action relations by observation," The International Journal of Robotics Research, vol. 30, no. 10, pp. 1229–1249, 2011.

[3] A. J. Ijspeert, J. Nakanishi, and S. Schaal, "Movement imitation with nonlinear dynamical systems in humanoid robots," in Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292), vol. 2. IEEE, 2002, pp. 1398–1403.

[4] R. Dillmann, T. Asfour, M. Do, R. Jäkel, A. Kasper, P. Azad, A. Ude, S. R. Schmidt-Rohr, and M. Lo¨sch, "Advances in robot programming by demonstration," KIT-Künstliche Intelligenz, vol. 24, no. 4, pp. 295–303, 2010.

[5] A. Ude, "Trajectory generation from noisy positions of object features for teaching robot paths," Robotics and Autonomous Systems, vol. 11, no. 2, pp. 113–127, 1993.

[6] D. Lee and Y. Nakamura, "Stochastic model of imitating a new observed motion based on the acquired motion primitives," in 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2006, pp. 4994–5000.

[7] S. Calinon, F. Guenter, and A. Billard, "On learning, representing, and generalizing a task in a humanoid robot," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 37, no. 2, pp. 286–298, 2007.

[8] T. Luksch, M. Gienger, M. Mühlig, and T. Yoshiike, "A dynamical systems approach to adaptive sequencing of movement primitives," in ROBOTIK 2012; 7th German Conference on Robotics. VDE, 2012, pp. 1–6.

[9] S. Park and J. K. Aggarwal, "A hierarchical Bayesian network for event recognition of human actions and interactions," Multimedia systems, vol. 10, no. 2, pp. 164–179, 2004.

[10] S. Park and J. K. Aggarwal, "Semantic-level understanding of human actions and interactions using event hierarchy," in 2004 Conference on Computer Vision and Pattern Recognition Workshop. IEEE, 2004, pp. 12–12.

[11] Y. Shi, Y. Huang, D. Minnen, A. Bobick, and I. Essa, "Propagation networks for recognition of partially ordered sequential action," in Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004, vol. 2. IEEE, 2004, pp. II–II.

[12] D. Minnen, I. Essa, and T. Starner, "Expectation grammars: Leveraging high-level expectations for activity recognition," in 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings, vol. 2. IEEE, 2003, pp. II–II.

[13] M. Sridhar, A. G. Cohn, and D. C. Hogg, "Learning functional object categories from a relational spatio-temporal representation," in ECAI 2008: 18th European Conference on Artificial Intelligence (Frontiers in Artificial Intelligence and Applications). IOS Press, 2008, pp. 606–610.

[14] S. Hongeng, R. Nevatia, and F. Bremond, "Video-based event recognition: activity representation and probabilistic recognition methods," Computer Vision and Image Understanding, vol. 96, no. 2, pp. 129–162, 2004.

[15] M. S. Ryoo and J. K. Aggarwal, "Semantic representation and recognition of continued and recursive human activities," International journal of computer vision, vol. 82, no. 1, pp. 1–24, 2009.

[16] Y. Yang, C. Fermuller, and Y. Aloimonos, "A cognitive system for human manipulation action understanding," in the Second Annual Conference on Advances in Cognitive Systems (ACS), vol. 2. Citeseer, 2013.

[17] E. E. Aksoy, M. Tamosiunaite, F. Wörgötter, Model-free incremental learning of the semantics of manipulation actions, Robotics and Autonomous Systems 71 (2015) 118–133.