2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Congress Center Hamburg
Sept 28 - Oct 2, 2015. Hamburg, Germany

Semantic Parsing of Human Manipulation Activities using On-line Learned Models for Robot Imitation

E. E. Aksoy, M. J. Aein, M. Tamosiunaite, and F. Wörgötter

Abstract— **Human manipulation activity recognition is an important yet challenging task in robot imitation. In this paper, we introduce, for the first time, a novel method for semantic decomposition and recognition of continuous human manipulation activities by using on-line learned individual manipulation models. Solely based on the spatiotemporal interactions between objects and hands in the scene, the proposed framework can parse not only sequential and concurrent (overlapping) manipulation streams but also basic primitive elements of each detected manipulation. Without requiring any prior object knowledge, the framework can furthermore extract object-like scene entities that are performing the same role in the detected manipulations. The framework was evaluated on our new egocentric activity dataset which contains 120 different samples of 8 single atomic manipulations (e.g.** *Cutting* **and** *Stirring*) **and 20 long and complex activity demonstrations such as "***making a sandwich***" and "***preparing a breakfast***". We finally show that parsed manipulation actions can be imitated by robots even in various scene contexts with novel objects.**

I. INTRODUCTION

Understanding continuous human action sequences is of great interest in computer vision and robotics. Conventional approaches essentially apply motion pattern analyses and appearance-based feature matching methods for automatic temporal segmentation and monitoring of human actions. Here, we go beyond this by introducing a novel method that relies only on the spatiotemporal interactions between the existing objects in the scene and allows manipulation sequence modeling, segmentation, recognition, as well as manipulation object categorization to be used in robot imitation, all in the same coherent Semantic Event Chain framework.

"Semantic Event Chains" (SECs) were introduced in [1] as a possible encoding scheme for manipulation actions. From a visual input stream, the SEC framework extracts sequences of changes between the spatial relations of the objects and hands in the scene. SECs store these descriptive change-patterns, which remain the same for a given manipulation type even when there are large variations in trajectory, pose, velocity, and objects. Thus, SECs can be used to classify manipulations as well as to categorize manipulated objects, as shown earlier in [1], [2].

In this paper we expand the approach by showing how: (a) semantic action models can be on-line learned from observation in a grounded and unsupervised way without requiring prior action knowledge; (b) the semantic information embedded in the event chains of long and complex manipulation activities (e.g. "*preparing a breakfast*") can be used to parse atomic actions (e.g. *cutting* or *stirring*) performed either sequentially or concurrently; (c) combinatorial explosions in the manipulation recognition and object categorization phases can be avoided by excluding the object recognition step; (d) semantically important entities (e.g. manipulators) common to all manipulation actions can be identified also in a model free way. Moreover, we show that extracted semantic action models can be further employed for imitating parsed human demonstrated manipulations with robots even in various scene contexts. The general idea of employing the semantics for manipulation analysis has a well grounded theoretical background which has been deeply investigated in [3] by using 26 different manipulation types.

Fig. 1 depicts the block diagram of our approach. The main inputs for the framework are image streams captured from human demonstrated manipulation activities. Each object in the recorded *RGB-D* image stream is uniquely segmented and separately tracked by employing the method from [4]. Spatiotemporal interactions between tracked image segments are further encoded with the concept of SECs [1]. The main contribution of our paper, as indicated by a dashed line in Fig. 1, is the semantic decomposition and recognition of the encoded long manipulation activities using previously learned SEC models. For this purpose, the extracted SEC representation of a long activity is first scanned to estimate the manipulator and manipulated objects, i.e. hand and knife, without employing any visual feature-based object recognition method. Based on the interactions between the hand and objects, the SEC is partitioned into fragments to extract serial and parallel manipulation streams. Each parsed manipulation stream is compared with model SECs in the recognition process. SEC models, stored in the library, are learned in an on-line unsupervised fashion using the semantics of manipulations derived from a given set of training data in order to create a large vocabulary of single atomic manipulations. Finally, learned SEC models are employed for reproducing detected manipulations with robots by using the method in [5].

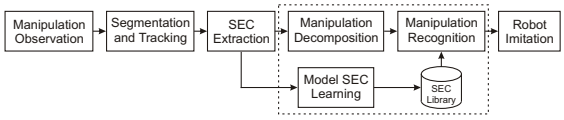

Fig. 1. Block diagram of the algorithm.

*The research leading to these results has received funding from the European Community, Seventh Framework Programme FP7/2007-2013 (Programme and Theme: ICT-2011.2.1, Cognitive Systems and Robotics) under grant agreement no. 600578, ACAT.

Third Institute of Physics & BCCN, University of Göttingen, Friedrich-Hund Platz 1, D-37077, Göttingen, Germany [eaksoy,maein,minija,worgott] at gwdg.de

II. State of the Art

There is a large corpus of work in both temporal decomposition [6], [7], [8] and recognition [9], [10], [11] of actions in computer vision and robotics. Most of the previous works in action recognition [12], [13], [14] consider only intrinsic hand configurations or body movements (e.g. *walking or running*) as the main features. To date, there are only a few approaches [15], [9], [10], [16], [11] attempting to reach the semantics of manipulation actions (e.g. *pushing or cutting*) in conjunction with the manipulated objects.

The work in [7] introduced a probabilistic graphical model with additional substructure transition and discriminative boundary models to tackle the problem of continuous action segmentation and recognition. An unsupervised hierarchical bottom-up framework was presented in [8] for temporal partitioning of human motion into disjoint segments. Such approaches are based on bottom-up continuous motion patterns which, however, have high variability in each individual demonstration of an action. Therefore, these methods require fully labeled large training data and suffer from the generalization problem in the case of having limited amount of training data. The computational complexity, as seen in [8], also limits their applicability to long activities.

The recent work in [6] proposed an event parsing approach based on stochastic event grammar by employing binary spatial relations (e.g. touch or near) between objects and agents in the scene. This approach heavily relies on the semi-supervised object recognition results to infer intended actions. A 4D spatiotemporal human-object interaction model was proposed in [17] in order to relate human pose and object in 3D space. Although approaches in [15], [9], [10], [11] consider interactions between objects and hands in manipulation activities, they are not suitable for detecting and recognizing parallel (e.g. overlapping) action streams since the applied models can only assign one label to each parsed action segment. The work in [16] stores the entire manipulation stream in a single and large activity graph. This method is rather for short lasting actions and difficulty of finding exact graph isomorphisms in the resulting complex structure makes this framework expensive and sensitive to noise. In contrast to HMM-based generative approaches, as seen in [14], the SEC framework obeys the Markovian assumption with the difference that all states, i.e. columns of SECs, are observable and represent topological changes in the scene.

To our best knowledge, our work is the first study that applies semantic reasoning to decompose chained manipulation activities and to recognize embedded sequential and parallel manipulations in conjunction with the manipulated objects without employing prior object knowledge. The theoretical background for the idea of using touching (or grasping) event sequences, as encoded in SECs, had been analyzed independently by two different groups [3], [18], who showed that such sequences can be considered as an action grammar very similar to that found in human language. These works form the scientific grounding of our approach.

III. Method

In the following, we provide detailed descriptions of the different algorithmic steps shown in Fig. 1.

A. Manipulation Observation

Since the proposed framework focuses on the spatiotemporal interactions between the manipulated objects and hands, every demonstrated human manipulation is recorded from the subject's own point of view with a static *RGB-D* camera. The top row in Fig. 2 depicts some original images from a sample *Cutting* manipulation demonstration.

B. Segmentation and Tracking

The recorded image sequences are first pre-processed by a real-time, color and depth-based image segmentation method to uniquely identify and track all objects including hands in the scene. Since segmentation and tracking approaches are not in the core of this paper and were comprehensively described elsewhere [4], we omit details here.

C. Semantic Event Chain (SEC) Extraction

During the segmentation and tracking phase, a foreground background segmentation is applied to ignore the supporting surface in the scene. Each consistently segmented image is then represented by a graph: nodes represent segment centers and edges indicate whether two segments touch each other. By employing an exact graph matching method, the continuous graph sequence is discretized into decisive main graphs, i.e. "key frames", each of which represents a topological change in the scene. All extracted main graphs form the core skeleton of the SEC, which is a matrix where rows are spatial relations (e. g. touching) between object pairs and columns describe the scene configuration when a new main graph occurs. Possible spatial relations are *Not touching (N)*, *Touching (T)*, and *Absence (A)*. N means that there is no edge between two spatially separated object segments, T represents a touching event between two objects, and the absence of an object yields A. Fig. 2 depicts the SEC representation of a sample *Cutting* demonstration, in which a hand is first taking a knife, cutting a cucumber and then withdrawing. For instance, the first row of the SEC represents the spatial relations between graph nodes 9 and 6 which are the right hand and knife, respectively. Note that, although the scene involves more objects, the SEC representation only encodes object pairs that produce at

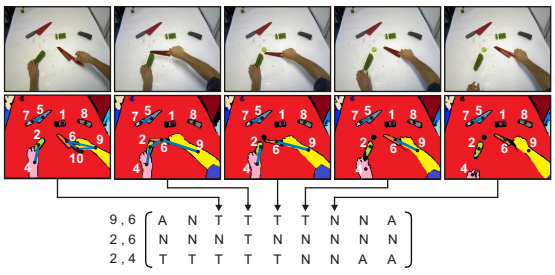

Fig. 2. SEC representation of a sample *Cutting* action.

least one relational change from N to T or vise versa since all other pairwise relations (e.g. between the left and right hands) are static and irrelevant. On top of Fig. 2 sample *key frames* including original images, respective tracked segments (colored regions), and corresponding main graphs are given to illustrate the topological configurations at the related SEC columns. The concept of SECs, briefly explained here, was introduced in [1].

D. Model SEC Learning

The main aim of the learning is to generate a vocabulary of single atomic manipulations like *Cutting*, *Stirring*, or *Pushing*. Such a vocabulary can then be employed to monitor the decomposed long manipulation activities.

Fig. 3 illustrates the on-line unsupervised learning framework which is triggered whenever a new manipulation sample is observed. At start, an individual manipulation is shown and the first extracted SEC sample is assumed to be the first "model" and stored in a "library". We then encode the manipulation that follows again by a SEC. Next, we compare it with all existing SEC models in the library. For this purpose, the framework measures semantic similarities (δ) between the new SEC sample and the existing models by employing the method described in [1], which compares rows and columns of two SECs using sub-string search and counting algorithms. Computed semantic similarity values between all existing models and the new sample are stored in a matrix, called the similarity matrix (ζ_{sim}), which is then converted into a histogram ($\mathcal{H}$) representing the distribution of similarities. We apply the conventional Otsu's method [19] to the normalized histogram to distinguish low from high similarities. We take the average of the high similarities to estimate a threshold τ to classify the currently observed SEC sample against the existing models.

If similarity (δ) is higher than τ, then the new sample will be assigned to the best fitting (highest similar) model and this model will be updated with additional rows or columns that might exist in the new SEC sample [1]. In this way, the model SECs will only consist of those rows and columns observed frequently in all type-similar manipulations. If similarity (δ) is lower than τ, the novel SEC sample will be used as a new model. In addition, we merge learned SEC models, which have higher semantic similarities (φ) than τ, as they are likely representing the same manipulation.

E. Manipulation Decomposition

In the decomposition phase, we partition long manipulation activities into chunks. Fig. 4 (a) depicts the extracted event chain for a sample manipulation sequence, in which a hand is first replacing a bucket, putting an apple down and then hiding it with the same bucket. The extracted long SEC pattern is decomposed by considering the changes in the spatial interactions between graph nodes, i.e. image segments. We assume that each manipulation is composed of three main elements: *manipulator*, *primary* and *secondary objects*. The *manipulator* is the main actor, e.g. a *hand*, which plays the main role by frequently interacting with objects in the scene. The *primary object* is the one which is primarily touched by the *manipulator*. The *knife* can be considered as the *primary object* in a *cutting* action. All other objects interacting with the *primary object* are named *secondary objects*, e.g. the cucumber to be cut.

There is, however, no object recognition approach so far involved to identify graph nodes, i.e. segments, in the extracted event chains. We, therefore, first apply probabilistic reasoning to identify the naked graph nodes as *manipulator*, *primary* or *secondary objects* based only on changes in their spatiotemporal relations in the SECs. We assume that the SEC representation of a meaningful manipulation must have at least a row with a sequence of $[N, T, \cdots, T, N]$ spatial relations between the *manipulator* and *primary object*. Such a row emphasizes the entire dynamics of the manipulation by indicating that the *manipulator* is first not touching (N) the *primary object*, then touches (T) the *primary object* to apply a certain task on it. Depending on the manipulation type, the temporal length of the touching (T) event can vary. Finally, the *manipulator* releases (N) the *primary object* and continues with a different *primary object*. The sequences of such relational changes between the *manipulator* and *primary object* are then used for the temporal decomposition of long manipulation sequences. In the next two subsections we elaborate on the manipulator estimation, following with the final decomposition process.

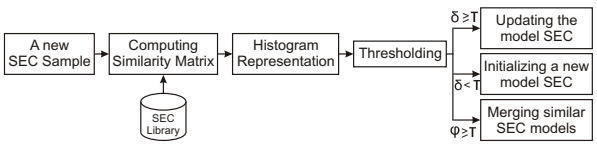

Fig. 3. Overview of the proposed on-line learning framework.

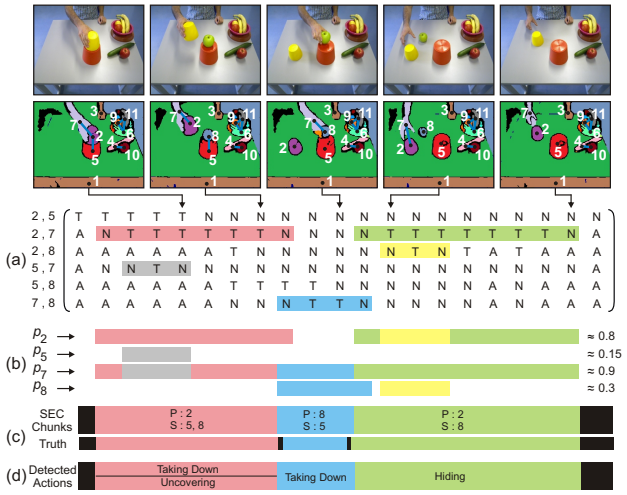

Fig. 4. (a) The extracted SEC with colored blocks highlighting a sequence of $[N, T, \cdots, T, N]$ relations. (b) Computed segment probabilities to estimate the *manipulator*. Final probability values are given on the right. (c) Decomposed SEC fragments with respect to the ground truth. Black blocks represent null actions. *P* and *S* stand for the estimated *primary* and *secondary objects*. (d) Recognized manipulations at each fragment.

1) Estimating the Manipulator: During the activity, the manipulator (hand) mainly oscillates between N and T events and meanwhile produces repetitive touching-relations T in the SEC. By counting those touching events, the manipulator can be found by employing Algorithm 1. For a given segment s_k and all rows in the SEC that contain it, we consider only sequences of $[N, T, \cdots, T, N]$ relations and the algorithm determines a probability value p_k based on how wide the touching relations T stretch along the temporal duration of the SEC. The *manipulator* is finally estimated as the segment with the highest probability value, i.e. most T's encoded after normalization. For the sake of simplicity, we consider *single-hand* manipulations, since the second hand is mainly used as a supporter.

For instance, the colored blocks shown in the SEC in Fig. 4 (a) highlight the sequences of $[N, T, \cdots, T, N]$ detected in each row. Fig. 4 (b) links those blocks to the corresponding segments in the SEC to indicate which segment has the longest block, i.e. highest p_k probability value. For example, as the segment number 2 exists only in the first three rows of the SEC, detected blocks only in these rows are superimposed and assigned as a probability value to that segment. On the right side of Fig. 4 (b), final probability values computed from Algorithm 1 are given. Since the blocks that linked to the segment number 7 (i.e. the right hand) cover the widest area in the temporal phase, it is correctly estimated as the *manipulator*.

2) Decomposing SECs: The correct estimation of the *manipulator* triggers the final decomposition process of a long event chain. In the decomposition stage, we consider the detected $[N, T, \cdots, T, N]$ sequences that belong to the *manipulator* as cutting points, because any change from N to T and from T to N defines the natural start and end points of the manipulation. It is here important to note that we cannot directly assume each $[N, T, \cdots, T, N]$ sequence as a fragment due to spurious spatial relations propagated from noisy segmentation and tracking information. Therefore, we first apply a low pass filter to those rows with the estimated *manipulator* and then label each time interval between an $[N, T]$ and $[T, N]$ change as a potential action segment, i.e. fragment. Each fragment is assigned a confidence value indicating the rate of the existing touching events. Those confident action fragments that are encapsulated by others or share a common temporal zone are merged to converge to the ultimate partitioning of the manipulation. A predefined threshold is here introduced to define the lowest confidence value and minimum common temporal zone for merging fragments.

In the same example illustrated in Fig. 4 (b), there exist, for instance, four candidate action fragments which are the red, blue, green, and gray blocks of the *manipulator*, i.e. segment number 7. However, the gray block is merged with the red one since it is entirely surrounded. Thus, the remaining three blocks construct the ultimate temporal points at which the manipulation is cut. Fig. 4 (c) illustrates the final decomposition results together with the ground truth defined by a human. Compared to the ground truth, the frame-wise decomposition accuracy is computed as 96%. Note that, the end point of each block in Fig. 4 (c) is considered as the beginning of the next consecutive one.

Algorithm 1 Computing the probability value p_k. The n and m values are the row and column numbers in the SEC ξ.

for all segment s_k in the SEC ξ **do**
  p_k = 0 (Initiation!) , δ_k = [ ] (An empty array!)
  for r=1 to n (go through all rows!) **do**
    if s_k exists in this row! **then**
      t_{Start} , t_{End} = 0 (Initiate time points!)
      for c=1 to $m-2$ (go through columns! **do**
        if $\xi(r, c:c+1)$= $[N, T]$ **then**
          t_{Start} = c (Starting time point!)
          for f=c+2 to m **do**
            if $\xi(r, f)$= $[N]$ **then**
              t_{End} = f (Ending time point!)
              $break$
          if $t_{End} > t_{Start}$ **then**
            δ_k (t_{Start},t_{End}) = 1
            t_{Start} , t_{End} = 0
  p_k = sum(δ_k)/m (Compute the final probability!)

F. Manipulation Recognition

In the recognition phase, we first distinguish the manipulated *primary* and *secondary object* segments in each parsed SEC fragment. For instance, in the temporal interval of the red block (between the second and eighth columns of the SEC) in Fig. 4 (c), the segment number 2 (the yellow bucket) is estimated as the *primary object* since it has most touching events with the previously detected *manipulator*, i.e. segment 7. Next, segments 5 and 8 (the apple and red bucket) are estimated as *secondary objects* because they are the only segments sharing a touching event with the *primary object* within the same temporal interval. Fig. 4 (c) shows all estimated *primary* and *secondary objects* in SEC fragments.

By reformulating the manipulations with interactions between the *manipulator*, *primary*, and *secondary objects*, we can diagnose the parallel streams of concurrent manipulations because of the fact that each manipulation has to have a unique *secondary object*. Detection of multiple *secondary objects* indicates either noisy segments, i.e. graph nodes, in the SEC or existence of parallel atomic manipulations, e.g. *Putting* and *Pushing*, in the demonstration. In this regard, we apply a brute force combinatorial process which treats each combination of the *manipulator*, *primary* and *secondary objects* as a separate manipulation hypothesis. The crucial rule here is that each hypothesis must consist of the entire *secondary object* set. The best hypothesis that has the highest semantic similarity with the learned SEC models is then considered as the final recognition result.

For instance, the first parsed SEC fragment, depicted by the red block in Fig. 4 (c), has two *secondary objects* (segment numbers 5 and 8). Fig. 5 illustrates the computed two hypotheses (shown in different colors) each of which

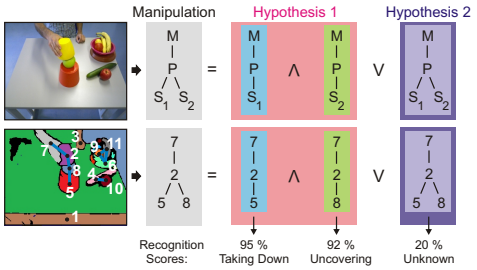

Fig. 5. Detection of parallel manipulation streams in a decomposed SEC fragment that has two *secondary objects*. *M*, *P*, and *S* stand for the *manipulator*, *primary*, and *secondary objects*.

has a different object combination, i.e. manipulation stream. The first hypothesis is composed of two separate (parallel) manipulation streams, each utilizes one *secondary object* as indicated by unique colors, whereas the next hypothesis employs both *secondary objects* together as one manipulation stream. Note that, even though the scene involves many more objects, the number of hypotheses is remaining small due to the consideration of only those objects that are sharing touching events with the *primary object*. Thus, our approach does not suffer from combinatorial explosion. Fig. 5 also shows the final similarity scores of each hypothesis stream when we compare them with the model SECs in the library. Since the first hypothesis has much higher average recognition rate, our proposed approach successfully returns two parallel manipulation streams; *Taking Down* the yellow bucket (segment 2) from the red bucket (segment 5) while *Uncovering* the green apple (segment 8). Fig. 4 (d) illustrates the final recognized manipulation types at each decomposed SEC fragment for the chained manipulation sequence depicted in Fig. 4 (a).

G. Robot Imitation

So far, we have elaborated on the activity perception. We, now, rather discuss how a recognized manipulation can be reproduced by a cognitive agent.

As highlighted in section III-C, SECs discretize the continuous action and extract temporal anchor points, each of which corresponds to one SEC column. These anchor points indicate unique and descriptive scene states, i.e. topological changes in the manipulation. Hence, we consider each transition from one SEC column to the next as a movement primitive, such as *Approach* or *Grasp*. Fig. 6 pictures a human demonstrated *Putting* action sample with the SEC representation and corresponding movement primitives. Note that these primitives are symbolic, but, on the other hand, are fully grounded at the signal level with uniquely tracked image segments. This forms the main novelty coming with our proposed framework.

In the imitation phase, we enrich the raw symbolic SEC primitives with additional object and trajectory information. Each image segment, classified as *manipulator*, *primary* and *secondary objects* (see section III-E), is identified with the method from [20]. We also encode the trajectory pattern of the manipulator with the modified Dynamic Movement Primitives (DMPs, [21]) and attach it to the respective

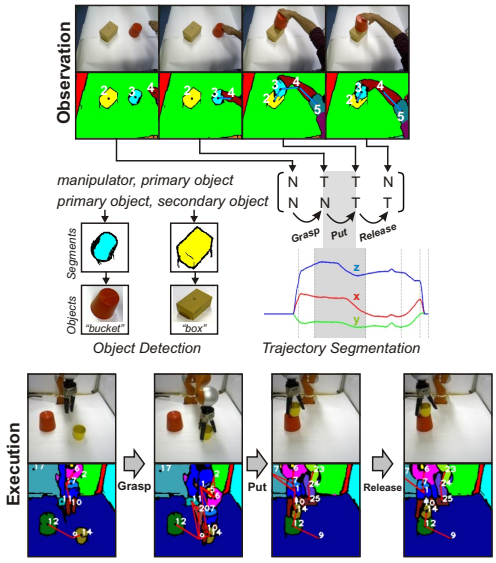

Fig. 6. Robot execution of a human demonstrated *Putting* action.

primitive in the SEC. Fig. 6 shows detected *primary* and *secondary objects* together with the trajectory segment of the *manipulator* during the *Put* primitive (shown in gray box).

Once the SEC representation is augmented with action descriptive object and trajectory parameters, we employ the state machine introduced in [5] which allows the robot to transit from one primitive the next by applying the embedded trajectory pattern to any given object. Bottom row in Fig. 6 depicts the robot execution of the recognized *Putting* action even with novel objects in a new scene context. Consequently, the semantic action descriptive features obtained with our proposed framework give rise to the reproduction of the action with robots.

IV. Results

To quantitatively evaluate the proposed framework, we created a large dataset with eight different atomic manipulation actions: *Hiding*, *Pushing*, *Putting*, *Stirring*, *Cutting*, *Chopping*, *Taking*, and *Uncovering*. Fig. 7 shows a sample frame for each manipulation type. We recorded 15 different versions for each of these manipulations with 5 different subjects who demonstrated each manipulation 3 times using in total 30 different objects in various scene contexts. The supplementary videos show sample demonstrations that highlight variations in trajectory, velocity, object type and pose.

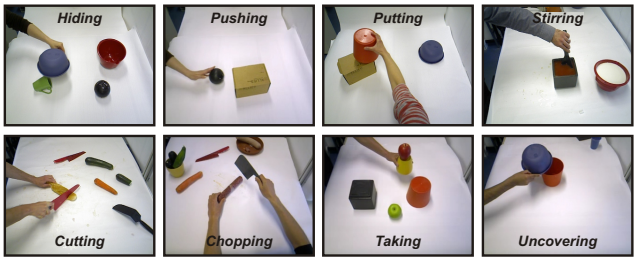

Fig. 7. Sample frames from eight different single manipulations.

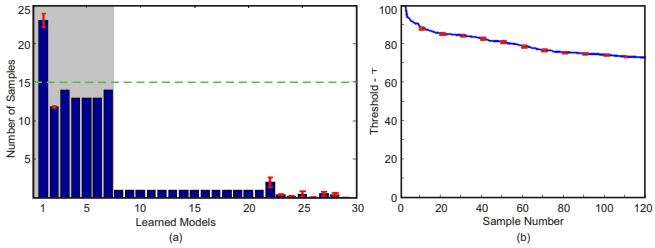

Fig. 8. Total number of learned models after 100 independent trials. (a) Learned 29 SEC models with corresponding number of trained samples. (b) Development of τ during the observation. Red bars depict the non-zero standard error of the mean.

All 120 demonstrations were recorded with the Microsoft Kinect sensor as it provides accurate depth information. To the best of the authors knowledge, this is the largest dataset [22] that only investigates egocentric manipulation activities. Note that unique background color does not play any role as the supporting surface in the scene is removed in the preprocessing step. Using distinct object colors is only for improvement of the segmentation and tracking accuracies which are not in the core of our paper.

A. Model Learning

In the first stage, we want to create a vocabulary of atomic manipulations by learning a model SEC for each manipulation type. For this, we applied our on-line unsupervised learning framework to 120 manipulation demonstrations that were presented in a random order. To investigate the robustness of the framework, we repeated the same learning experiment 100 times independently from each other and computed differences between the incrementally learned SEC models. In each trial, the framework produced minimally 21 and maximally 23 various models. However, when we compared all these models extracted in 100 trials, we saw only 29 different ones. Fig. 8 (a) depicts the distribution of all 29 learned models with corresponding number of observation samples employed for updating each. The green dashed line indicates the expected sample numbers as the ground truth. Although the framework learns in total 29 models, only 7 of them, those in the gray box, contain more than 10 samples and the rest hold at most 2 samples.

We can consequently state that the learning approach always converged to the same 7 models given 8 manipulation types. The reason is that our framework retrieved one single SEC model by naturally merging both *Cutting* and *Chopping* manipulation samples due to measuring high semantic similarity between each type. This is because both manipulations have the same fundamental action primitives, i.e. similar columns in the event chains, and the only differences are mostly in the followed trajectories and velocity of the movements which are not captured by SECs. The learning framework also generated additional models outside the gray box because some demonstrated samples have enormous variations or noise that led to a similarity lower than τ with any other models. In Fig. 8 (b), we can see the average behavior of τ after 100 independent learning trials. It was initiated with 100 and after updating at each observation of a new SEC sample it always converged to approximately 72%. Fig. 8 consequently suggests that without using any human intervention the proposed learning framework can automatically and robustly retrieve the demonstrated 8 manipulation types two of which are naturally merged.

B. Decomposition and Recognition

We additionally provided 20 long and complex manipulation activities, such as "*making a sandwich*" or "*preparing a breakfast*" as shown in Fig. 9. These chained sequences have in total 103 different versions of the trained 8 single atomic manipulations together with some novel tasks, e.g. *Pouring*. All these chained manipulations were performed in different orders, either sequentially or parallelly, with novel objects in various scene contexts to make the decomposition and recognition steps more challenging. As Fig. 9 indicates such long activities also include flickering noisy segment labels due to occlusion, e.g. segment number 3 (representing the spoon) in the fourth row switches to segment 13 as the hand passes over in the seventh frame.

To bootstrap the decomposition process, we first searched for the main *manipulator* in each scenario. We acquired 100% correct *manipulator* estimation rate in all 20 chained sequences. After applying the proposed decomposition

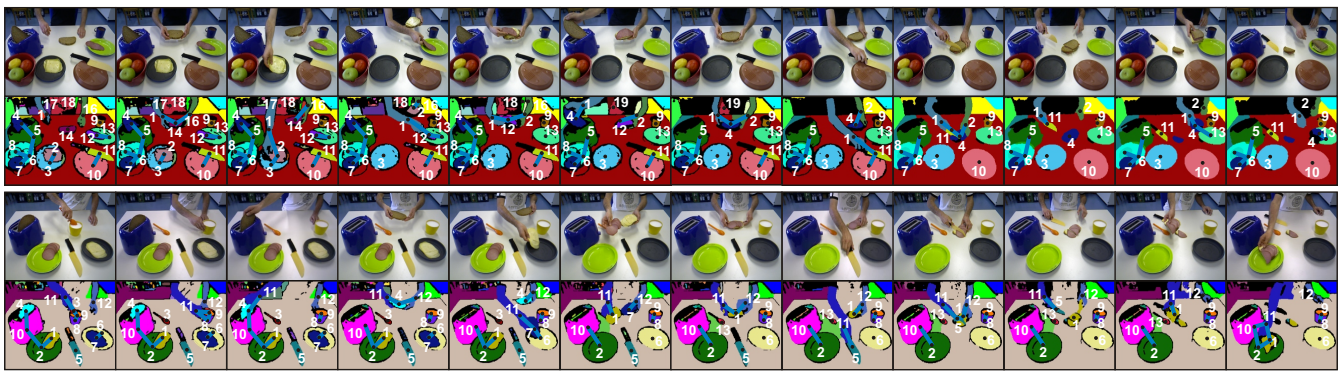

Fig. 9. Sample frames with tracked segments and graphs extracted from two different chained manipulation activities. On the top and bottom frames subjects are performing complex tasks such as "*making a sandwich*" and "*preparing a breakfast*".

method, we computed the frame-wise accuracy by comparing our decomposition results with the human defined ground truth. Fig. 10 displays the final decomposition rates for each chained sequence. The mean decomposition accuracy over all 20 sequences was computed as 91%.

Next, we evaluated the recognition rates of the sequential and parallel manipulation streams detected in the chained sequences. We compared each parsed manipulation with the learned 7 SEC models. In the case of measuring higher semantic similarity than τ, which was learned as 72%, with any known models, the parsed manipulation was assigned with the label of the respective highest similar model. If the similarity values with all known models were always less than τ, then the parsed manipulation was treated as *Unknonwn*. Fig. 11 (a) shows the confusion matrix depicting the recognition accuracies of all 103 manipulation streams, detected in the 20 chained sequences, with respect to the learned 7 SEC models. We obtained minimum 82% accurate classification rate which is for the *Taking* manipulation and maximum 10% misclassification rate as observed only for the *Pushing* manipulation. *Cutting* and *Chopping* samples were mainly classified under the same manipulation model due to the reason clarified above. It is also interesting to note that the novel *Pouring* manipulations demonstrated in the chained sequences were never confused with any of the known SEC models because of having a distinct semantics and, thus, were always classified as *Unknown*. We obtained an average recognition accuracy of 92%.

As stated in section III-F, we can also estimate the *primary* and *secondary objects* utilized in each perceived manipulation without employing any visual feature-based object recognition method. Fig. 11 (b) displays the estimated *primary object* types that were frequently manipulated in the detected manipulations. For instance, *Spoon* was the only object type primarily employed in the detected *Stirring* manipulations, whereas *Knife* and *Cleaver* were heavily preferred in the *Cutting* and *Chopping* tasks. Note that object labels are here given by a human for the sake of simplicity. Results here consequently show that our framework can highlight the direct link between actions and objects by using the encoded semantics in the event chains. Note that the usage rates of the likewise estimated *secondary objects* are not shown due to the lack of space.

Fig. 12 illustrates the final decomposition and recognition results of all chained sequences together with the human labeled ground truth. This side-by-side comparison strongly indicates the success of the perception of the parallel manipulations. Note that the lengths of chained sequences are normalized for the sake of clarity in the display.

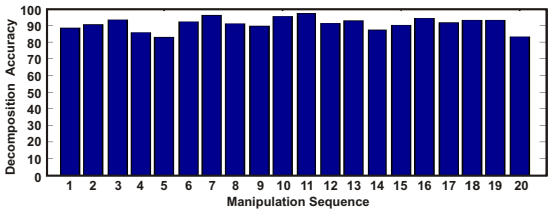

Fig. 10. Manipulation decomposition accuracies.

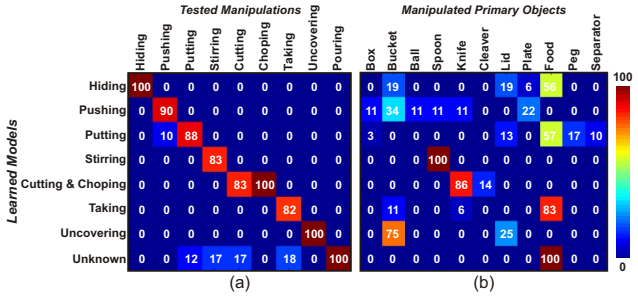

Fig. 11. Confusion matrix showing (a) the recognition accuracies of the manipulation types detected in the 20 chained sequences and (b) the usage rate of different *primary objects* in the recognized manipulation types.

C. Robot Imitation

In the imitation phase, we address how the robot executes a new scenario, e.g. "*preparing a salad*", by considering the augmented SEC models. Our robotic platform involves a single KUKA Light Weight Robot (LWR) arm equiped with a Schunk gripper for object manipulation and a Kinect sensor for the scene perception. In the detection of object relations, not only vision but we also employ the force feedback to decide whether the expected state (i.e. SEC column) is achieved. Beside the 7 atomic manipulations learned in section IV-A, we also added two more action classes *Unloading* and *Pouring* required for the salad preparation. Fig. 13 pictures sample frames from a robot execution of the chained action sequence "*preparing a salad*".

We here note that with the aid of our semantic reasoning framework, the robot now employs the same semantic structure, i.e. SECs, for both action perception and execution. Furthermore, the symbolic primitives derived form SECs are well grounded with continuous image segments and allow robots to replace manipulated objects in the execution phase.

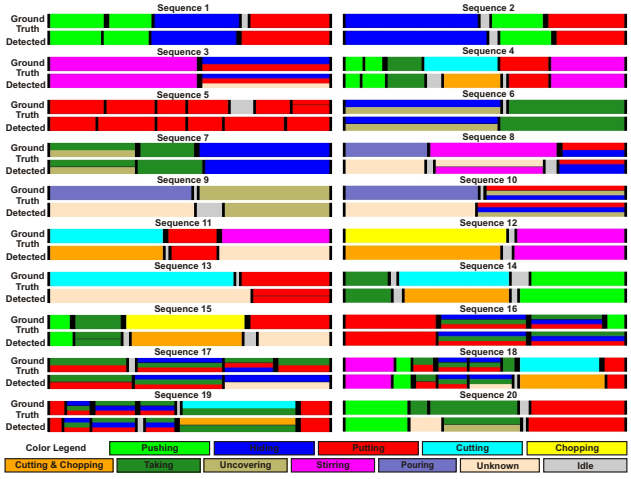

Fig. 12. Automatic decomposition and recognition results of the 20 chained manipulation sequences versus human labeled ground truth. The action fragments are color coded. Black frames indicate the border of each manipulation stream.

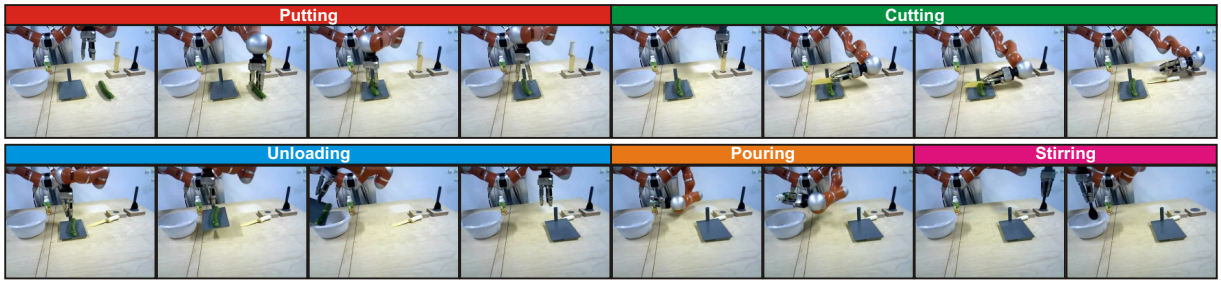

Fig. 13. Robot execution of a chained action sequence "*preparing a salad*".

V. Conclusion

Our novel semantic reasoning approach differs from existing approaches [23], [16], [11] in the sense of detecting parallel (overlapping) manipulation streams and extracting object roles in an unsupervised fashion. Generative or discriminative model-based probabilistic approaches [15], [9], [23], [10], however, cannot assign multiple labels to detected parallel manipulation streams. Compared to the recent work in [24] where each learned action is coupled with visual object features, our approach is also independent from the manipulated object type, hence, can prevent any combinatorial explosion during the recognition. This is a very strong contribution as the semantic information encoded in SECs is invariant to trajectory, velocity, object type and pose variations in the manipulations.

In contrast to the other state of the art action recognition approaches, SECs also provide movement primitives, i.e. temporal anchor points, which largely ease the task of action execution with robots. With our semantic reasoning approach, we can acquire symbolic movement primitives from continuous signals and let robots imitate the observed primitives even with altered objects.

As we are interested in only egocentric manipulations, actions including whole body movements (e.g. jumping) or deformable objects (e.g. turntable) were not in the scope of this paper. Our high-level reasoning approach can be affected by *segment discontinuity* which can happen during the segmentation and tracking. To minimize such problems, we feed our framework with accurate *RGB-D* data streams. Hence, excluding the depth component in the image streams could harm our approach.

References

[1] E. E. Aksoy, A. Abramov, J. Dörr, K. Ning, B. Dellen, and F. Wörgötter, "Learning the semantics of object-action relations by observation," *IJRR*, vol. 30, no. 10, pp. 1229–1249, 2011.

[2] E. E. Aksoy, A. Abramov, F. Wörgötter, and B. Dellen, "Categorizing object-action relations from semantic scene graphs," in *IEEE ICRA*, may 2010, pp. 398–405.

[3] F. Wörgötter, E. E. Aksoy, N. Krüger, J. Piater, A. Ude, and M. Tamosiunaite, "A simple ontology of manipulation actions based on hand-object relations," *IEEE TAMD*, 2013.

[4] A. Abramov, K. Pauwels, J. Papon, F. Wörgötter, and B. Dellen, "Depth-supported real-time video segmentation with the kinect," in *WACV*, 2012, pp. 457–464.

[5] M. J. Aein, E. E. Aksoy, M. Tamosiunaite, J. Papon, A. Ude, and F. Wörgötter, "Toward a library of manipulation actions based on semantic object-action relations," in *IROS*, 2013.

[6] M. Pei, Z. Si, B. Yao, and S.-C. Zhu, "Video event parsing and learning with goal and intent prediction," *Computer Vision and Image Understanding*, vol. 117, no. 10, pp. 1369–1383, 2013.

[7] Z. Wang, J. Wang, J. Xiao, K.-H. Lin, and T. Huang, "Substructure and boundary modeling for continuous action recognition," in *IEEE CVPR*, 2012, pp. 1330–1337.

[8] F. Zhou, F. De la Torre Frade, and J. K. Hodgins, "Hierarchical aligned cluster analysis for temporal clustering of human motion," *IEEE PAMI*, vol. 35, no. 3, pp. 582–596, March 2013.

[9] A. Gupta, A. Kembhavi, and L. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE PAMI*, vol. 31, pp. 1775–1789, 2009.

[10] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *The Inter. Jour. of Robotics Research*, vol. 32, no. 8, pp. 951–970, 2013.

[11] Y. Yang, C. Fermüller, and Y. Aloimonos, "Detection of manipulation action consequences (mac)," in *Computer Vision and Pattern Recognition*, 2013, pp. 2563–2570.

[12] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[13] C. Sminchisescu, A. Kanaujia, and D. Metaxas, "Conditional models for contextual human motion recognition," *Comp. Vision and Image Understanding*, vol. 104, no. 2-3, pp. 210–220, 2006.

[14] H. L. U. Thuc, S.-R. Ke, J.-N. Hwang, P. V. Tuan, and T. N. Chau, "Quasi-periodic action recognition from monocular videos via 3d human models and cyclic hmms," in *Int. Conf. on Adv. Tech. for Com.*, 2012, pp. 110–113.

[15] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in *Proceedings of the 2011 International Conference on Computer Vision*, 2011, pp. 407–414.

[16] M. Sridhar, G. A. Cohn, and D. Hogg, "Learning functional object-categories from a relational spatio-temporal representation," in *Euro. Conf. on Artificial Intelligence*, 2008.

[17] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, "Modeling 4d human-object interactions for event and object recognition," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 3272–3279.

[18] Y. Yang, A. Guha, C. Fermüller, and Y. Aloimonos, "A cognitive system for understanding human manipulation actions," *Advances in Cognitive Systems*, pp. 67–86, 2014.

[19] N. Otsu, "A Threshold Selection Method from Gray-level Histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[20] M. Schoeler, S. Stein, J. Papon, A. Abramov, and F. Wörgötter, "Fast self-supervised on-line training for object recognition specifically for robotic applications," in *VISAPP*, 2014.

[21] T. Kulvicius, K. J. Ning, M. Tamosiunaite, and F. Wörgötter, "Joining movement sequences: Modified dynamic movement primitives for robotics applications exemplified on handwriting," *IEEE Transactions on Robotics*, vol. 28, no. 1, pp. 145–157, 2012.

[22] E. E. Aksoy, M. Tamosiunaite, and F. Wörgötter, "Model-free incremental learning of the semantics of manipulation actions," *Robotics and Autonomous Systems*, vol. 71, pp. 118 – 133, 2015.

[23] H. Kjellström, J. Romero, and D. Kragić, "Visual object-action recognition: Inferring object affordances from human demonstration," *Computer Vision and Image Understanding*, vol. 115, pp. 81–90, 2011.

[24] A. Behera, A. Cohn, and D. Hogg, "Real-time activity recognition by discerning qualitative relationships between randomly chosen visual features," in *British Machine Vision Conference*, 2014.