Semantic Analysis of Manipulation Actions Using Spatial Relations

Fatemeh Ziaeetabar[1], Eren Erdal Aksoy[2], Florentin Wörgötter[1], and Minija Tamosiunaite[1, 3]

Abstract— **Recognition of human manipulation actions together with the analysis and execution by a robot is an important issue. Also, perception of spatial relationships between objects is central to understanding the meaning of manipulation actions. Here we would like to merge these two notions and analyze manipulation actions using symbolic spatial relations between objects in the scene. Specifically, we define procedures for extraction of symbolic human-readable relations based on Axis Aligned Bounding Box object models and use sequences of those relations for action recognition from image sequences. Our framework is inspired by the so called Semantic Event Chain framework, which analyzes touching and un-touching events of different objects during the manipulation. However, our framework uses fourteen spatial relations instead of two. We show that our relational framework is able to differentiate between more manipulation actions than the original Semantic Event Chains. We quantitatively evaluate the method on the MANIAC dataset containing 120 videos of eight different manipulation actions and obtain 97% classification accuracy which is 12 % more as compared to the original Semantic Event Chains.**

Index Terms—**Spatial relations, manipulation actions, semantic analysis, action semantics, action classification.**

I. Introduction

Action recognition and human activity analysis are the most active and challenging domains in computer vision and robotics. They play an important role in human-human as well as human-robot interactions. Also, it has many other applications in different fields such as video surveillance systems or video retrieval. Most of the researches in this area focus on full-body action categorization [1] [2], but there are a lot of tasks that an agent (human or robot) performs only using his hands (i.e., manipulation actions). Manipulation actions make a big proportion of applications both in industrial and service robotics. Intelligent robots could use observation of manipulation actions for learning how to manipulate. However, there are many ways to perform a single manipulation and it would be very inefficient to store a large set of observed examples that is not easy to generalize. The paper addresses the problem of representing manipulations in a compact and efficient way. It describes actions in terms of changes of spatial relations in the scene, while ignoring the diversity of scenes, objects and small details in the trajectory for doing the same action.

Spatial relations are abstract and functional relationships between entities in space [3]. One way of representing them is in the way humans speak about space [4] [5], e.g. "Top", "Bottom" or "Above", "Below". A correct understanding of object-wise spatial relations for a given action is essential for a robot to perform an action successfully [6]. Suppose, we ask a robot to put some object on the top of the other object. For a successful execution, in addition to the recognition of those two objects, the robot should have knowledge about "Above" and "Top" relations. It should take the first object and rise it to the "Above" of the second object and then put it on the "Top" of it. Definition of a robot action through appropriate spatial relations would lead to an accurate and generalizable performance in the robot execution.

In this regard, we apply qualitative spatial reasoning to each object pair in the scene. We use camera axes and create an Axis Aligned Bounding Box (AABB) around of each object. In the AABB representation, all box sides are parallel to the directions of axes. Next, we evaluate static and dynamic spatial relations, where the static relations set includes "Touching", "Non-touching", "Above", "Below", "Around", "Top", "Bottom" and the dynamic relations set includes "Getting Close", "Moving Apart", "Move Together", "Stable" and "Halt Together" for all pairwise objects. We design heuristic rules for evaluation of those relations and track changes in those relations during continuous video-frames.

The computed relations are embedded into the so called "Enriched Semantic Event Chain" representation, which is the extension of the original Semantic Event Chain approach [7] developed to semantically compare and identify actions [8]. We benchmark the proposed approach for accuracy in action recognition based on the MANIAC dataset [8] that includes 8 different manipulation actions (overall 120 videos performed by three different actors). To address wider action variety, we also show that the Enriched Semantic Event Chains in principle can differentiate between more actions as compared to the original Semantic Event Chains based on a 26 action set presented in [9].

II. RELATED WORKS

There has been a great deal of research in the field of spatial representation and reasoning because of its multifaceted applications in robot planning and navigation [10], interpreting visual inputs [11], computer-aided design [12], cognitive science where models of spatial skills help to

[1] Institute for Physics 3- Biophysics and Bernstein Center for Computational Neuroscience, Georg August University, Göttingen, Germany (e-mails: {fziaeetabar, worgott}@gwdg.de, minija.tamosiunaite@phys.uni-goettingen.de), [2] Karlsruher Institute for Technologies (KIT), Karlsruhe, Germany (e-mail: eren.aksoy@kit.edu). [3] Faculty of Informatics, Vytautas Magnus University, Kaunas, Lithuania.

explain human performance [13], geographic information systems (GIS) [14], and understanding natural languages [15]. All of these cases need to represent and reason about spatial aspects of the world. Spatial reasoning is studied using both quantitative and qualitative approaches. According to [16], quantitative reasoning is the developed (human) ability to analyze quantitative information and to determine which skills and procedures can be applied to a particular problem to arrive at a solution while a qualitative approach creates non-numerical descriptions of physical systems and their behavior, preserving important behavioral properties and qualitative distinctions. Qualitative spatial reasoning (QSR) provides representational primitives and inference mechanisms about space. In fact, QSR aims at capturing human-level concepts of space by using finite sets of relations to model particular spatial aspects such as topology, orientation and distance while quantitative spatial models rely on numeric calculations. Here, we would like to apply a qualitative approach because it is closer to how humans represent and reason using commonsense knowledge. It can overcome the indeterminacy problems, by allowing inference from incomplete spatial knowledge and it also offers a compact representation that is supposed to enable complex decision tasks.

Spatial reasoning techniques in artificial intelligence attempt to emulate human reasoning during navigation and other spatial planning tasks. For example, [18] applies results of brain research to obtain geometrical factors or [19] suggests a model in the form of spatial templates and prototypes (both quantitative spatial reasoning). A method of performing qualitative spatial reasoning on robots is proposed in [20].

Robotics is a domain much influenced by methods of spatial reasoning. One of the key aspects which is needed to understand commands such as "go in front of the closet door", is the ability of reasoning about spatial directions in a qualitative manner. In other words, the robot needs to be able to reason about an object with respect to another object in a given reference frame [20]. Therefore, finding spatial relations between objects in a scene is fundamental in execution of tasks by robots. In this work, we limit our study on manipulation actions that define actions which are done by hands. Because of large variation of ways for performing manipulation actions and also many occlusions in the visual scenes, manipulation action recognition is still an open and challenging problem. Meanwhile, hand movements as such have been widely investigated, but for a slightly different purpose: hand gesture recognition, for human-computer interfaces or sign language recognition [21].

In this study we concentrate on analysis of manipulation actions via the relations of manipulated objects. Only a couple of studies exist doing this type of analysis. In [22] visual semantic graphs were introduced for recognition of action consequence according to the changes in the topological structure of the manipulated objects. The study presented in [23] represents an entire manipulation by an activity graph which holds spatiotemporal interaction between objects, however, the activity graph requires complicated processing for extraction of semantic level knowledge. The work in [24] modeled human activities by involving some information about human skeleton and tracking the segments of manipulated objects. The authors of [25] use hand trajectories and hand-object interactions in a Bayesian model for manipulation observation. All the studies mentioned above introduce representations which don't abstract from multiple execution details, while we attempt to describe manipulation actions through abstract relations. The already mentioned "Semantic Event Chain" (SEC) approach [7] is introduced as a possible generic descriptor for manipulation actions, which encodes the sequence of spatio-temporal changes in relations between manipulated objects. But it only takes into account touching and not-touching relations and does not consider other spatial information, therefore it has limitations in action recognition, as well in its usability for guiding execution by a robot. Here we would like to extend the SEC framework by considering qualitative static and dynamic spatial relations between objects and make a novel more accurate framework for classification of manipulation actions based on symbolic spatial relations.

III. OUR APPROACH

A. Overview of our method

A brief description of the steps involved in our approach is provided in Fig.1 and the details will be discussed in the following sections.

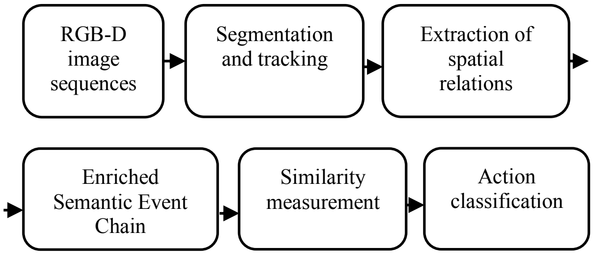

RGB-D image sequences
Segmentation and tracking
Extraction of spatial relations
Enriched Semantic Event Chain
Similarity measurement
Action classification

Fig. 1. Steps of our spatial reasoning approach

In order to semantically identify and compare manipulation actions, we present a new algorithm based on qualitative spatial relations. The input of our algorithm is an RGB-D video of a manipulation action. In this work, we use the videos of the MANIAC dataset which includes 8 different manipulation actions ***(Pushing, Hiding, Putting, Stirring, Cutting, Chopping, Taking, and Uncovering)*** [8].

A segmentation algorithm is applied on the scene at the first frame and objects are tracked during the rest of frames (section III-B). Spatial relations like "on top", "above", "below", are extracted as described in section III-C and so called Enriched Semantic Event Chains (ESEC) are defined in section III-D. Finally, our similarity measures and classification procedure is described in section III-E. The discriminative ability of the ESECs for different actions is evaluated in section IV. Results are compared to analogous results obtained using the original Semantic Event Chains (SECs) as presented in [8, 9].

B. Point cloud segmentation and tracking

As the first step, the recorded video frames are pre-processed by an image segmentation procedure based on color and depth information as described in [8]. In this procedure objects (and hands) in the scene are extracted as separate segments. A sample of a MANIAC dataset frame before and after segmentation is shown in Fig.2. Segments are tracked using a persistent super voxel world-model which is updated, rather than replaced, as new frames of data arrive as described in [26].

Each object in a scene after the aforementioned procedures is a point cloud, i.e., a set of points in a three-dimensional coordinate system (X, Y, Z). We define the scene at frame f as a set of point clouds: $\{\alpha_1^f,\dots,\alpha_N^f\}$, where N is the number of objects in the f_{th} frame of the action. Object α_i^f represents the point cloud of object i at frame f, $i \in\{1,\dots, N\}$ and can be tracked throughout the frames sequence.

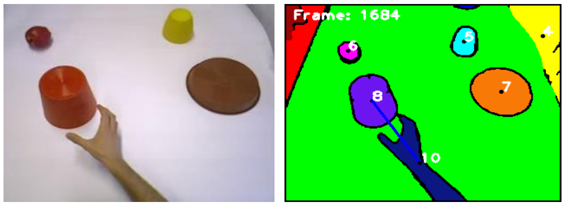


Fig. 2: A frame in MANIAC dataset (a) before and (b) after the scene segmentation. Segments are identified by different colors and segment numbers.

C. *Extraction of spatial object relations*

In this work, we define two types of spatial relations. The first type includes *static* relations which describe the directional ordering of objects in a scene and the second type contains *dynamic* relations between objects.

We define the following static spatial relations between objects in the scene: "Above" (**Ab**), "Below" (**Be**), "Right" (**R**), "Left" (**L**), "Front" (**F**), "Back" (**Ba**) and "Between" (**Bw**).

"**To**" and "**Bo**" explain top and bottom relations, respectively, which incorporate "Above" and "Below" with touching (**Ab** + **T** = **To**; **Be** + **T** = **Bo**). We gather all of these relations in a set and name it *Rel_static*. Thus, *Rel_static* = {**Ab**, **Be**, **R**, **L**, **F**, **Ba**, **Bw**, **To**, **Bo**}.

Dynamic relations are the second type of relations in the current study which are collected in a *Rel_dynamic* set. When an object starts moving and the distance between its central point and another object's central point decreases in a time interval they are "Getting Close" (**GC**) and when this distance increases, it means these two objects are "Moving Apart" (**MA**). We also observe "**MT** = Move Together" (here we mean only moving together when being in touching (**T**) relation), "**HT** = Halt Together" (touching but not moving) and "**S**=Stable" (non-touching (**N**), but keeping the same distance). Thus, *Rel_dynamic* = {**GC**, **MA**, **MT**, **HT**, **S**}. Note, the relations "Touching" (**T**) and "Non-touching" (**N**) making the backbone of the original Semantic Event Chain framework [7] are used in some of the definitions of our new relations (e.g. **To**, **Bo**, **MT**, **HT**, **S**) as described above.

Further we explain in more detail how the introduced relations are calculated in real scenes. The touching (**T**) and non-touching (**N**) relations are determined by applying the "kd-tree algorithm" on two point clouds [5] and evaluating occurrence (or non-occurrence) of collision between the point clouds.

For definition of the other relations we need to first introduce our object model. We define the coordinate axes according to the direction of the camera axes. Our coordinate system is shown in Fig.3. The z axis corresponds to perceived depth (front/back) direction, while the x and y axes define directions of right/left and above/below, respectively. Table 1 defines directions of six spatial relations in terms of the coordinate system axes.

For each point cloud (object) we create an Axis Aligned Bounding Box (AABB). In the AABB all sides are parallel to the directions of the coordinate system axes (Fig.3(b)).

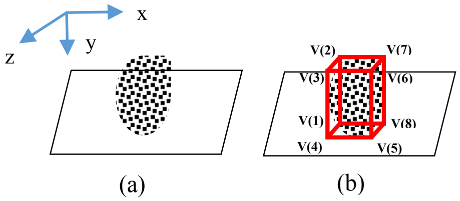


Fig.3. (a) Coordinate system, (b) A sample of AABB around a point cloud based in the defined coordinate system.

Suppose object α_i^f is the i_{th} object in the f_{th} frame represented as a point cloud and consisting of $P_{N\alpha i}$ points. As an object α_i^f model we define the AABB by the following set of vertices:

$$\begin{aligned}
V_i^f(1) &= [x_{min(i)}^f,\ y_{max(i)}^f,\ z_{min(i)}^f],\\
V_i^f(2) &= [x_{min(i)}^f,\ y_{min(i)}^f,\ z_{min(i)}^f],\\
V_i^f(3) &= [x_{min(i)}^f,\ y_{min(i)}^f,\ z_{max(i)}^f],\\
V_i^f(4) &= [x_{min(i)}^f,\ y_{max(i)}^f,\ z_{max(i)}^f],\\
V_i^f(5) &= [x_{max(i)}^f,\ y_{max(i)}^f,\ z_{max(i)}^f],\\
V_i^f(6) &= [x_{max(i)}^f,\ y_{min(i)}^f,\ z_{max(i)}^f],\\
V_i^f(7) &= [x_{max(i)}^f,\ y_{min(i)}^f,\ z_{min(i)}^f],\\
V_i^f(8) &= [x_{max(i)}^f,\ y_{max(i)}^f,\ z_{min(i)}^f].
\end{aligned}$$

where $x_{min(i)}^f$, $x_{max(i)}^f$, $y_{min(i)}^f$, $y_{max(i)}^f$, $z_{min(i)}^f$ and $z_{max(j)}^f$ are the minimum and maximum values between the points of object α_i^f in the x, y and z axes, respectively. We calculate spatial relations only for objects which are "neighbors" in the scene where the neighborhood is defined in the following way: suppose O_i^f shows the central point of the AABB of object α_i^f; we define $\Omega\ (\alpha_i^f, \alpha_j^f) = \|O_i^{\ f} - O_j^f\|$ to be a two argument function for measuring the Euclidean distance between the objects α_i and α_j in f_{th} frame. Objects are considered to be neighbors in case $\Omega\ (\alpha_i^f, \alpha_j^f) \leq$ Ʈ. In this study we define a threshold Ʈ of 1 m, which makes most of the objects in our table-top manipulation neighbors (only extremely distant objects, e.g. those that are beyond the table are excluded).

Each relation is defined by a set of rules and those rules are evaluated for each neighboring object pair. We start with specifying the rules set for static spatial relations. Let us consider the relation "Right": *SR* (α_i^f, α_j^f) = **R** (object α_i is to the right of object α_j in frame f) **if** $x_{max}(\alpha_i^f) > x_{max}(\alpha_j^f)$ as well as all the following (exception) conditions are *not* true:

$y_{min}(\alpha_i^f) > y_{max}(\alpha_j^f)$; $y_{min}(\alpha_j^f) > y_{max}(\alpha_i^f)$; $z_{min}(\alpha_i^f) > z_{max}(\alpha_j^f)$; $z_{min}(\alpha_j^f) > z_{max}(\alpha_i^f)$. The exception conditions exclude from the relation "Right" those cases when two object-AABBs do not overlap in altitude (y direction) or front/back (z direction). Several examples of objects holding relation *SR* (red, blue) = **R**, when the size and shift in y direction varies, are shown in in Fig. 4.

SR (α_i^f, α_j^f) = **L** is defined by $x_{min}(\alpha_i^f) < x_{min}(\alpha_j^f)$ and the same set of exception conditions. The relations "**Ab**", "**Be**", "**F**", "**Ba**" are defined in an analogous way. For "**Ab**" and "**Be**" the emphasis is on the "y" dimension, while for the **F**", "**Ba**" the emphasis is on the "z" dimension.

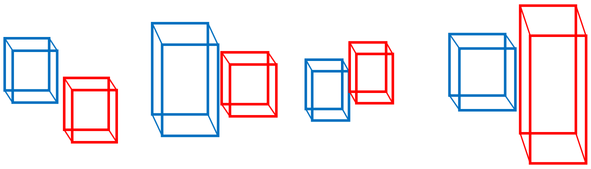


Fig.4. Possible states of Right- Left relations between two AABBs when size and y positions vary.

Next we will define the "**Bw**" (Between) relation (see Fig 5). First we define so called "Between space" for two objects. This space is obtained by extending the AABBs of two non-overlapping objects towards each other along the pre-defined axis and taking the intersection of those extensions. Whenever the third object's AABB completely stays in the "Between space" of the two other objects', it is assumed that the third object is in "Between" (**Bw**) of the two objects. The rules for this relation in the case the "Between space" is on the *X* axis is defined below (the object α_3^f is in between of objects α_1^f and α_2^f):

$$SR\,(\alpha_1^f, \alpha_2^f, \alpha_3^f) = \mathbf{Bw},$$
$$\textbf{If } (x_{min(3)}{}^f > \text{maximum}\,(x_{min(1)}{}^f, x_{min(2)}{}^f)\ \&\&\ (x_{max(3)}{}^f < \text{minimum}\,(x_{max(1)}{}^f, x_{max(2)}{}^f))$$
$$\textbf{If } (y_{min(3)}{}^f > \text{minimum}\,(y_{min(1)}{}^f, y_{min(2)}{}^f)\ \&\&\ (y_{max(3)}{}^f < \text{maximum}\,(y_{max(1)}{}^f, y_{max(2)}{}^f))$$
$$\textbf{If } (z_{min(3)}{}^f > \text{minimum}\,(z_{min(1)}{}^f, z_{min(2)}{}^f)\ \&\&\ (z_{max(3)}{}^f < \text{maximum}\,(z_{max(1)}{}^f, z_{max(2)}{}^f))$$

Two objects can have more than one static spatial relation regarding each other: e.g. one object's AABB can be both to the right and in front of other object's AABB. However, for forming the ESEC (as will be explained in III-D) we need only one relation per object pair. Here we propose a solution for this problem.

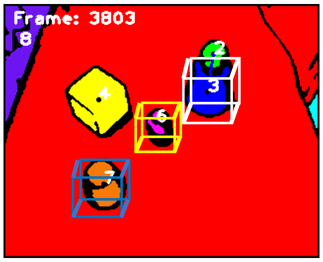


Fig.5. Defining betweenness by AABBs. In this scene, yellow AABB is between white and blue AABBs.

Each AABB is a cube with 6 rectangles. Let us label them as top, bottom, right, left, front and behind based on their positions in our scene coordinate system. Whenever object α_i is in the right of object α_j, one can make a projection from the left rectangle of object α_i onto the right rectangle of object α_j and consider only the rectangle intersection area which we will call "*shadow*" in this work.

Suppose *SR* $(\alpha_i^f, \alpha_j^f) = \{\gamma_1,\ldots,\gamma_k\}$ while $\{\gamma_1,\ldots,\gamma_k\} \in$ *Rel_static* and we have calculated *shadow* $(\alpha_i^f, \alpha_j^f, \gamma)$ for all relations γ between the objects α_i^f and α_j^f. The relation with the biggest *shadow* is chosen as the main static relation for the two objects:

$$SR\,(\alpha_i^f, \alpha_j^f) = \gamma_n\,,\ \textbf{If}\ shadow\,(\alpha_i^f, \alpha_j^f, \gamma_n) = \max_{1 \le m \le k}(Shadow\,(\alpha_i^f, \alpha_j^f, \gamma_m)).$$

The static relations around objects are highly dependent on the viewpoint and their changes, also do not make a human-notable difference in the performance of manipulation actions. For instance, when picking a knife to cut a cucumber we do not note if the knife is picked from the right or the left side of the cucumber. Thus we define a new relation called "Around" (**Ar**) and map the set of relations {**L**, **R**, **F**, **Ba**} onto it. In fact, "Ar" (Around) includes the space located on lateral sides of an object in a limited radius equal to threshold Ʈ. This space does not cover the vertical neighborhood areas like "Above" or "Below".

TABLE 1: Definition of spatial relation directions

Directions	Right	Left	Front	Back	Above	Below
Relevant vector	+x	-x	+z	-z	-y	+y

Now we switch to explaining dynamic relations *DR* which we define as a two argument function where arguments are objects in the scene. When the distance between two objects' AABB decreases during a time segment (let us say within ϴ frames; we have used ϴ=10 in our experiments, given the 30 fps recording), they are "Getting Close" (**GC**) and when this distance increases, these two objects are "Moving Apart" (**MA**). Formal definition is given next, where the threshold τ is kept at 0.1 m:

$$\underset{(i \neq j)}{DR\,(\alpha_i^f, \alpha_j^f):}\ \begin{cases} \textbf{GC}, & \text{if: } \Omega\,(\alpha_i^{f+\Theta}, \alpha_j^{f+\Theta}) - \Omega\,(\alpha_i^f, \alpha_j^f) < \tau \\ \textbf{MA}, & \text{if: } \Omega\,(\alpha_i^{f+\Theta}, \alpha_j^{f+\Theta}) - \Omega\,(\alpha_i^f, \alpha_j^f) > \tau \end{cases}$$

When calculating **GC** and **MA** we are also checking the touching relations $SR_{touch}\,(\alpha_i^f, \alpha_j^f)$ = (**T** or **N**) between the two objects. Based on SR_{touch}, we define two conditions required for calculating the remaining dynamic relations:

P1: *Rel_touch* (α_i^f, α_j^f) =**T &&** *Rel_touch* $(\alpha_i^{f+\theta}, \alpha_j^{f+\theta})$ =**T**

P2: *Rel_touch* (α_i^f, α_j^f) =**N &&** *Rel_touch* $(\alpha_i^{f+\theta}, \alpha_j^{f+\theta})$ =**N**

The third condition is on object α_i, α_j movement:

P3: $O_i^{\,f} \neq O_i^{\,f+\Theta}$ **&&** $O_j^{\,f} \neq O_j^{\,f+\Theta}$

The dynamic relations **MT**, **HT** and **S**, based on the three conditions above are defined in the following way:

$$DR(\alpha_i^f, \alpha_j^f) \; (i \neq j) \begin{cases} \textbf{MT}, & \text{if: } P1 \text{ and P3} \\ \textbf{HT}, & \text{if: } P1 \text{ and} \sim \text{P3} \\ \textbf{S}, & \text{if: } P2 \text{ and } \Omega(\alpha_i^{f+\Theta}, \alpha_j^{f+\Theta}) - \Omega(\alpha_i^f, \alpha_j^f) \\ & < \tau \end{cases}$$

D. Enriched SEC framework (ESEC)

As mentioned in the introduction, the Enriched SEC framework is inspired by the original Semantic Event chain (SEC) approach [4]. The original SECs check touching (T), not-touching (N) and absence (A) relations between each pair of objects in all frames of a manipulation scene and focus on transitions (change) in these relations. The extracted sequences of relational changes (represented in a form of a matrix, see first matrix in Fig. 6) are used in the manipulation action recognition. In the Enriched SEC framework the wealth of relations described in section III-C are embedded into a similar matrix-form representation, showing how the set of relations changes throughout the action. We expect to be able to differentiate actions in more details this way.

As the first step of making an Enriched SEC, we recognize so called "fundamental objects" among all of the other objects in a manipulation scene. Definition of these objects are based on the original SEC relations and given in Table 2. This way we exclude distractor objects which are present in the scene but do not perform any role in the manipulation.

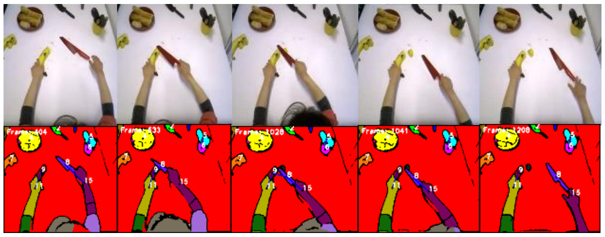
TABLE 2. Definition of fundamental objects during manipulation action

Object	Definition	Relation
Hand	The object that performs the action	Not touching anything at the beginning and at the end of the action. It touches at least one object
Main	The object which is directly in contact with the hand	Not touching the hand at the beginning and at the end of the action. It touches the hand at least once
Primary	The object from which the main object separates	Initially touches the main object. Changes its relation to not touching during the action
Secondary	The object to which the main object joins	Initially does not touch the main object. Changes its relation to touching during the action

As ESEC representation we introduce two matrices: one for representing the sequence of the static spatial relations *Rel_static* between the fundamental manipulated objects and one for describing the sequence of dynamic relations *Rel_dynamic* between the objects. We calculate static and dynamic relations in the sequence of the video frames of a manipulation action and add a new column to both (static and dynamic relation) matrixes whenever any static or dynamic relation has changed. This way we obtain a notation in matrix form as shown in Fig. 7 (middle is the static relation matrix and bottom is the dynamic relation matrix).

Alternatively, we can interpret our matrixes as sequences of graphs, where fundamental objects are connected by edges with the labels of static and dynamic relations. Each column in each matrix represents one graph, and the sequence of columns shows the time-development of those graphs.

One can observe (compare top representation in Fig. 6 for the original SEC with the bottom representation for the ESEC), that the ESEC has more columns as compared to the original SEC.



$$\begin{array}{c} \textbf{H, M} \\ \textbf{P, S} \\ \textbf{M, S} \end{array} \begin{pmatrix} A & N & T & T & T & T & N & N & A \\ T & T & T & T & T & T & T & T & T \\ N & N & N & T & N & N & N & N & N \end{pmatrix}$$

$$\begin{array}{c} \textbf{H, M} \\ \textbf{P, S} \\ \textbf{M, S} \end{array} \begin{pmatrix} A & Ar & Ab & To & To & To & To & To & Ab & Ar & A & A \\ To & To & To & To & To & To & To & To & To & To & To & To \\ Ar & Ar & Ar & Ar & Ar & Ab & To & Ab & Ar & Ar & Ar & Ar \end{pmatrix}$$

$$\begin{array}{c} \textbf{H, M} \\ \textbf{P, S} \\ \textbf{M, S} \end{array} \begin{pmatrix} A & GC & GC & GC & MT & MT & MT & MT & MA & MA & MA & A \\ HT & HT & HT & HT & HT & HT & HT & HT & HT & HT & HT & HT \\ S & S & S & S & GC & GC & S & MA & MA & MA & S & S \end{pmatrix}$$

Fig.6. Description of a cutting action in SEC and Enriched SEC frameworks. First row: frames from the manipulation video for visualization of the action; second row: segmentation of the frames above, third row: SEC matrix, fourth row: ESEC: Static relation matrix; fifth row: ESEC: dynamic spatial relation matrix; knife is the main object (M), table is the primary object (P), cucumber is the secondary object (S).

E. Similarity measures and classification procedure

For calculating similarity "*sim"* between two ESECs we use a measure based on Longest Common Subsequence (LCS) as described in [8]. For discriminating different actions, we define, a threshold (Ɛ) according to the minimum similarity value between the ESEC matrices of the same manipulation actions in real data: $\mathcal{E} = \mathrm{Min}_k (\mathrm{Min}_q (sim\ (A_{kq}, A_{kq})))$. Here A_k is a representative of a manipulation type (e.g. in the MANIAC data set we are using in further experiments) and A_{kq} indicates the q-th scenario of that action, in the dataset. For action classification we follow the online procedure defined in [8] in a slightly simplified way (see pseudocode in Fig. 7).

```
For (1 ≤ i ≤ 120)
  If (i=1)
     Make "cluster one" and assign ESEC_i to "cluster one"
  Else
     For (1 ≤ j ≤ Number of existing clusters)
        For (1≤ k ≤ Number of members in cluster j)
           S_ijk = sim (ESEC_i, m_jk) // calculate similarity to cluster j
member k
        S_i = max ({S_ijk})   // find maximum
        J = arg(max(S_ijk)) // find to which cluster maximum belongs
        If (S_i> Ɛ)
           Assign ESEC_i to cluster J
        Else
        Create new cluster and assign ESEC_i to the new cluster
```

Fig.7. Pseudocode for ESEC clustering.

We take ESECs extracted for each dataset video in a random order. The first ESEC is assigned to cluster one. For the second randomly selected ESEC we calculate the similarity *sim* to the first ESEC. If the similarity is above the threshold Ɛ, we assign the ESEC to the same cluster. Otherwise, we assign the ESEC to a new cluster. When more than one ESEC is already assigned to some cluster, we calculate the maximum similarity between the cluster members and the new ESEC. In case more than one cluster show above-threshold similarity, the ESEC is assigned to the cluster with the highest similarity. The procedure is continued until the dataset is exhausted. Afterwards class labels are assigned to clusters using the ground-truth labels, according to the majority in that cluster and the classification error is calculated in comparison to the ground-truth labels.

IV. EXPERIMENTS

A. Data Sets

Our action classification experiments were performed on the MANIAC dataset. It includes 8 different manipulation actions (Pushing, Hiding, Putting, Stirring, Cutting, Chopping, Taking, and Uncovering), each of which is presented in 15 different versions performed by 5 different human actors (overall 120 demonstrations). Actors were performing actions in different order, choosing from a set of 30 different objects and performing in differently configured scenes. Manipulation instances of each action have big variations in terms of manipulated objects, their poses, and followed trajectories.

To address a wider action variety, we have conducted additional experiments on a 26 actions set presented in [9]. Here, however, we did not have data recordings and thus were working on hand-made action models following the methodology suggested in [9].

B. Spatial relations accuracy

Here we begin with a brief evaluation of the performance of our spatial relation model. We asked three persons to indicate static spatial relations between pairs of objects from the set *Rel_static* on a set of 120 selected scenes in the MANIAC dataset. We have accepted the human-labeled relations to be the ground-truth in those cases where a majority vote was possible (2 matching human evaluations). We then calculated the relations using the algorithms introduced in III-C, including the extraction of the main relation, in case several relations were true, and compared with the ground truth. The obtained false positive rate is **FPR=4.725%** and the obtained false negative rate is **FNR=5.262%**.

C. Action Classification

We performed action classification on the MANIAC dataset as described in section III-E. The threshold Ɛ used for action discrimination for the MANIAC dataset is Ɛ=57%.

Table 3 compares action classification results of our novel ESEC representation to the results of the SEC framework as indicated in [8]. The classification accuracy for all actions is higher in ESECs. Totally, in average the spatial reasoning method has 97% accuracy in action classification which makes 12% improvement in compare of the previous method. This supports the notion that ESEC is a more powerful tool for classifying manipulation actions, as compared to the original SEC approach.

D. Discriminative ability of the Enriched SEC framework in an extensive actions set

A manipulation action ontology was designed in [9] where the hierarchical relations of 26 single-hand manipulation actions were based on the SEC framework (as well as pose and velocity considerations). However, it was shown that the discriminative ability of SECs alone is not enough to differentiate all those actions from each other. Here we will take the 26 manipulation actions analyzed in [26] and measure how much the discriminative ability increases when we use the ESECs for that purpose.

TABLE 3. Accuracy of classification on the MANIAC dataset in ESEC and SEC frameworks

Actions	ESEC	SEC[8]
Hiding	100%	87%
Pushing	94%	93%
Putting	100%	87%
Stirring	93%	93%
Cutting	91%	80%
Chopping	100%	93%
Taking	95%	87%
Uncovering	100%	80%
Average	**97%**	**85%**

The study [9] divides the 26 manipulation actions into six groups, where within one group all actions are similar or identical based on the SEC representation. Actions can be differentiated with SECs only *across* groups. Different from this, here we show how the Enriched SECs can now also differentiate actions *within* each group. Two groups are analyzed in Tables 4 and 5. To allow for fair comparison, we use as discrimination threshold 65% as this had been used in [8, 9]. As a consequence, in Tables 4 we see an action group where the ESECs differentiate the same number of actions as the SECs. However, ESECs can observe *sub-threshold differences* between the first three actions in the group, while the SECs indicate those actions as fully identical (similarity 100%).

In Table 5 we see an action group where ESECs can differentiate an additional action. Actions "cut" and "scoop" are 100% identical in the SEC representation, while the ESECs can differentiate those (with only **41%** similarity). We also see sub-threshold improvement when differentiating "Cut" from "Scissor cut".

In addition, ESECs can differentiate actions "Put over" from "Push over" (**48%** similarity vs. 66% in SECs), "Break" from "Uncover by pick&place" (**18%** vs 69% in SECs), "Break" from "Uncover by pushing" (**19%** vs 69 in SECs), "Uncover by pick&place" from "Uncover by pushing" (**54%** vs. 67 in SECs).

V. DISCUSSION

In this paper, we have introduced a representation for manipulations and called the Enriched Semantic Event Chain, which focuses on spatial relations between objects in a scene. We divided possible spatial relations into "static" and "dynamic" ones. For each action, the sequences of these static and dynamic spatial relations create a semantic descriptor of the manipulation action. The obtained descriptors are used to discriminate between different actions using real video sequences from the MANIAC data set (8 different actions) as well as sequences from the 26 actions from [9].

TABLE 4. ESECs showing differences in actions, when SECs indicate those as 100% similar (identical). "Hit&more" action set includes: Hit, Flick, Poke, Rub and Bore actions. Similarity values allowing action differentiation are shown in bold font.

ESEC					
Actions	**Hit&more**	**Push**	**Pull**	**Stir**	**Knead**
Hit&more	100%	83%	83%	**29%**	**46%**
Push		100%	83%	**19%**	**22%**
Pull			100%	**29%**	**46%**
Stir				100%	**0%**
Knead					100%

SEC					
Actions	**Hit&more**	**Push**	**Pull**	**Stir**	**Knead**
Hit&more	100%	100%	100%	**30%**	**60%**
Push		100%	100%	**30%**	**60%**
Pull			100%	**30%**	**60%**
Stir				100%	**44%**
Knead					100%

TABLE 5. ESECs differentiating between additional pair of actions, as compared to SECs. Similarity values allowing action differentiation are shown in bold font.

ESEC				
Actions	**Cut**	**Scissor cut**	**Draw**	**Scoop**
Cut	100%	83%	**52%**	**41%**
Scissor cut		100%	**14%**	**21%**
Draw			100%	**12%**
Scoop				100%

SEC				
Actions	**Cut**	**Scissor cut**	**Draw**	**Scoop**
Cut	100%	100%	**63%**	**100%**
Scissor cut		100%	**63%**	**42%**
Draw			100%	**36%**
Scoop				100%

Action differentiation by ESECs is compared to our earlier method based only on touching and not-touching events encoded in the older SEC (Semantic Event Chain) framework [8]. Both frameworks do not require object recognition and they ignore movement trajectories. Because in the original SECs touching and not-touching are the only defined spatial relations, the discriminative power of SECs is more limited than that of the here proposed Enriched SECs. This is shown by the difference between 96.625% action recognition accuracy for ESECs as compared to 87.5% for SECs using MANIAC. Also for the data from [9] we find improved performance and 5 more actions can be discriminated with ESECs. In addition, we found that several actions that had been 100% similar using the SEC framework begin to show differences when using ESECs (e.g. 83% similarity only). All this clearly shows that ESECs have a higher discriminative power than SECs. Because of this ESEC are necessarily also more robust against noise during action observation.

Evidently, there are some actions that can only be distinguished when considering dynamics, too (e.g. push versus hit). Those are not covered by the (E)SEC frameworks. In our older works [8,9] we had argued for a level-based semantic understanding of manipulations, where (E)SECs represent one certain symbolic level of understanding which can be supplemented by "finer" sub-symbolic layers (such as differentiating actions on the grounds of their different movement characteristics). ESECs help this process, because – having more transitions than SECs – they are breaking down an action into more (symbolic) components. Suppose we want to put a cup on the top of a box. In the original SEC, the relation between cup and box is initially "not-touching" and later "touching". With an ESEC representation there are additional phases where the cup is "Getting close" or is "Above", etc. These phases are now quite fine-grained and this should allow defining and joining trajectories for each phase. As the ESEC framework describes the sequence of required object relations based on quantitatively measured object (and manipulator) positions, it is possible to use the entries in the columns of the ESECs to provide quantitative start and end points for the manipulator trajectory. We had designed such a procedure using the older SEC framework coupled to DMPs [27] for trajectory generation in [28, 29] and we can now do the same in an improved way using the finer-grained representation of ESECs instead of the SECs.

Acknowledgements. The research leading to these results has received funding from the DFG grant WO 388/13-1 and the EU Horizon 2020 research and innovation program under grant agreement No. 680431, ReconCell.

References

[1] Y. Yacoob, M. Black, "Parameterized modeling and recognition of activities," in *International Conference on Computer Vision*, 1998, pp.120-12.

[2] L. Lo Presti and M. La Cascia, "3D skeleton-based human action classification: A survey," *Pattern Recognition*, vol. 53, pp. 130–147, May 2016.

[3] B. Rosman and S. Ramamoorthy, "Learning spatial relationships between objects," *The International Journal of Robotics Research*, vol. 30, no. 11, pp. 1328–1342, Sep. 2011.

[4] J. Goguen, "Mathematical models of cognitive space and time," in *Reasoning and Cognition: Proceedings of the Interdisciplinary Conference on Reasoning and Cognition*, 2006, pp. 125–128.

[5] M. Aiello and B. Ottens, "The Mathematical Morpho-logical View on Reasoning About Space," in *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, San Francisco, CA, USA, 2007, pp. 205–211.

[6] K. Zampogiannis, Y. Yang, C. Fermüller, and Y. Aloimonos, "Learning the spatial semantics of manipulation actions through preposition grounding," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 1389–1396.

[7] 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, Sep. 2011.

[8] 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, Sep. 2015.

[9] 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 Transactions on Autonomous Mental Development*, vol. 5, no. 2, pp. 117–134, Jun. 2013.

[10] T. M. Crockett, M. W. Powell, and K. S. Shams, "Spatial planning for robotics operations," in *2009 IEEE Aerospace conference*, 2009, pp. 1–7.

[11] J. A. Park, Y. S. Kim, and J. Y. Cho, "Visual reasoning as a critical attribute in design creativity," in *Proceedings of International Design Research Symposium*, 2006.

[12] M. Contero, F. Naya, P. Company, and J. L. Saorín, "Learning Support Tools for Developing Spatial Abilities in Engineering Design," *International Journal of Engineering Education*, vol. 22, no. 3, pp. 470–477, Jun. 2006.

[13] H. Schultheis, S. Bertel, and T. Barkowsky, "Modeling Mental Spatial Reasoning About Cardinal Directions," *Cognitive Science*, vol. 38, no. 8, pp. 1521–1561, Nov. 2014.

[14] S. Eagleson, F. Escobar, and I. Williamson, "Hierarchical spatial reasoning theory and GIS technology applied to the automated delineation of administrative boundaries," *Computers, Environment and Urban Systems*, vol. 26, no. 2–3, pp. 185–200, Mar. 2002.

[15] Y. Wei, E. Brunskill, T. Kollar, and N. Roy, "Where to go: Interpreting natural directions using global inference," in *IEEE International Conference on Robotics and Automation, 2009. ICRA '09*, 2009, pp. 3761–3767.

[16] B. Bredeweg and P. Struss, "Current topics in qualitative reasoning," *AI Magazine*, vol. 24, no. 4, p. 13, 2003.

[17] J. Renz and B. Nebel, "Qualitative Spatial Reasoning Using Constraint Calculi," in *Handbook of Spatial Logics*, M. Aiello, I. Pratt-Hartmann, and J. V. Benthem, Eds. Springer Netherlands, 2007, pp. 161–215.

[18] 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)*, 2008, pp. 606–610.

[19] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *The International Journal of Robotics Research*, vol. 32, no. 8, pp. 951–970, Jul. 2013.

[20] G. Gemignani, R. Capobianco, and D. Nardi, "Approaching Qualitative Spatial Reasoning About Distances and Directions in Robotics," in *AI*IA 2015 Advances in Artificial Intelligence*, M. Gavanelli, E. Lamma, and F. Riguzzi, Eds. Springer International Publishing, 2015, pp. 452–464.

[21] H. Kjellström, J. Romero, D. Martínez, and D. Kragić, "Simultaneous Visual Recognition of Manipulation Actions and Manipulated Objects," in *Computer Vision – ECCV 2008*, D. Forsyth, P. Torr, and A. Zisserman, Eds. Springer Berlin Heidelberg, 2008, pp. 336–349.

[22] K. Nagahama, K. Yamazaki, K. Okada, and M. Inaba, "Manipulation of multiple objects in close proximity based on visual hierarchical relationships," in *2013 IEEE International Conference on Robotics and Automation (ICRA)*, 2013, pp. 1303–1310.

[23] K. Ramirez-Amaro, E.-S. Kim, J. Kim, B.-T. Zhang, M. Beetz, and G. Cheng, "Enhancing human action recognition through spatio-temporal feature learning and semantic rules," in *2013 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2013, pp. 456–461.

[24] 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)*, 2013, vol. 2.

[25] D. R. Faria, R. Martins, J. Lobo, and J. Dias, "Extracting data from human manipulation of objects towards improving autonomous robotic grasping," *Robotics and Autonomous Systems*, vol. 60, no. 3, pp. 396–410, Mar. 2012.

[26] J. Papon, A. Abramov, M. Schoeler, and F. Wörgötter, "Voxel Cloud Connectivity Segmentation Supervoxels for Point Clouds,"in *2013 IEEE conference on computer vision and pattern recognition,* 2013, pp. 2027–2034.

[27] T. Kulvicius, K. Ning, M. Tamosiunaite and F. Worgötter, "Joining Movement Sequences: Modified Dynamic Movement Primitives for Robotics Applications Exemplified on Handwriting," in *IEEE Transactions on Robotics*, vol. 28, no. 1, pp. 145-157, Feb. 2012.

[28] M J. Aein, E E. Aksoy, M. Tamosiunaite, J. Papon, A. Ude and Wörgötter, F. "Toward a library of manipulation actions based on Semantic Object-Action Relations," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems,* 2013, pp. 4555 - 4562 .

[29] M J. Aein, E E. Aksoy, F. Wörgötter, "Library of Actions: Implementing a Generic Robot Execution Framework by Using Manipulation Action Semantics", Submitted to *International Journal of Robotic Research (IJRR)*.