# A Summarized Semantic Structure to Represent Manipulation Actions

Tobias Strübing, Fatemeh Ziaeetabar and Florentin Wörgötter[a]
*Göttingen University, Institute for Physics 3 - Biophysics and Bernstein Center for Computational Neuroscience,*
*Friedrich-Hund-Platz 1, 37077 Göttingen, Germany*

Keywords: Enriched Semantic Event Chain, Statistical Analysis, Spatial Relations, Semantic Action Representation.

Abstract: To represent human manipulation actions in a simple and understandable way, we had proposed a framework called enriched semantic event chains (eSEC) which creates a temporal sequence of static and dynamic spatial relations between objects in a manipulation. The eSEC framework has so far only been used in manipulation actions consisting of one hand. As the eSECs descriptors are in the form of huge matrices, we need to have a concise version of them. Here, we want to extend this framework to interactions which involve more hands. Therefore, we applied statistical and semantic analyses to summarize the current eSEC while preserving its important features and introducing an enhanced eSEC (e$^2$SEC). This summarization is done by reducing the number of rows in an eSEC matrix and merging semantic spatial relations between manipulated objects. Eventually, we presented the new e$^2$SEC framework which has 20% fewer rows, 16.7% less static spatial and 11.1% less dynamic spatial relations while still maintaining the eSEC efficiency in recognition and differentiation of manipulation actions. This simplification paves the way for a simpler recognition and predicting complex actions and interactions in a shorter time and is beneficial in real time applications such as human-robot interactions.

## 1 INTRODUCTION

Thanks to the experience of many training samples in their lives, humans can recognize and even predict actions well. However, this task is very difficult for a machine. A machine knows neither the objects that are used for an action, nor the reason and outcome of the action. Even simple actions like cut, stir or push are difficult to understand. Unless the machine is equipped with an efficient algorithm to represent and recognize human actions, many important challenges such as human computer interactions, visual surveillance and video indexing will remain unattainable. Therefore, various methodologies have recently been developed to resolve this issue. The majority of these proposed approaches, recognize human actions through statistical, syntactic or semantic analyses. Recently, We have introduced a semantic framework which represents manipulation actions (as an important category of human actions) in terms of thirty-row matrices whose rows indicate the chains of static and dynamic spatial relations between each pair of fundamental manipulated objects during the video frame sequences (Ziaeetabar et al., 2018; Ziaeetabar et al., 2017; Wörgötter et al., 2020). This approach extracts qualitative spatio-temporal relations between objects without further knowledge of their type which makes it superior to other approaches.

Although eSEC is a useful method for simple one-handed manipulation actions, its efficiency decreases as the number of involved hands increases or the actions become more complex. It also fails to represent the interactions effectively. On the other hand, we intend to extend our work by integrating the eSEC framework with the basic components of body limbs (Borràs et al., 2017) and create a conjoint framework to represent whole body human actions. But as long as our matrices are so huge (with thirty rows and many columns), we will not succeed in combining other features of body which inevitably lead to larger matrices. Because the huge matrices increase the complexity of the computations and slow down the performance of the algorithm.

Therefore, in this paper an enhanced version of ESEC (e$^2$SEC) is proposed to keep the eSEC matrices smaller and simpler with almost the same amount of information. The procedure is as follows: (a) computing the level of importance for each row in an eSEC matrix by an extensive statistical analysis to remove the less important rows as well as (b) shrinking the

[a] https://orcid.org/0000-0001-8206-9738

Strübing, T., Ziaeetabar, F. and Wörgötter, F.
A Summarized Semantic Structure to Represent Manipulation Actions.
DOI: 10.5220/0010258803700379
In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 4: VISAPP*, pages 370-379
ISBN: 978-989-758-488-6

set of spatio-temporal relations (semantics) to keep the new framework even simpler.

## 2 RELATED WORKS

In this paper, we applied spatial reasoning between objects to represent manipulation actions performed by humans. This type of reasoning has been previously presented in numerous other domains, including robot planning and navigation (Crockett et al., 2009), interpreting visual inputs (Park et al., 2006), computer aided design (Contero et al., 2006) and natural language understanding (Wei et al., 2009).

To represent manipulation actions semantically, various methodologies have been proposed. (Qi et al., 2019) used an attentive semantic recurrent neural network to understand individual actions and group activities in videos. To encode interactions between objects, (Sridhar et al., 2008) extracted functional object categories from spatio-temporal patterns. The next ability intelligent systems must be equipped with after representing actions is to be able to recognize them. Recently, (Khan et al., 2020) used deep neural networks, with features from a convolutional Neural Network model, and multiview features to recognize human actions. Other studies utilized RGB-D data to classify actions through a Bag-of-Visual-Words model (Avola et al., 2019; Fei-Fei and Perona, 2005), a multi-class Support Vector Machine classifier and a Naive Bayes Combination method (Kuncheva, 2004) to recognize human actions.

Among the existing methods, approaches that use a semantic perspective are more widely used, due to their simplicity in perception and similarity with the human cognitive system. In this regard, (Aksoy et al., 2011) introduced the semantic event chain (SEC) which considers the sequence of transitions between touch and non-touch relations between manipulated objects to represent and recognize actions. We further improved this method using a computational model, named the enriched Semantic Event Chain (eSEC) (Ziaeetabar et al., 2017), which incorporates the information of static (e.g. top, bottom) and dynamic spatial relations (e.g. moving apart, getting closer) between objects in an action scene. This led to a significant accuracy in recognition and prediction of manipulation actions (Ziaeetabar et al., 2018). The predictive power of humans and the eSEC framework was compared in (Wörgötter et al., 2020). Here, we intend to upgrade the current eSEC framework to cover other new and important applications of manipulation actions in every-day life.

This paper is organized into the following sections: First, we introduce the eSEC framework to continue with its enhanced version (e$^2$SEC) in 3.1. Then, the similarity measurement algorithm is proposed in 3.2. Next, the importance of rows in an eSEC matrix is computed in 3.3 and the updated semantics are presented in 3.4. The results are discussed following the methods section in 4 and finally, the paper is concluded by providing a conclusion and outlook to future work.

## 3 METHODS

### 3.1 eSEC

The eSEC framework has been completely introduced in our previous papers (Ziaeetabar et al., 2018; Ziaeetabar et al., 2017; Wörgötter et al., 2020). Here, we mention its basics.

The Enriched SEC framework is inspired by the original Semantic Event chain (SEC) approach (Aksoy et al., 2011) which check touching (T) and not-touching (N) relations between each pair of objects in all frames of a manipulation scene and focus on transitions (change) of these relations. The extracted sequences of relational changes which are represented in the form of a matrix will then used in the manipulation action recognition. In the enriched SEC framework the wealth of relations described below are embedded into a similar matrix-form representation, showing how the set of relations changes throughout the action.

A practical application would be human-robot interaction where a human performs an action while a robot observes it and performs the suitable response as soon as possible (Ziaeetabar et al., 2018).

#### 3.1.1 Spatial Relations

The details on how to calculate static and dynamic spatial relations have been provided in (Wörgötter et al., 2020). Here we only define these relations.

- Touching and non-touching relations (TNR) between two objects are defined according to collision or no-collision between their corresponding point clouds.

- Static spatial relations (SSR) include: “Above” (**Ab**), “Below” (**Be**), “Right” (**R**), “Left” (**L**), “Front” (**F**), “Back” (**Ba**), “Inside” (**In**), “Surround” (**Sa**). Since “Right”, “Left”, “Front” and “Back” depend on the viewpoint and directions

of the camera axes, we combined them into "Around" (**Ar**) and used it at times when one object was surrounded by another. Moreover, "Above" (**Ab**), "Below" (**Be**) and "Around" (**Ar**) relations in combination with "Touching" were converted to "Top" (**To**), "Bottom" (**Bo**) and "Touching Around" (**ArT**), respectively, which corresponded to the same cases with physical contact. If two objects were far from each other or did not have any of the above-mentioned relations, their static relation was considered as Null (**O**). This led to a set of nine static relations in the eSECs:
**SSR** = {Ab, Be, Ar, Top, Bo, ArT, In, Sa, O}
The additional relations, mentioned above: **R**, **L**, **F**, **Ba** are only used to define the relation Ar=around, because the former four relations are not view-point invariant.

- Dynamic Spatial Relations (DSR) require to make use of the frame history in the video. We used a history of 0.5 seconds, which is an estimate for the time that a human hand takes to change the relations between objects in manipulation actions. DSRs included the following relations: "Moving Together" (**MT**), "Halting Together" (**HT**), "Fixed-Moving Together" (**FMT**), "Getting Close" (**GC**), "Moving Apart" (**MA**) and "Stable" (**S**). MT, HT and FMT denote situations when two objects are touching each other while: both of them are moving in a same direction (MT), are motionless (HT), or when one object is fixed and does not move while the other one is moving on or across it (FMT). Case **S** denotes that any distance-change between objects remained below a defined threshold of ($\xi = 1$ cm) during the entire action. In addition, **Q** is used as a dynamic relation between two objects when their distance exceeded the defined threshold $\xi$ or if they did not have any of the above-defined dynamic relations. Therefore, dynamic relations make a set of seven members:
**DSR** = {MT, HT, FMT, GC, MA, S, Q}

To distinguish between touching/non-touching, we measure the distance between the closest points of two objects and set a touching relation if this distance is smaller than a predefined threshold ($\eta = 1$ cm). To facilitate the computation of spatial relations between objects, we use camera axes and create an Axis Aligned Bounding Box (AABB) surrounding each object's point cloud. In the AABB representation, all box sides are parallel to the directions of axes. An example of an object's point cloud and its corresponding AABB is shown in figure 1. By taking the AABBs around the objects' point clouds (instead of their original shape), the computation is much easier but also reliable.

### 3.1.2 Fundamental Object Roles

Computing the spatial relations described above between all pairs of objects is time consuming and useless. Therefore, we recognize the so-called "fundamental objects" among all of the other objects in a manipulation scene. The definition of these objects is based on the original SEC relations and given in Table 1. This way we exclude distractor objects which are present in the scene but do not perform any role in the manipulation.

Table 1: Definition and remarks of all objects in the eSEC framework. (Ziaeetabar et al., 2018).

| Object | Definition | Remarks |
|---|---|---|
| Hand | Hand interacts with the objects in the scene. | In the beginning and the end not touching anything. Interacts at least with one object during the manipulation. |
| Ground | The Support of all objects except the hand. | A ground plane extracted from a visual scene. |
| 1 | The first object which has a transition from N to T. | This object will have its first transition with hand. |
| 2 | The second object which has a transition from N to T. | Can have a change from N → T or from T → N. |
| 3 | The third object which has a transition from N to T. | Can have a change from N → T or from T → N. |

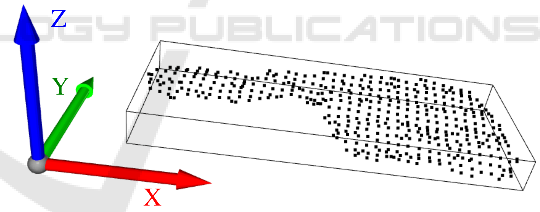


Figure 1: An AABB around a point cloud. The box is parallel to the axes x,y and z.

## 3.2 Similarity Measurement

The extracted sequences of spatial relational changes (produced in the form of a matrix, see the left matrix in figure 6) are used in the representation as well as recognition of manipulation actions. An eSEC matrix always consists of 30 rows while the top, middle and bottom 10 rows indicate the sequence of Touching and non-touching, static spatial and dynamic spatial relations between each pair of the fundamental manipulated objects during a manipulation contentious frames, respectively. Although the number of rows is constant, the number of columns varies depends on the number of frames. With the change of spatial relations between objects a new column is created.

Action representation by eSEC matrices allows us to measure the (diss)similarity between them through mathematical computations. To this end, each 30-row eSEC matrix is transformed to a 10-row matrix Θ that consists of triples containing (TNR, SSR, DSR) like seen in the following equation (Ziaeetabar et al., 2018):

$$\Theta = \begin{pmatrix} (a_{1,1}, a_{11,1}, a_{21,1}) & (a_{1,2}, a_{11,2}, a_{21,2}) & \cdots & (a_{1,n}, a_{11,n}, a_{21,n}) \\ (a_{2,1}, a_{12,1}, a_{22,1}) & (a_{2,2}, a_{12,2}, a_{22,2}) & \cdots & (a_{2,n}, a_{12,n}, a_{22,n}) \\ \vdots & \vdots & \ddots & \vdots \\ (a_{10,1}, a_{20,1}, a_{30,1}) & (a_{10,2}, a_{20,2}, a_{30,2}) & \cdots & (a_{10,n}, a_{20,n}, a_{30,n}) \end{pmatrix}.$$

With the three relations categories $\mathrm{L}^{1:3}$(Ziaeetabar et al., 2018)

$$L^1_{i,j} = \begin{cases} 0, & \text{if } \mathrm{a}_{i,j} = \mathrm{b}_{i,j} \\ 1, & \text{otherwise} \end{cases}$$

$$L^2_{i,j} = \begin{cases} 0, & \text{if } \mathrm{a}_{i+10,j} = \mathrm{b}_{i+10,j} \\ 1, & \text{otherwise} \end{cases}$$

$$L^3_{i,j} = \begin{cases} 0, & \text{if } \mathrm{a}_{i+20,j} = \mathrm{b}_{i+20,j} \\ 1, & \text{otherwise} \end{cases}$$

a compound difference $d_{i,j}$ can be defined:

$$d_{i,j} = \frac{\sqrt{L^1_{i,j} + L^2_{i,j} + L^3_{i,j}}}{\sqrt{3}}. \tag{1}$$

This leads to a compound matrix, which holds the difference values between each pair of the corresponding items in two eSEC matrices, and finally obtains the amount of dissimilarity between the two manipulation actions ($D_{\Theta_1,\Theta_2}$) (Ziaeetabar et al., 2018):

$$D_{(10,k)} = \begin{pmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,k} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ d_{10,1} & d_{10,2} & \cdots & d_{1,k} \end{pmatrix}$$

$$D_{\Theta_1,\Theta_2} = \frac{1}{k \cdot 10} \sum_{j=1}^{k} \sum_{i=1}^{10} d_{i,j}. \tag{2}$$

## 3.3 Importance of eSEC Matrix Rows

Some rows of an eSEC matrix depend on the importance of the objects that reflect the spatial relations between them contain more information than others. For instance, obviously the interaction between the hand and ground pair is less important than the one between hand and object 1 (which is the first object touched by hand and usually refers to a tool with the essential role in an action's execution).

According to our previous studies, we have defined and represented 35 manipulation actions by the eSEC framework (Ziaeetabar et al., 2018). To investigate the effectiveness and importance of a row, we considered all combinations of size 2 from our predefined 35 possible manipulation actions. Comparing the two matrices, when the number of columns is not equal, we repeated the last column in the matrix with fewer columns until the number of columns was the same. Then we initialized a counter to zero, and during the comparison, each member of the $i_{th}$ row ($1 \leq i \leq 30$) of each manipulation eSEC matrix compared with its corresponding member in the other matrix, and if they were not equal, the counter value was increased by one. This counter value was assigned to each row in any comparison. Next, the counter values of 30 rows were ranked from the lowest to highest. An example can be seen in table 2. Then we did this for each pair of the possible permutations between the predefined manipulations: $C(35,2) = 595$, and finally defined a parameter called the "degree of importance of the row" by calculating the mean (median) of the ranks (which were computed according to the counter values) for each row using all the permutation computations. An example of these comparison between action 1 to action 35: {(1,2), (1,3),...,(35,34)} is shown in table 3.

Table 2: An example of the calculation for the importance of rows. If one value in a row is same for manipulation 1 and manipulation 2, the dissimilarity counter rises by one. The rank is calculated with the percentages.

| Row | Manipulation 1 | | | | | | | Manipulation 2 | | | | | | | Dis. | % | Rank |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | U | U | T | T | T | N | N | U | U | T | T | N | N | N | 1 | 20 | 2 |
| · | · | · | · | · | · | · | · | · | · | · | · | · | · | · | 0 | 0 | 1 |
| 11 | U | U | Ar | ArT | ArT | Ar | O | U | U | Ab | To | To | Ab | O | 4 | 80 | 3 |
| · | · | · | · | · | · | · | · | · | · | · | · | · | · | · | 0 | 0 | 1 |
| 30 | U | U | U | U | U | U | U | U | U | U | U | U | U | U | 0 | 0 | 1 |

Table 3: An example for the calculation the mean and median of the ranks.

| Manipulations | Row 1 | · | Row 11 | · | Row 30 |
|---|---|---|---|---|---|
| 1,2 | 2 | · | 3 | · | 1 |
| · | · | · | · | · | · |
| 35,34 | 1 | · | 2 | · | 7 |
| Mean | 3.14 | · | 5.24 | · | 2.42 |
| Median | 3 | · | 5 | · | 1 |

The mean/median value is obtained at the end and is directly related to the "degree of importance of the rows". Because the row that produces the most distinction among all possible action permutations and causes more counter-value is of higher importance.

### 3.3.1 Removing Unimportant Rows

With the evaluation of "degree of importance of the rows", the less important rows can be deleted while all 35 predefined manipulations are still distinguishable from each other.

If there is more than one row that is less important according to 3.3, all possible combination of those must be considered to see if all 35 manipulations are still distinguishable after removing those rows or not. Therefore, we defined the number of rows that are least important according to the analysis explained in 3.3 as “$n$”. Initially, only one row is deleted ($k = 1$), then all possible combinations of two rows ($k = 2$), next three rows ($k = 3$) and so on, while every combination is considered. The resulting combinations are calculated with the binomial coefficient.

$$\binom{n}{k}_{1\leq k\leq n} = \frac{n!}{k!(n-k)!}. \qquad (3)$$

After considering all possible combination of rows to delete, we calculate the dissimilarity value for each pair of the predefined 35 manipulations, using equation 2. Given this results, we plot a huge dissimilarity matrix (size: 35×35) (figure 7), which displays the dissimilarity values between each pair of predefined manipulations after removing rows. In this way, $\binom{n}{k}_{1\leq k\leq n}$ dendrograms are produced. Despite the removal of the less importance rows, to make sure that the actions are still significantly different, we select the combination from which the most distinction between the existing manipulation actions is produced.

### 3.3.2 Dissimilarity Measure of Groups

We had categorized manipulation actions into 6 groups based on their nature in the figure 6 of (Ziaeetabar et al., 2018). To obtain more information about the “degree of importance of rows”, we introduced the dissimilarity measure of those groups. Therefore, groups were determined using an unsupervised clustering between different manipulation actions using their dissimilarity (Ziaeetabar et al., 2018). These groups can be seen in figure 7.

Using these groups, we calculated the dissimilarity between each member of one group with each member of another group, using equation 2. A calculation example can be seen in figure 2.

Finally, we calculated the minimum, maximum, mean and median for the dissimilarity of the groups which can be represented in a dissimilarity matrix. We select the rows that lead to the minimal information loss while removing.

## 3.4 Updated Semantics

So far, we summarized the eSEC manipulations descriptor by reducing the number of rows without compromising the uniqueness of the actions. To make the e$^2$SEC tables even simpler, we decided to shrink the huge set of static and dynamic spatial relations as well, and merge some of the current semantics. Our purpose is to ensure every manipulation is still distinguishable from each other while the set of spatial relations has been summarized. To this end, we combined the items that seemed most logical to integrate. For example, it is reasonable that the relations “Above” and “Below” can be merged in some way but “Inside” and “Around” have no relation to each other. Since there are more than one semantics we wanted to merge, we had to consider every possible combinations, using equation 3. For further analysis, we applied our merged semantics with the removed rows from chapter 3.3.1 & 3.3.2 and once more calculated the dissimilarity between the groups according to chapter 3.3.2.

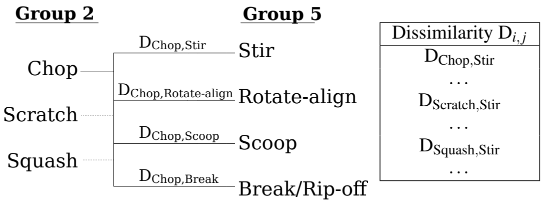


Figure 2: A dissimilarity value is calculated between each member of two groups (left). To determine the minimal information loss, we measured the mean, median, min and max values for all calculated dissimilarities between two groups (right).

Eventually, the maximum, minimum, mean and median values of the dissimilarities were computed and accordingly the merged semantics that leads to minimal information loss were selected.

## 4 RESULTS

In this paper, we divided the problem of simplification of the eSEC manipulation action descriptors into two parts.

- Determining the degree of importance for the rows and removing the less important ones.
- Integrating some spatio-temporal spatial relations with each other and shrinking the set of semantics.

To achieve the first purpose, we measured the importance of eSEC rows according to chapter 3.3, for every combination of the 35 predefine manipulation actions, which leads to C(35,2) = 595 calculations. To specify which rows can be removed with the least information loss, we plotted the median and mean values of the resulting ranks. As shown in figure 3, there are five rows with a lower rank and thus of less importance. These row numbers are 3, 4, 6, 8 and 10. Since

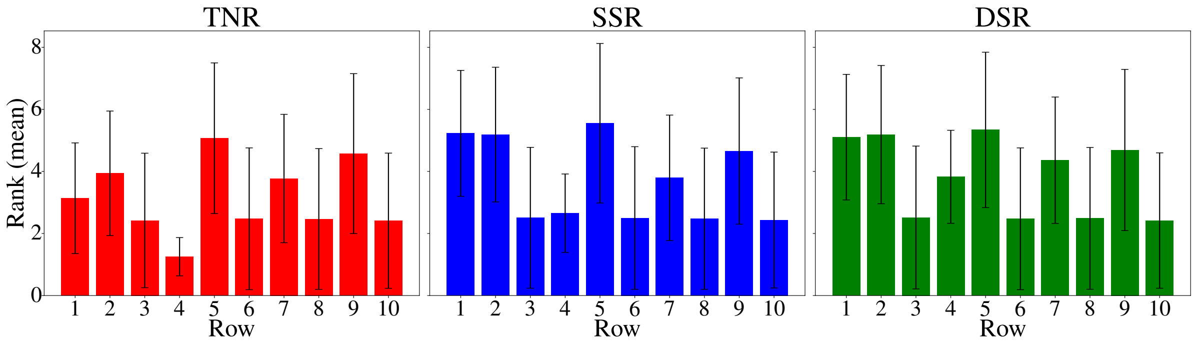


Figure 3: Mean values of the row ranks from TNR, SSR and DSR with corresponding error bars. A row is more important if the rank is high and vice versa.

we have $n = 5$ rows in total, to investigate further, we check the combination of one($k = 1$), two($k = 2$), three($k = 3$), four($k = 4$) and five($k = 5$) rows. Using equation 3, we reach a total of $5+10+10+5+1 = 31$ combinations to remove. We discovered that all pairs of manipulations are still distinguishable, even when all five rows were removen together. Therefore, we needed to calculate the dissimilarity between the groups of manipulations to find the combinations of the less important rows, by the removal of which, the different actions are still the most different from each other.

We started by removing all combinations of rows as mentioned before. Then, we calculated the dissimilarity between each group, leading to dissimilarity values of each group member $\{D_{i,j},\ldots,D_{n,m}\}$ while $i,\ldots,n \in$ group x and $j,\ldots,m \in$ group y, as shown in figure 2. We used these values to calculate the minimum, maximum, mean and median to plot them in a dissimilarity matrix, shown in figure 4.

We determined the best suitable combinations to remove by calculating the minimal cost between the original dissimilarity matrix ($A$) and the dissimilarity matrix with removed rows ($B$) by using the following equation:

$$\tfrac{1}{4}\left(|\bar{A}-\bar{B}|+|\tilde{A}-\tilde{B}|+|\max(A)-\max(B)|+|\min(A)-\min(B)|\right).$$

The variables $\bar{A}/\bar{B}$ are defined as the mean and $\tilde{A}/\tilde{B}$ as the median of the matrices. We discovered that removing rows 4 (relation between the hand and the ground) and 10 (relation between object 3 and the ground) result in a minimal information loss.

Once the special pairs of fundamental objects with less meaning are identified, we next integrate semantics to further simplify our e$^2$SEC framework.

In total, we discovered four semantics($n$=4) that were possible candidates to merge, shown in the following list:

- “Above”(**Ab**) & “Below”(**Be**) → “Vertical Around”(**VArT**)

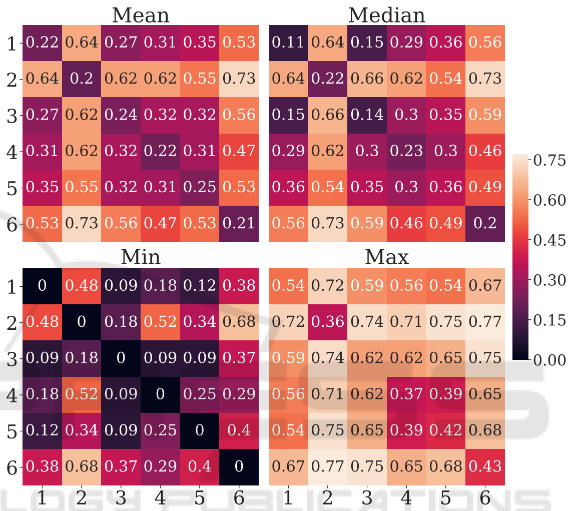


Figure 4: An example of the mean, median, min and max dissimilarity values of all groups with removed rows 4 and 10.

- “Top”(**To**) & “Bottom”(**Bo**) → “Vertical Around With Touch”(**VArT**)
- “Moving Together”(**MT**) & “Fixed-Moving Together”(**FMT**) → “Moving Together”(**MT**)
- “Getting Close”(**GC**) & “Moving Apart”(**MA**) → “Moving Around”(**MA**)

We once more check all combinations for one(k=1), two(k=2), three(k=3) and four(k=4) merged semantics. Using equation 3, we reach a total of $4+6+4+1 = 15$ combinations. In detail, we check all the combinations listed in table 4. We discovered that the minimal information loss is obtained by merging “MT+FM”, “Ab+Be” and “To+Bo”.

Furthermore, some semantics were renamed to remain consistent:

- “Around”(**Ar**) → “Horizontal Around”(**HAr**)
- “Around With Touch”(**ArT**) → “Horizontal Around With Touch”(**HArT**)

Table 4: All possible combinations of semantics for the analysis.

| Combination | Relations |
|---|---|
| One | To+Bo; |
| | Ab+Be; |
| | MA+GC; |
| | MT+FMT; |
| Two | Ab+Be, To+Bo; |
| | MA+GC, To+Bo; |
| | MA+GC, Ab+Be; |
| | MT+FMT, To+Bo; |
| | MT+FMT, Ab+Be; |
| | MA+GC, MT+FMT |
| Three | MA+GC, To+Bo, Ab+Be; |
| | MT+FMT, To+Bo, Ab+Be; |
| | MT+FMT, MA+GC, To+Bo; |
| | MT+FMT, MA+GC, Ab+Be; |
| Four | Ab+Be, To+Bo, MT+FMT, MA+GC |

In conclusion, we obtain the following new relations:

- **TNR**: {T, N, U, X}
- **SSR**: {VAr, In, Sa, Bw, HAr, VArT, HArT, U, X, O}
- **DSR**: {MT, HT, GC, MA, S, U, X, Q}

which can be observed in figure 5.

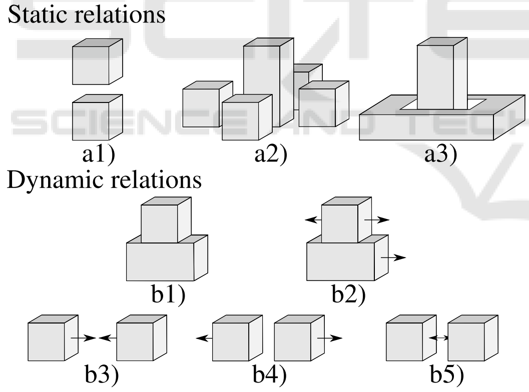


Figure 5: Final static spatial and dynamic spatial relations of the new e$^2$SEC framework.
a) Static Spatial Relations: a1) Vertical Around, a2) Horizontal Around, a3) Inside/Surround. b) Dynamic Spatial Relations: b1) Halting Together, b3) Moving Together, b4) Getting Close, b5) Moving Apart, b6) Stable.

At the appendix, we show an example to present the difference between the eSEC and e$^2$SEC matrices.

The figure 6 includes an eSEC and e$^2$SEC table of the "Scratch" manipulation. We Selected this action out of the 35 analyzed eSEC matrices because it benefits the most from the new framework. The images on top of the tables are the frames, which correspond to the process of this manipulation. First, the pencil (object 1) is touched by hand. After scratching on the paper, the pencil lead breaks (object 2) and separates from the remaining part of the pencil (object 3). The hand moves away from the lead and the pencil is put down by the hand. Finally, the hand moved out of the frame.

After applying our new e$^2$SEC approach, i.e., removing the rows containing spatial relation between (H,G) and (3,G) as well as merging the semantics, we are able to reduce the number of columns by $\approx 27\%$ in comparison to the eSEC table. Furthermore, the number of rows is reduced by 20%. Specifically, columns 1/2, 6/7 and 9/10 are equal with our framework and can therefore be removed. If we consider all 35 manipulation we get a mean of $\approx$ 12% reduced rows.

## 5 CONCLUSIONS AND OUTLOOK

In this paper, we improved our previously defined action representation framework, the so-called eSEC (enriched semantic event chain) and produced the enhanced version of it, called e$^2$SEC framework to represent human actions in a simple and concise way.

The traditional eSEC performed well in recognition and prediction of simple manipulations which were performed only by one hand (Ziaeetabar et al., 2018; Wörgötter et al., 2020) but was not efficient enough when we aimed to represent and recognize complex actions as well as interactions when two (or more) hands were involved. Because their eSEC tables were growing in size and the computations became heavier. This disadvantage was mostly considerable in real time applications such as prediction. The new e$^2$SEC simplifies the previous eSEC and provides a new possibility for the analysis of complex actions as well as the interactions that are most common in humans' every-day life.

In the e$^2$SEC framework, the number of rows was reduced to 20%. Moreover by merging the semantics in the set of spatial relations, we reduced the amount of static and dynamic spatial relations to 16.7% and 11.1%, respectively. This simplification allows us to combine manipulation descriptors with features of body limbs (Mandery et al., 2015) and create an integrated framework for full-body human action representation.

## REFERENCES

Aksoy, E., Abramov, A., Dörr, J., Ning, K., Dellen, B., and Wörgötter, F. (2011). Learning the semantics of object-action relations by observation. *I. J. Robotic Res.*

Avola, D., Bernardi, M., and Foresti, G. L. (2019). Fusing depth and colour information for human action recognition. *Multimedia Tools and Applications*.

Borràs, J., Mandery, C., and Asfour, T. (2017). A whole-body support pose taxonomy for multi-contact humanoid robot motions. *Science Robotics*.

Contero, M., Naya, F., Company, P., and Saorín, J. (2006). Learning support tools for developing spatial abilities in engineering design. *International Journal of Engineering Education*.

Crockett, T. M., Powell, M. W., and Shams, K. S. (2009). Spatial planning for robotics operations. In *2009 IEEE Aerospace conference*.

Fei-Fei, L. and Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*.

Khan, M., Javed, K., Saba, T., and Habib, U. (2020). Human action recognition using fusion of multiview and deep features: An application to video surveillance. *Multimedia Tools and Applications*.

Kuncheva, L. I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience.

Mandery, C., Terlemez, O., Do, M., Vahrenkamp, N., and Asfour, T. (2015). The kit whole-body human motion database. In *International Conference on Advanced Robotics (ICAR)*.

Park, J. A., Kim, Y. S., and Cho, J. Y. (2006). Visual reasoning as a critical attribute in design creativity. In *Proceedings of International Design Research Symposium*.

Qi, M., Wang, Y., Qin, J., Li, A., Luo, J., and Van Gool, L. (2019). stagnet: An attentive semantic rnn for group activity and individual action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*.

Sridhar, M., Cohn, A., and Hogg, D. C. (2008). Learning functional object-categories from a relational spatio-temporal representation. In *ECAI*.

Wei, Y., Brunskill, E., Kollar, T., and Roy, N. (2009). Where to go: Interpreting natural directions using global inference. In *2009 IEEE International Conference on Robotics and Automation*.

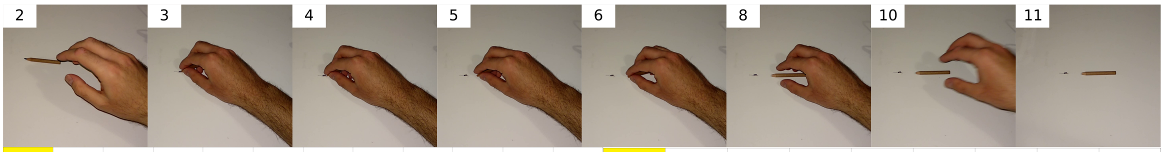
Wörgötter, F., Ziaeetabar, F., Pfeiffer, S., Kaya, O., Kulvicius, T., and Tamosiunaite, M. (2020). Humans predict action using grammar-like structures. *Scientific reports*.

Ziaeetabar, F., Aksoy, E. E., Wörgötter, F., and Tamosiunaite, M. (2017). Semantic analysis of manipulation actions using spatial relations. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*.

Ziaeetabar, F., Kulvicius, T., Tamosiunaite, M., and Wörgötter, F. (2018). Prediction of manipulation action classes using semantic spatial reasoning. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Ziaeetabar, F., Kulvicius, T., Tamosiunaite, M., and Wörgötter, F. (2018). Recognition and prediction of manipulation actions using enriched semantic event chains. *Robotics and Autonomous Systems*.

# APPENDIX



| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| H, 1 | U | U | T | X | X | X | X | X | X | X | X |
| H, 2 | U | U | U | T | N | N | N | N | N | N | N |
| H, 3 | U | U | U | T | T | T | T | N | N | N | N |
| H, G | N | N | N | N | N | N | N | N | N | N | N |
| 1, 2 | U | U | U | X | X | X | X | X | X | X | X |
| 1, 3 | U | U | U | X | X | X | X | X | X | X | X |
| 1, G | U | U | T | X | X | X | X | X | X | X | X |
| 2, 3 | U | U | U | N | N | N | N | N | N | N | N |
| 2, G | U | U | U | T | T | T | T | T | T | T | T |
| 3, G | U | U | U | N | N | N | T | T | T | T | T |
| H, 1 | U | U | To | X | X | X | X | X | X | X | X |
| H, 2 | U | U | U | To | Ar | Ar | Ar | Ar | Ar | Ar | O |
| H, 3 | U | U | U | ArT | ArT | ArT | ArT | Ar | Ab | Ab | O |
| H, G | O | Ab | Ab | Ab | Ab | Ab | Ab | Ab | Ab | Ab | O |
| 1, 2 | U | U | U | X | X | X | X | X | X | X | X |
| 1, 3 | U | U | U | X | X | X | X | X | X | X | X |
| 1, G | U | U | To | X | X | X | X | X | X | X | X |
| 2, 3 | U | U | U | Be | Ar | Ar | Ar | Ar | Ar | Ar | Ar |
| 2, G | U | U | U | To | To | To | To | To | To | To | To |
| 3, G | U | U | U | Ab | Ab | Ab | To | To | To | To | To |
| H, 1 | U | U | FMT | X | X | X | X | X | X | X | X |
| H, 2 | U | U | U | FMT | MA | S | S | MA | MA | MA | Q |
| H, 3 | U | U | U | MT | MT | MT | MT | MA | MA | MA | Q |
| H, G | Q | S | S | S | S | GC | S | S | S | MA | Q |
| 1, 2 | U | U | U | X | X | X | X | X | X | X | X |
| 1, 3 | U | U | U | X | X | X | X | X | X | X | X |
| 1, G | U | U | HT | X | X | X | X | X | X | X | X |
| 2, 3 | U | U | U | MA | MA | S | S | S | S | S | S |
| 2, G | U | U | U | HT | HT | HT | HT | HT | HT | HT | HT |
| 3, G | U | U | U | S | S | GC | HT | HT | HT | HT | HT |
| Frame | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |

eSEC

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| H, 1 | U | T | X | X | X | X | X | X |
| H, 2 | U | U | T | N | N | N | N | N |
| H, 3 | U | U | T | T | T | N | N | N |
| 1, 2 | U | U | X | X | X | X | X | X |
| 1, 3 | U | U | X | X | X | X | X | X |
| 1, G | U | T | X | X | X | X | X | X |
| 2, 3 | U | U | N | N | N | N | N | N |
| 2, G | U | U | T | T | T | T | T | T |
| H, 1 | U | VarT | X | X | X | X | X | X |
| H, 2 | U | U | VarT | Har | Har | Har | Har | O |
| H, 3 | U | U | HarT | HarT | HarT | Har | Var | O |
| 1, 2 | U | U | X | X | X | X | X | X |
| 1, 3 | U | U | X | X | X | X | X | X |
| 1, G | U | VarT | X | X | X | X | X | X |
| 2, 3 | U | U | Var | Har | Har | Har | Har | Har |
| 2, G | U | U | VarT | VarT | VarT | VarT | VarT | VarT |
| H, 1 | U | MT | X | X | X | X | X | X |
| H, 2 | U | U | MT | MA | S | MA | MA | Q |
| H, 3 | U | U | MT | MT | MT | MA | MA | Q |
| 1, 2 | U | U | X | X | X | X | X | X |
| 1, 3 | U | U | X | X | X | X | X | X |
| 1, G | U | HT | X | X | X | X | X | X |
| 2, 3 | U | U | MA | MA | S | S | S | S |
| 2, G | U | U | HT | HT | HT | HT | HT | HT |
| Frame | 1 | 3 | 4 | 5 | 6 | 8 | 9 | 11 |

e²SEC

Figure 6: Comparison of eSEC (left) and e$^2$SEC (right) matrices for the manipulation “Scratch”. The important frames are shown on top of the tables.

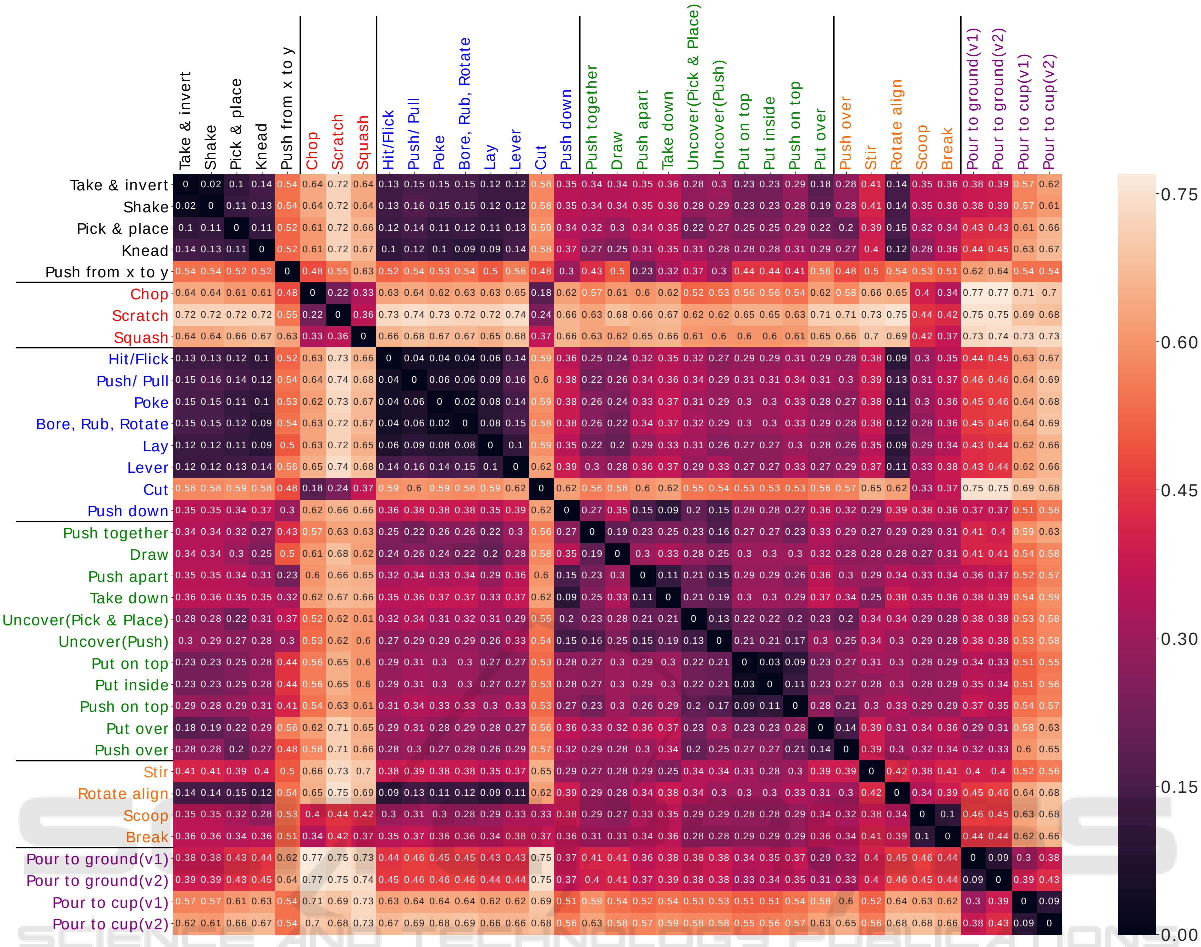


Figure 7: Dissimilarity matrix for all manipulations without removed rows. The color of the manipulation names represent the groups. Group 1: black, Group 2: red, Group 3: blue, Group 4: green, Group 5: orange, Group 6: purple