Who Can Predict Faster? Human or Robot?

Fatemeh Ziaeetabar[1], Stefan Pfeiffer[1], Minija Tamosiunaite[1,2], Tomas Kulvicius[1], Florentin Wörgötter[1]

[1] III. Physics Institute, University of Göttingen, Friedrich-Hund-Platz 1, 37077 Göttingen, Germany. fziaeetabar@gwdg.de
[2] Faculty of Informatics, Vytautas Magnus University, Lithuania.

Human activity analysis is one of the most important and challenging domains in computer vision which can have wide applications, including surveillance systems, patient monitoring, human computer interaction, video indexing, etc. Although human activity recognition itself is beneficial in some offline analysis but is not sufficient in lots of the real time applications. In real world operations, we need to have a predictive action recognition algorithm which can recognize actions early in time.

In this study we have focused on prediction of manipulation actions, which is important for industrial as well as service robotics and also play an important role in Human-Robot Interaction (HRI). To achieve this we develop the so-called Enriched semantic Event Chain (ESEC) framework [1] which is a much extended version of the Semantic Event Chain (SEC) [2]. SEC only considers two relations "Touching" and "Not touching" (T/N) between manipulated objects while ESEC uses 16 spatial relations. Spatial relations are divided into static and dynamic relations.

Static Spatial Relations (SSR) depend on the relative position of two objects in space and include "Above", "Below", "Right", "Left", "Front", "Back", "Inside" and "Surround". Right, Left, Front and Back are merged into "Around". The relations "Above", "Below" and "Around" are assumed to happen in case the relation "Not touching" holds. When paired with the "Touching" relation (that is, two objects are in physical contact with each other), the corresponding relations are called: "Top", "Bottom" and "Touching Around".

Dynamic Spatial Relations (DSR) define the spatial relation of two objects during movement of either or both of them. Here, different from SSR, some information from the previous K frames (e.g., distance related parameters) between each pair of objects is necessary. Dynamic relations consist of "Getting close", "Moving apart", "Stable", "Moving together", "Halting together" and

"Fixed moving together". Fig.1. represents static and dynamic spatial relations between two objects in terms of cubes.

ESEC creates a temporal sequence of static and dynamic spatial relations between the objects that take part in the manipulation aiding early action recognition. Mathematically speaking, ESECs are transition matrices that symbolically encode the relational static and dynamic changes between (unspecified) objects. Each row of an ESEC matrix represents the sequence of the spatial relations between each pair of manipulated objects attained during the continuous video. Whenever a change occurs in any of those spatial relations a new column is created. As a consequence, every column reflects at least one such change. In order to facilitate the spatial relations computations, we model each object in a simple AABB (Axis-Aligned Bounding Box) and perform calculations based on the relationships between the AABBs [1].

The special way of manipulation actions representation in ESEC method by using static and dynamic spatial relations allows us to use the ESEC action matrices for action prediction. For this, the T/N, SSR, and DSR relations are computed for each pair of so called "fundamental" objects. We consider the object to belong to the set of fundamental objects if this object is being touched or untouched by some other object during the action. For action prediction, we perform column-wise comparison of the matrix of that action to the matrices from the training data set (in this case we use several action matrices as models for each action class) until all actions are categorized into a set which consists of the action members from the same class, or where there are no identical columns with any of the actions. In the latter case, we compute the similarity measure as presented in [3] for those incomplete action tables and predict the label based on the maximum similarity score. In case scores are identical for several action from different classes we proceed to the next column until a unique class is obtained.

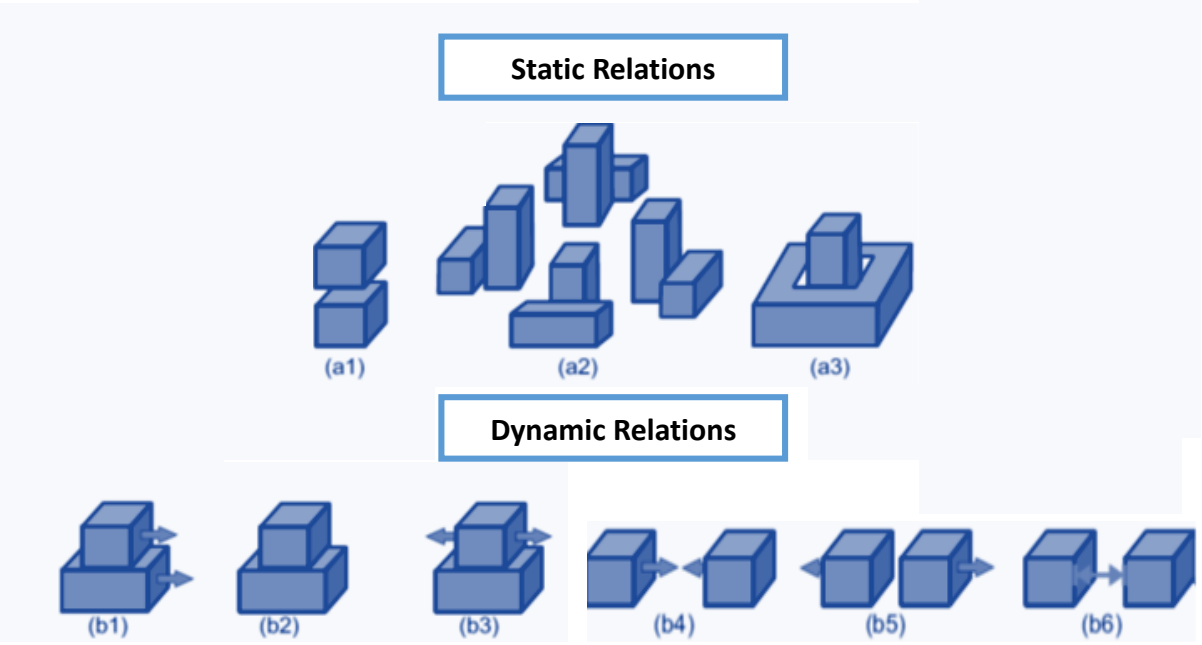

Figure 1: Spatial relations between the cubes:

a) Static Spatial Relations: (a1) Above/Below, (a2) Around, (a3) Inside/Surround.

(b) Dynamic Spatial Relations: (b1) Moving together, (b2) Halting together, (b3) Fixed-Moving Together, (b4) Getting Close, (b5) Moving Apart, (b6) Stable.

All details about ESEC prediction method and similarity measurement algorithms have been completely discussed in [3]. Moreover, we compared predictability power of ESEC framework with a state of the art HMM based trajectory recognition method and obtained that ESECs outperform HMM [3]. Now, we are going to compare the predictability power of manipulations between ESEC and humans in a Virtual Reality (VR) experiment.

We have defined 35 manipulation actions theoretically in our previous study [3]. Manipulations can be divided into three main groups (Fig.2.): "Hand-Only Actions", "Separation Actions" and "Release Determined Actions". Hand-Only Actions are actions where the hand alone acts on a target object (or first grasps a tool and then the tool acts on the target object). According to their goals and effects on the scene they can be subdivided into "Rearranging" (like stirring) and "Destroying" (like cutting) actions. Separation Actions denote actions where the hand manipulates one object to either destroy it or remove it from another object. This group is also divided into two cases: "Break" (ripping-off) and "Take-Down" (taking down one object from another one).

Finally, there are so-called Release Determined Actions, which include all actions where the hand manipulates an object and combines it with another object. This type of actions is subdivided into "Hide" (covering an object with another one) and "Construct" (building a tower) [4].

In this study we selected 10 actions, including Push, Shake, Lay, Stir, Cut, Chop, Take down, Uncover, Put on top and Hide, which are distributed in all possible groups and subgroups of manipulations (see Fig.2). Then we made 30 sample scenarios of each of which in Virtual Reality (totally 300 scenarios), each scenario with a different geometrical and coloring setup.

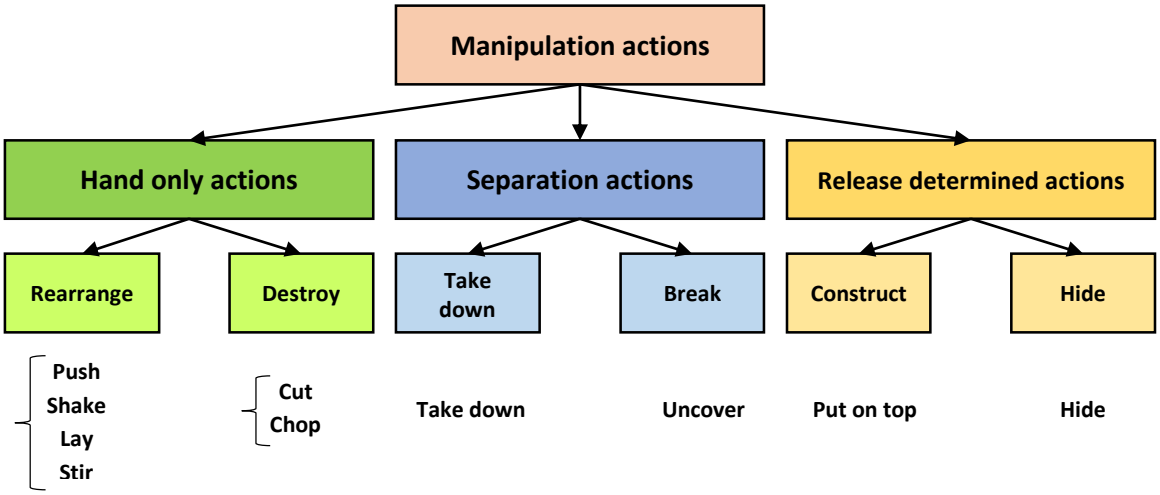

Figure 2. Theoretical categorization of manipulation actions according to [4] and the list of our selected manipulations in their groups.

As ESEC prediction method does not use object recognition and considers each object as an AABB cube, we used only simple cubes with different sizes and colors instead of true objects in our VR design. Thus, we were asking humans to recognize actions without knowledge on concrete objects. A sample of "Hide" action in the VR design is shown in Fig.3. The red cube is a "hand". The hand with the green cube is in the process of hiding the purple cube.

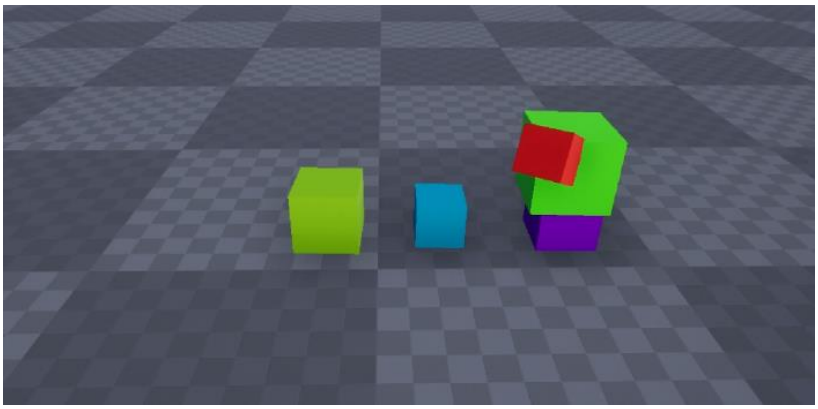

Figure 3. A sample of "Hide" action in a VR scenario

The procedure of the VR experiment is the following: the scenarios are shown to a person randomly (10*30=300 scenarios for each person). Whenever a person predicts the type of an action, he has to press a button in the virtual reality controller and stop the scenario. Then, the person has to select the action name from the list of ten actions (note, the list is always in the background of the scenario, but only after the button-press the participant is given access to select from the list). The time of pressing the VR controller button is recorded as the time of the person's prediction and the *Prediction Power (P.P.)* is calculated:

$$P_{1\leq i\leq 300}(\alpha_i) = \left(1 - \frac{T(\alpha_i)}{L(\alpha_i)}\right) * 100$$

Where $P(\alpha_i)$, is the total time of the video during execution of an action α_i and denotes the duration of the action and $T(\alpha_i)$ is the moment of prediction.

We had 50 persons performing VR experiments and calculated prediction power, both for human participants and for the ESEC algorithm, as explained in [3]. In Fig. 4 we show comparison of the prediction power between humans and the ESEC algorithm. Average and median results are provided. The indicated prediction power for humans is first averaged (or processed by taking a median) for each participant when predicting 30 samples of the same action and afterwards averaged one more time (or, alternatively, processed by taking the median) among 50 participants. For the algorithm average (or median) are taken for the 30 samples of the action. The comparison

shows, that in the given set-up the ESEC-based algorithm performs not worse than a human in action prediction.

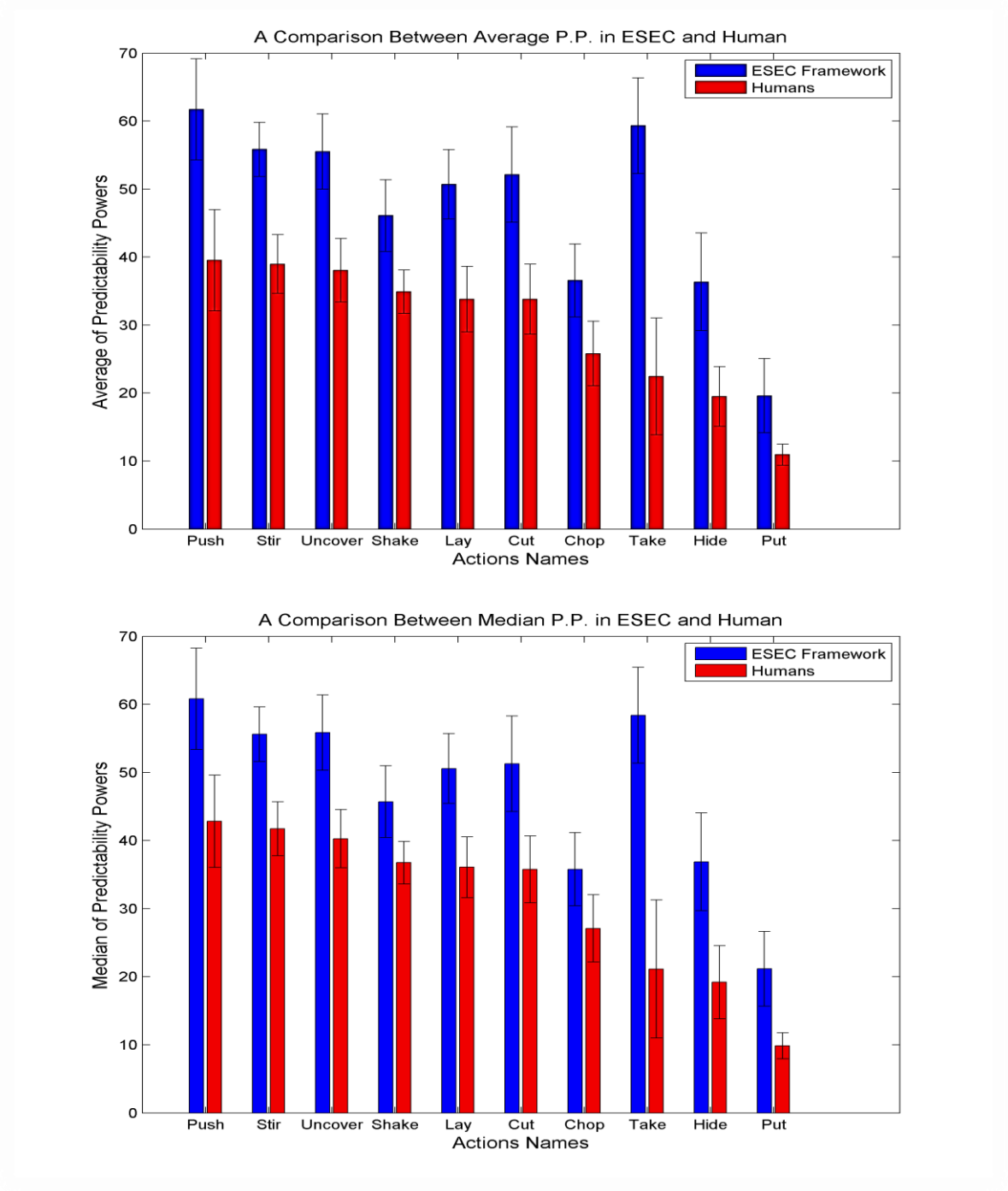

Figure 4. A comparison between average and median of each action predictability power between ESEC and the people

References

[1] F. Ziaeetabar, E. E. Aksoy, F. Wörgötter, M. Tamosiunaite, Semantic analysis of manipulation actions using spatial relations, in: IEEE Int. Conf. on Robotics and Automation (ICRA), 2017, pp. 4612-4619.

[2] E. E. Aksoy, A. Abramov, J. Dörr, K. Ning, B. Dellen, F. Wörgötter, Learning the semantics of object-action relations by observation, The International Journal of Robotics Research 30 (10) (2011) 1229-1249.

[3] F. Ziaeetabar, T. Kulvicius, M. Tamosiunaite, F. Wörgötter, Recognition and prediction of manipulations using enriched semantic event chains, Robotics and Autonomous Systems (110) (2018) 173-188.

[4] F. Wörgötter, E. E. Aksoy, N. Krüger, J. Piater, A. Ude, M. Tamosiunaite, A simple ontology of manipulation actions based on hand-object relations, IEEE Trans. on Autonomous Mental Development 5 (2) (2013) 117-134.