In: IEEE International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil 2018

Recognition and Prediction of Human-Object Interactions with a Self-Organizing Architecture

Luiza Mici, German I. Parisi and Stefan Wermter
Knowledge Technology, Department of Informatics, University of Hamburg, Germany
{mici, parisi, wermter}@informatik.uni-hamburg.de

***Abstract*—The recognition and prediction of human actions are challenging perception tasks that require reasoning upon a large space of fine-grained body motion patterns. In this work, we propose a hierarchical self-organizing architecture which jointly learns to recognize and predict human-object interactions from RGB-D videos. Our model consists of a hierarchy of Grow-When-Required (GWR) networks which process and learn co-occurring actions and objects from the training data. Our goal is to learn prototype body motion patterns when manipulating objects as well as to internally store prototype transitions of body postures over time. The architecture can generate sequences of arbitrary length given an observed initial motion pattern as well as predict future action labels. Experimental results on a dataset of daily activities demonstrate that our architecture recognizes ongoing actions and predicts the upcoming ones with high accuracy. The generated body pose trajectories demonstrate that our architecture is suitable to be further applied to the problem of the look-ahead planning of a robotic response in a human-robot interaction scenario.**

***Index Terms*—self-organization, action prediction, action recognition, human-object interaction**

I. Introduction

Body action recognition and prediction are key components of the humans' capability to interact and cooperate with others. Applied to intelligent systems and robotic agents, such a capability can lead to a wide range of real-world applications such as health-care and assistive technologies as well as human-robot interaction and cooperation. Assistive robots that are able to predict an activity before it has been fully executed, can act anticipatorily and not just reactively or when given a command. For instance, when a robot sees a person holding a water carafe, it could infer that the person wants to drink and, consequently, it would react by fetching a cup. There has been extensive work on human action recognition since the early 1990s [1]. In the last decade, the field has moved towards the recognition of complex realistic human activities involving objects or multiple persons. However, most of the existing approaches have focused on recognizing activities after a full observation, leaving action prediction an open challenge [2, 3].

We approach the task of human activity prediction from the view of the hierarchical compositionality of human activities, i.e., the spatio-temporal segmentation of the activities in atomic actions. For instance, the *drinking* activity can be segmented into the atomic actions of picking up the cup and bringing the cup towards the mouth. The segmentation of actions into smaller sub-sequences that can be learned and ideally reused for encoding different actions is the basis of the motor schemata theory in the research field of learning by imitation [4]. Despite important advances in this field, the automatic extraction of such sub-sequences is still an open problem [5, 6].

Neurobiological studies [7] have shown that the human brain can understand activities by observing only a few of its composing motor acts. Neuroscientists explain this fact with the mirror-neuron system found in the mammalian brain that responds not only when observing an action but also when a similar action is being executed. Thus, an internal simulation of the observed action based on the observer's motor repertoire allows him/her to better understand others' intentions [8]. The underlying mechanism for both the execution and recognition of the actions in the brain is believed to be the propagation of the activity within *neuronal chains* encoding subsequent motor acts [9, 10]. Moreover, this chain activation mechanism during action observation was found to be strictly modulated by the visual cues in the environment. From the experiments with a monkey observing demonstrations of the task of grasping to eat and grasping to place an object, it was found that the observer could only make predictions based on the object's identity [8]. For instance, the presence of food led the monkey to anticipate the action of eating, while the presence of another object led more often to placing.

In this paper, we propose a hierarchical architecture which learns prototypical segments of body pose and motion during human-object manipulations. The architecture is equipped with a temporal association mechanism for learning consecutive body motion patterns. Due to this mechanism, the architecture is able to receive only the action segment(s) starting a learned sequence and to carry out the most likely completion of the given sequence over time. The architecture presented in this paper is an extension of our previous work [11, 12] on the recognition of human-object interactions from RGB-D videos, which exhibited state-of-the-art performance on a benchmarking dataset. The building block of our architecture is the Grow-When-Required (GWR) algorithm [13], an incremental self-organizing network with no pre-defined topological structure and a dynamic number of neurons. Static topological arrangements of neurons, as in the Self-Organizing Map algorithm (SOM) [14], have shown a considerable negative impact on the resulting input data mappings, especially in the case of high-dimensional data which are usually obtained from the visual perception. Moreover, the GWR provides us with a flexible neuron insertion strategy which we adapt for the

purpose of creating specialized object-directed chains of action segments, which will be described in detail in Section III-B. The architecture proposed in this paper is novel in two main aspects: First, our learning mechanism can unequivocally map objects with possible actions. This allows for the bidirectional retrieval of the information, i.e., it is possible to retrieve the appropriate object given a body action as well as to retrieve body motion patterns for manipulating a given object. Second, we use the same learning mechanism, i.e, the temporal Hebbian connectivity, for both learning the spatio-temporal dynamics of the body motion and the temporal order of the action sequences in longer activities.

We evaluate our architecture with a dataset of RGB-D videos containing daily human-object manipulations [12]. Our experimental results show a high accuracy of the architecture in learning and anticipating plausible future actions. We also demonstrate the architecture's capability to synthesize body motion, thereby allowing for anticipating the way the next action will be performed. For robotic applications, the latter becomes relevant especially when robot planning takes place in shared environments.

II. Related Work

Recognition of human-object interactions: The understanding of human-object interactions from visual perception requires the integration of cues about human body pose and motion and cues from the environment, e.g. the identity of the manipulated objects [1]. From a computational perspective, it is unclear how to combine architectures specialized in object recognition and pose/motion estimation. Recently, Fleischer et al. [15] proposed a computational model based on dynamic neural fields for the recognition of actions such as grasping, placing, and holding. However, this model is more concerned with explaining physiological mechanisms of human visual processing rather than applications in real-world scenarios due to its limitation to an input with a uniform background.

Probabilistic approaches are quite popular for reasoning upon relationships and dependencies between objects and body motion. Gupta et al. [16] model hand trajectories with Hidden Markov Models (HMMs) and classify actions like *reach* or *manipulation* through a Bayesian network model. Graphical models and Conditional Random Fields (CRF) can also model the mutual context between objects and human pose [17, 18]. However, these types of models suffer from high computational complexity and require a fine-grained segmentation of the action sequences. Early attempts to apply neural networks for the problem of understanding human-object interactions from visual perception yielded promising results. Shimozaki and Kuniyoshi [19] proposed a SOM-based hierarchical architecture capable of integrating object categories, spatial relations, and movement and it was shown to perform well on simple 2D scenes of ball handling actions. However, the literature suggests that compared to the static image domain, there is limited work on understanding human-object relationships from video data sequences with neural network architectures [20, 21].

Action prediction: The goal of human activity prediction has been formally defined as the capability to infer an ongoing activity given an incomplete temporal observation [3]. Several approaches have been proposed, often referred to as *early activity recognition*, with the primary goal to infer the activity label from just the initial part of the video sequence [3, 22]. Lan et al. [23] proposed the hierarchical *movemes*, a new human motion representation that allows for describing motion at multiple levels of granularity and developed a learning framework on top of it for performing action prediction. SOM-based architectures have also been proposed for the purpose of action prediction [6] and motion sequence completion [24]–[26]. However, in contrast to these approaches, our goal is to learn motion patterns and the temporal order of the atomic actions composing an activity with a chain-like representation. The application of our architecture goes beyond the early recognition of an action while it is unfolding and aims for the prediction of what is likely to happen next. This would allow, for example, a better planning of a robot's response.

Typical hierarchical representations of human activity rely on hybrid approaches in which perceptual sequences are learned, e.g., through a neural network model, at the lower level and are combined into more complex sequences, or activities, by assigning them arbitrary symbols or rules [27, 28]. However, these symbols are usually fixed and defined a priori by the designers based on their knowledge.

Chersi et al. [10] have proposed a SOM-based computational model that combines principles of Hebbian learning and topological self-organization and reproduces the encoding of actions as well as language in the brain. The learning mechanism of this model establishes neural pools, i.e., neurons responding to similar visual input, and links among these pools which then form goal-directed neuronal chains. Although this model is quite interesting by considering human cognition mechanisms, no results have been reported on real-world action recognition applications. The work of Koppula et al. [29] is similar to ours in addressing the problem of anticipation of human actions at a fine-grained level of atomic actions. The authors also focus on predicting not only *what* comes next but also *how* it is performed. However, they demonstrate the capability of their architecture to generate object trajectories, whereas we focus on body postures.

III. Our Approach

The focus of this paper lies in the anticipation of what a human will do next given the current observation of his/her pose and the surrounding environment. For this purpose, we extend our previous architecture [11, 12] which was employed for the recognition of human-object interactions with a Hebbian association mechanism that allows for learning temporal dependencies of subsequent atomic actions. The architecture consists of three network streams processing separately visual features of the body pose-motion and of the objects being manipulated. The information coming from the three streams is then integrated in order to develop spatio-temporal representations of action segments. We implement a GWR

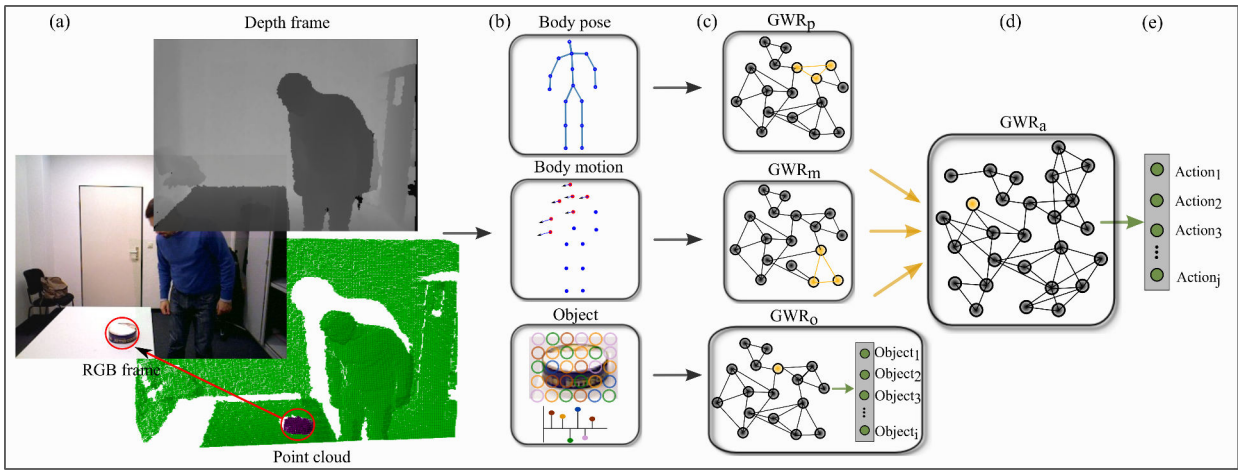

Fig. 1. Overview of the proposed architecture. (a) We use a point-cloud-based segmentation algorithm for automatically extracting the region of interest for the manipulated object. (b) A skeletal body representation is considered for the encoding of body pose and motion. A set of local features is extracted from the RGB image of the segmented object. (c) Three GWR networks are trained separately on each information channel and the *GWR*$_o$ learns to classify objects. (d) The last network, *GWR*$_a$, learns the combinations of body pose-motion and the object(s) involved in an action. Equipped with temporal Hebbian connections it learns the order of consecutive spatio-temporal segments. (e) Action labels are associated with each neuron in the *GWR*$_a$ network in order to evaluate the architecture's action classification and prediction performance.

algorithm [13] for the processing of the visual information and a modified GWR with the aforementioned temporal connections for the integration of the processed information. For evaluation purposes, we add a semantic layer including labels of the atomic actions and add associative connections between the integration module and this layer. An overview of the architecture is given in Fig. 1.

The proposed architecture has three main properties for the modeling of human-object interaction activities. First, activities are modeled as hierarchical structures in time, i.e., they are decomposed in sequences of atomic actions. Second, object identities are associated with actions in an unsupervised manner and serve as context information for disambiguating similar motion patterns. Third, human motion trajectories are internally stored and can be retrieved at any point in order to predict and simulate how an action can be performed on a given object.

A. Learning action-object segments

We adopt hierarchical GWR learning for the data processing and integration [26]. The hierarchical training is carried out layer-wise. We first extract visual features of body pose, A, body motion, B, and manipulated objects, O, from the training image sequences, as described in Section III-E. Then, we separately train the *GWR*$_p$ network with the body pose features, the *GWR*$_m$ with body motion, and the *GWR*$_o$ with the objects. After the training is completed, the *GWR*$_p$ will have created a set of prototype neurons representing typical pose configurations, the *GWR*$_m$ will have neurons for prototype body motion vectors and the *GWR*$_o$ network will have learned to classify objects appearing in each action sequence.

In order to encode spatio-temporal dependencies within the body features prototype space, we compute the neural activations of the *GWR*$_p$ and *GWR*$_m$, i.e., the best-matching units $b(\cdot)$, and apply the delay embedding technique [30]. For this, we take trajectories of neural activations over time and group them into vectors of the form:

$$\psi_i(\mathbf{x}) = \{b(\mathbf{x}_i), b(\mathbf{x}_{i-\xi}), ...b(\mathbf{x}_{i-(q-1)\xi})\}, i \in [q, k], \quad (1)$$

where k is the total number of training frames and q and ξ are the embedding parameters denoting the width of the time window and the lag or delay between two consecutive frames, respectively. The choice of q is not critical as long as it is large enough. The lag parameter ξ, on the other hand, is chosen in order to maximize the independence of the delay vector components. The embedding parameters are data-dependent and can be set following a heuristic method or, as in our case, can be chosen empirically. As a result, we obtain two sets of delay-embedded vectors with equal cardinality, one for the body pose $\psi_i(\mathbf{p})$ and one for the body motion $\psi_i(\mathbf{m})$, with $\mathbf{p} \in A$ and $\mathbf{m} \in B$.

The object data sample $\mathbf{y} \in O$ extracted at the beginning of each action sequence is provided as input to the *GWR*$_o$ network and the corresponding best-matching units $b(\mathbf{y})$ are computed. The label of the *GWR*$_o$ best-matching unit is represented in the form of a one-hot encoding, i.e., a vectorial representation in which all elements are zero except the ones with the index corresponding to the recognized objects' category. When more than one object appears in one action sequence, the object data processing and classification with *GWR*$_o$ is repeated as many times as the number of additional objects. The resulting one-hot-encoded labels are merged into one fixed dimension vector for the following integration step.

Finally, all information processed by the GWR networks in the first layer of the architecture is integrated into a higher dimensional vector:

$$\phi_i = \psi_i(\mathbf{p}) \oplus \psi_i(\mathbf{m}) \oplus l_o(\mathbf{y}), i \in [q, k-q], \quad (2)$$

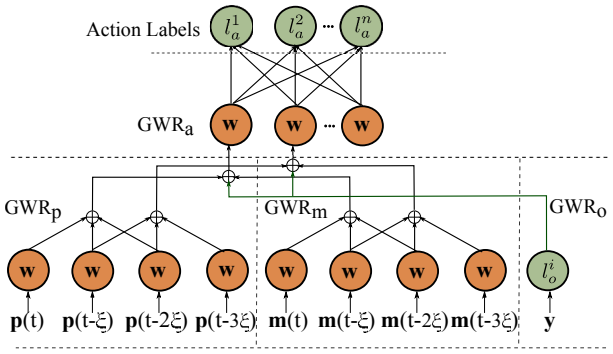

Fig. 2. Schematic description of the hierarchical learning and of the association of action labels (not all neurons and connections are shown). At each time step t, the body pose $\mathbf{p}(t)$ and body motion $\mathbf{m}(t)$ are represented by the weight $\mathbf{w}$ of the winner neurons in *GWR*$_p$ and *GWR*$_m$ respectively. Then each of these weight vectors are concatenated with the previous winner neuron weights (two previous neurons in this example) and the category label of the object l_o^i in order to compute the winner neuron in *GWR*$_a$. Each *GWR*$_a$ neuron is equipped with Hebbian connections to the semantic layer and the most frequently matched class will be the recognized action.

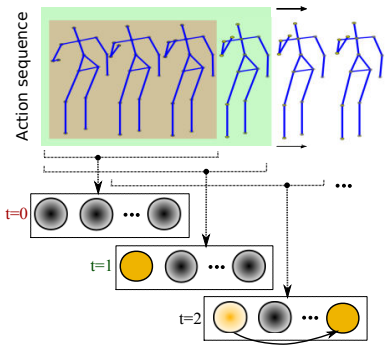

Fig. 3. An illustration of how the temporal connections are established between consecutive winners in the *GWR*$_a$ module. For simplicity, the action frames are depicted only as body skeletal configurations. At time $t = 0$ the delay embedded vector (highlighted in red) has a width lower than the defined time window (highlighted in green). Thus, no response is obtained from the network yet.

where $\oplus$ denotes the concatenation operator (see Fig.2). We will refer to the computed ϕ_i by the name *action-object segment*. Each segment is thus comprised of two parts:

1) the pre-processed visual sensor information about the body pose and motion,
2) the context information about the manipulated object, which is necessary to deal with ambiguities during the recognition and recall of segments which are shared among different action sequences.

The set of newly computed spatio-temporal vectors is then used for training the *GWR*$_a$ network.

B. Learning action chains

Now, we describe how we augment the GWR algorithm with two simple mechanisms in order to store and recall goal-directed action chains. For capturing the temporal aspects of human-object interaction sequences, we employ a time-delayed Hebbian learning rule in the *GWR*$_a$ network which develops asymmetric temporal connections among the neurons. This learning mechanism has been successfully applied to the problem of trajectory learning with a self-organizing network [25]. Interestingly, sequence completion driven by asymmetric connections between neurons is believed to be a feature of the human cortex [31].

We define a fully connected matrix of weighted connections Ω among the neurons of the *GWR*$_a$ network. The weights are adjusted when the winner neuron in each learning iteration is determined indicating the correct temporal order of the action-object segments (see Fig. 3). The learning rule is as follows:

$$\Delta\omega_{ij} = \mu \cdot a_j(t) \cdot a_i(t-1), \tag{3}$$

where $0 < \mu < 1$ is the temporal learning rate, i and j are the indices of the best matching units at time step $t-1$ and t and a denotes the activity of the GWR network. In other words, when the two neurons i and j are activated at time $t-1$ and t respectively, the temporal connection between them is strengthened in proportion to their activation, i.e., their similarity to the input data. The network's activity is computed as a nonlinear function of the Euclidean distance between the weight of the best-matching unit $\mathbf{w}_b$ and the input data sample that it matches at time step t:

$$a(t) = \exp(-||\mathbf{x}(t) - \mathbf{w}_b||). \tag{4}$$

Taking the network's activation into account for the update of the temporal connections can help alleviate the problem of high quantization errors during the first learning iterations, while the network's growth is still taking place. Therefore, the farther the consecutive winner neurons are from the input they match, the less is the temporal connection between them strengthened. The temporal weights are initialized to zero. Thus, non-zero connections are established only among consecutive winners and represent frequent transitions between action-object segments seen during training.

Since different trajectories, i.e., action sequences, should be handled by one single network, attention should be paid to have neurons responding to unique action-object pairs, e.g., *pick up mug* and not to *pick up phone* in order to have unambiguous goal-directed chains. For this purpose, we provide the objects' identity as a binary vector to each action-object segment as described in Section III-A. However, the GWR neuron competition during training, which is based on the Euclidean distance function, does not guarantee that neurons specialize in unique action-object pairs. Thus, we apply a weighted Euclidean distance function to consider equally the two composing components of the action-object segments. Additionally, we modify the neuron insertion strategy in the following way: if the weight vector of the best-matching unit computed at time step t contains the identity of an object, o_b, different from the matched input, $o_{x(t)}$, then a new neuron is created.

C. Activity classification

While leaving the learning of the *GWR*$_a$ network unsupervised, we simultaneously link each neuron to a symbolic action label $l \in L$, where L is the set of action classes. The *GWR*$_a$ will then have a many-to-many relation with the symbolic layer. The set of weights Π, which are initialized to zero, are updated according to a Hebbian learning rule:

$$\Delta\pi_{il_j} = \gamma \cdot a_i(t), \tag{5}$$

where $0 < \gamma < 1$ is the learning rate, $a_i(t)$ is the activity of the winner neuron at time step t and l_j is the target action label. After the training phase is complete, the weights are normalized by scaling them with the corresponding inverse class frequency and with the inverse neuron activation frequency. In this way, class labels that appear less during training are not penalized, and the vote of the neurons is weighed equally in spite of how often they have fired. The learning procedure of the GWR algorithm including the so far described modifications is illustrated in Algorithm 1.

At recognition time, given one temporal segment of a human-object interaction at the time step t, the best-matching unit $b(t)$ is computed in the *GWR*$_a$ module and the action label is given by:

$$l_j = arg\max_{l \in L}(\pi_{b,l}). \tag{6}$$

In order to classify an entire action sequence, a majority vote labelling technique is applied on the labels of its composing temporal segments.

D. Action prediction

During the prediction phase, each action sequence is presented to the trained architecture and the action-object segments are computed as described in Section III-B. As can be seen in Fig. 3, the first winner neuron of the *GWR*$_a$ is obtained at time $t = 1$, i.e., after the first q frames have been processed and the first composing temporal segment is available. The one-step-ahead prediction of the sequence can then be computed following the outgoing temporal connection with the maximal weight. In the case that the desired prediction horizon is greater than 1, the multi-step-ahead prediction can be obtained by recursively applying the one-step-ahead prediction computation. In both cases, the predicted action label for the last activated neuron is given by Eq. 6.

In contrast to our previous work where we focused mainly on the motion prediction task [32], here we are interested in the higher-level action prediction problem, which is often indeterministic. For instance, the action of *picking up can* can lead to both the action of *drinking* as well as the action of *pouring* from the can to a container like a mug. In the current architecture, such ambiguities are represented by multiple outgoing temporal connections with non-zero weights. In other words, the maximal temporal weight gives the most probable, but not the only possible transition, after the observed action-object segment.

The proposed architecture has the advantage of self-organizing and learning sequences of arbitrary lengths in an unsupervised manner. The recall of a sequence can start at any point given one component action-object segment. Finally, the architecture can learn ordered action sequences, in our case called atomic actions, as a single long sequence, thereby providing a mechanism to recall the atomic action following the observed one.

Algorithm 1 The modified GWR algorithm (used for training the *GWR*$_a$ module)

1) Create two random neurons with weights $\{\mathbf{w}_1, \mathbf{w}_2\}$
2) At each iteration t, generate an input sample $\mathbf{x}(t)$
3) Select the best and second-best matching neuron: $b = arg\min_{n \in A} ||\mathbf{x}(t) - \mathbf{w}_n||$, $s = arg\min_{n \in A/\{b\}} ||\mathbf{x}(t) - \mathbf{w}_n||$
4) Create a connection $E = E \cup \{(b, s)\}$ if it does not exist and set its age to 0.
5) If $(a(t) < a_T)$ and $(h_b < f_T)$) or $(o_b \neq o_{x(t)})$ then:
   - Add a new neuron r ($A = A \cup \{r\}$) with $\mathbf{w}_r = 0.5 \cdot (\mathbf{x}(t) + \mathbf{w}_b), h_r = 1$,
   - Update edges: $E = E \cup \{(r, b), (r, s)\}$ and $E = E/\{(b, s)\}$.
6) If no new neuron is added:
   - Update best-matching neuron and its neighbors i: $\Delta\mathbf{w}_b = \epsilon_b \cdot h_b \cdot (\mathbf{x}(t) - \mathbf{w}_b)$, $\Delta\mathbf{w}_i = \epsilon_i \cdot h_i \cdot (\mathbf{x}(t) - \mathbf{w}_i)$, with the learning rates $0 < \epsilon_i < \epsilon_b < 1$.
   - Increment the age of all edges connected to b by 1.
7) Update the temporal connection weight between $b(t)$ and $b(t-1)$ following Eq. 3.
8) Update the symbolic connection weight between $b(t)$ and the target action label l_j following Eq. 5.
9) Reduce the firing counters of the best-matching neuron and its neighbors i: $\Delta h_b = \tau_b \cdot \kappa \cdot (1 - h_b) - \tau_b$, $\Delta h_i = \tau_i \cdot \kappa \cdot (1 - h_i) - \tau_i$ with constant τ and κ controlling the curve behavior.
10) Remove all edges with ages larger than a pre-defined threshold and remove neurons without edges.
11) If the stop criterion is not met, repeat from step 2.

E. Feature extraction

The estimation and segmentation of body pose from RGB image sequences are challenging due to dynamic backgrounds, changes in ambient illumination, occlusion, and the difference in the point of view. For this reason, we use the Asus Xtion camera which, together with the OpenNI framework, provides us with reliable estimations of three-dimensional articulated body pose and motion even in real-world environments. This type of depth sensor operates at a reduced power consumption and is quite light, making it a more suitable choice than a Kinect camera for being placed on top of a humanoid robot. In this paper, we consider only the position of the upper body joints *shoulders*, *elbows*, *hands*, center of *torso*, *neck*, *head* and *hips*, given that they hold all necessary information about the human-object interactions we focus on in this paper. However, neither the number of considered joints nor the dimensionality of the input data limits the application of our architecture for

the task of recognition and prediction of body actions. From each video frame, we extract the (x, y, z) position of each joint and we translate them into a coordinate system having the torso as the origin and concatenate them into one vector $\mathbf{p}$, which will then represent the body pose.

We also consider the body motion vector $\mathbf{m}$, which we define as the differences in position of the upper-body joints between two consecutive frames. We assume that these motion vectors, which encode the velocity of the movement between frames, hold significant information about apparently similar motion patterns, e.g. *pick up can* for *drinking* or *pick up can* for *pouring* its liquid into a mug. Behavioral studies with human infants have shown that the hands' motion velocity plays an important role in action anticipation [33]. Finally, the objects are extracted from the scene through a table-top segmentation algorithm and encoded with a modified version of the dense SIFT features [12].

IV. Experiments

We present our results on a dataset of human-object interactions, that has been previously used to report the performance of our recognition architecture [11, 12]. The dataset consists of 4 simulated daily activities: *drinking* (from a container like a mug or a can), *eating* (an edible object like a biscuit), *pouring* (from a can into a mug) and *talking on phone* performed by 6 subjects. For the experiments reported in this paper, each sequence was segmented into fine-grained atomic actions which define the activity: *pick up mug* and *drinking*, *pick up phone* and *talking on phone*, *pick up biscuit* from the biscuits box and *eating*, *pick up can* and *pouring* the liquid inside it into a mug. Additionally, we synthetically build longer action sequences in which the actions of *pick up mug* and *drinking* from mug follows the action of *pouring* from can to mug. With this type of sequence, we want to assess the prediction of more complex action sequences that lead to the inference of higher-level activities like, for example, *having meal*.

A. Predicting the action label

We now assess the action label prediction capability of our architecture on the RGB-D dataset of human-object interactions. We follow the same cross-validation scheme described in Mici et al. [12], i.e., we train our architecture on activities performed by 5 subjects and test on activities of an *unseen subject*. The parameters used for training the architecture throughout our experiments were determined experimentally and are listed in Table I. We define a time window width $q = 15$ and a lag $\xi = 3$ for the computation of the delay embedded vectors. In this way, we obtain action-object segments with a temporal length of 15 video frames.

In this set of experiments, we fix a prediction horizon of 500ms (i.e, 15 frames at 30fps) and compare the predicted action labels with the ground truth. In our dataset, the length of each atomic action is variable ranging from very short sequences, like *picking up* and *pouring* which can last less than a second, to very long sequences like *talking on phone* which can last up to 15 seconds. Thus, a prediction horizon of 500ms is necessary not to penalize the short sequences in our dataset. The precision, recall, and F1-score for each action class, computed across all 6 folds, are reported in Fig. 4. Additionally, we report the confusion matrix both for the predicted and for the classified (ongoing) actions in Fig. 5 and Fig. 6 in order to further clarify the obtained results.

TABLE I
Training parameters for each GWR network in our architecture for the learning of human-object interaction sequences.

Parameter	Value
Activation Threshold	$a_T = 0.98$
Firing Threshold	$f_T = 0.1$
Learning rates	$\epsilon_b = 0.1, \epsilon_i = 0.01$
Firing counter behavior	$\tau_b = 0.3, \tau_i = 0.1, \kappa = 1.05$
Maximum edge age	{100,100,200}
Training epochs	50
Hebbian connections	$\mu = 0.3, \gamma = 0.5$

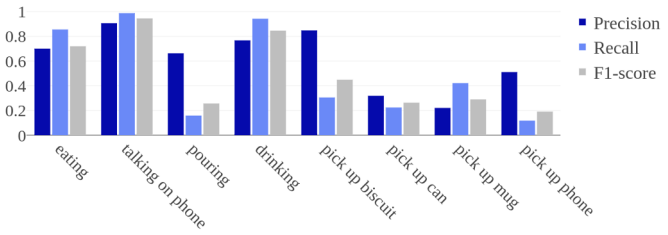

Fig. 4. Action label prediction results on the RGB-D human-object interactions dataset. Illustrated are precision, recall, F1-score, averaged over 6 trials of cross-validation.

Analyzing the confusion matrices we can observe that the actions of *eating*, *drinking* and *talking on phone* are predicted with high accuracy, even though the test sequences have never been seen during training. *Pouring*, on the contrary, is not predicted with the same accuracy. We assume that the reason for this is twofold: (1) the misclassification of the pouring frames (which is evident from the classification confusion matrix) due to the fact that the body pose for *pick up can* and *pouring* are very similar (see Fig. 7), (2) the architecture often predicts what comes after *pouring* already, like *pick up mug* in order to drink. In both cases, the considerable number of false negatives causes the drop of the *pouring* recall metric, as can be seen in Fig. 4.

As for the prediction of the *picking up* sub-sequences, the confusion matrix is farther from the diagonal. However, the results demonstrate the temporal ambiguity between consecutive atomic actions which can often lead to an imperfect segmentation. For instance, in the case of *pick up biscuit/mug/phone* the architecture predicts *eating*, *drinking* and *talking on phone*, respectively, quite early. However, this cannot be considered an error, but rather a desirable feature for real-time robotic applications, where the robot's response needs to be planned as much in advance as possible. Finally, there are obviously little to no implausible predictions such as *talking on phone* instead of *drinking* or *eating* and this shows that the architecture

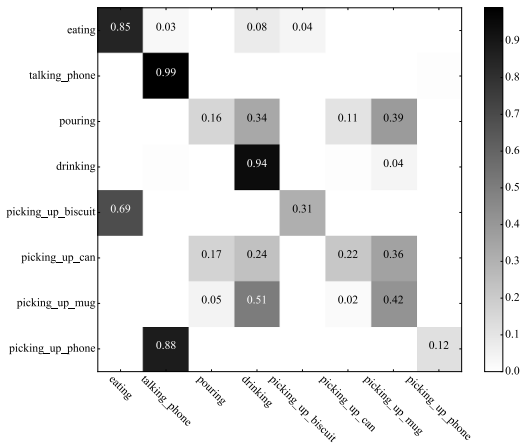

Fig. 5. Normalized confusion matrix of the **predicted** actions for the unseen subjects.

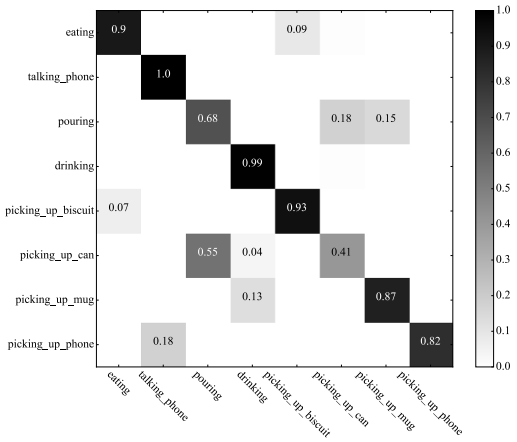

Fig. 6. Normalized confusion matrix of the **classified** actions for the unseen subjects.

successfully performs in the task that has been assigned to it.

B. Generating actions

We analyze the output of our architecture when simulating entire sequences of actions performed on a given object. While predicting action labels is important for a robotic platform when planning responses to those actions, predicting motion is crucial for planning robot motor commands in a shared workspace. In this round of experiments, we feed the trained architecture only the first action-object segment composing a learned sequence and rely on the GWR_a's temporal connections for completing the sequence automatically. In order to do so, we recursively compute the one-step-ahead action-object segment, as described in Section III-D. The iterations will stop when the current best-matching unit has no temporal connections to any other neurons - this indicates the end of the learned sequence. The performance of the architecture in this task is evaluated qualitatively, given that human motion

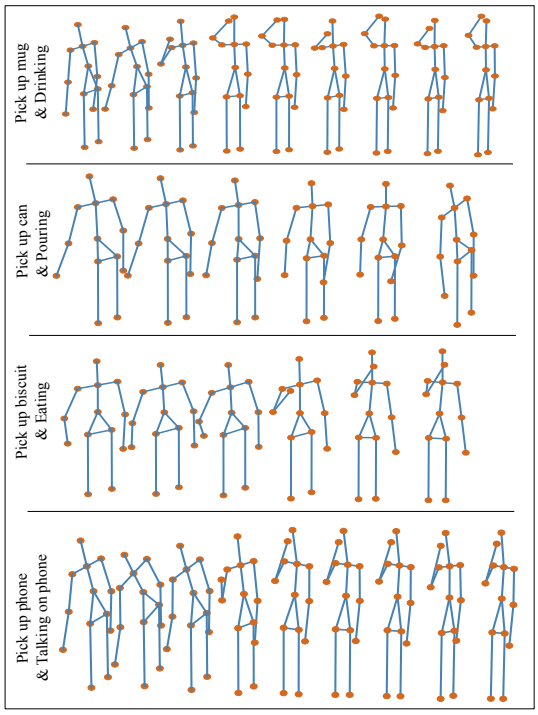

Fig. 7. Examples of body pose trajectories generated by the architecture when given only the starting action-object segment (not all frames are shown). Upper body joints are outputs of the GWR_a module, whereas the feet are added for illustration purposes.

synthesis is highly non-deterministic and its plausibility is hard to evaluate in a quantitative manner. Examples of the obtained sequence are depicted in Fig. 7.

One fundamental problem in generating such trajectories is to determine the correct temporal segment following a bifurcation point, i.e., a temporal segment shared between two different action sequences. For instance, without specifying a label the action of *drinking* may follow the action of *pick up can* instead of *pouring*, depending on which transition has been encountered more often. In this case, we make use of the action labels and predict the next segment considering also its label. In other words, among the outgoing temporal connections of the current action-object segment, we choose the one leading to the neuron with the specified action label.

V. Conclusions and Future Work

We presented a hierarchical neural network architecture for jointly learning to recognize and predict human-object interactions from RGB-D videos. In particular, we focused on how to extend a GWR learning algorithm in order to encode temporally ordered body motion patterns from sequences of arbitrary lengths. For this purpose, we employed a simple Hebbian learning mechanism which can associate consecutive network activations. Additionally, actions were represented as sequences of spatio-temporal segments consisting of body pose-motion and the identity of the manipulated object(s). The temporal association mechanism together with the GWR network's self-organizing capability led to the formation of

neural chains encoding goal-oriented actions. The formation of prototype neural chains resembles mechanisms for the execution and recognition of actions in the brain [9]. We showed that with the same underlying learning mechanism the architecture is able to predict body motion and scale to the prediction of action labels by learning ordered sequences of atomic actions. We evaluated our architecture with a dataset of human daily activities showing that it can predict plausible future actions with high accuracy albeit being tested on sequences never encountered before. Moreover, the action generation results showed that our architecture can deal with receiving only some initial sensory input and internally simulate the rest of the action without being fed any further input.

Our experimental results pointed out the need for fine-grained visual cues such as the pose of the hand and the objects, which could have led to a higher recognition accuracy, e.g., for the *pouring* action. To deal with this drawback, the 3D body skeletal representations should be complemented with additional computer vision pose estimation algorithms. Our dataset has a low inter-class variability, i.e., the motion patterns are very similar across different action classes. This made the action prediction more challenging and allowed us to focus on analyzing in detail the internal neural chain representations developed within our architecture. Future work will also focus on evaluating our architecture on larger-scale datasets composed of more complex sequences of actions.

ACKNOWLEDGMENT

The authors gratefully acknowledge partial support by the EU- and City of Hamburg-funded program Pro-Exzellenzia 4.0, the German Research Foundation DFG under project CML (TRR 169).

REFERENCES

[1] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, p. 16, 2011.

[2] N. P. Trong, H. Nguyen, K. Kazunori, and B. Le Hoai, "A comprehensive survey on human activity prediction," in *Proc. of ICCSA*. Springer, 2017, pp. 411–425.

[3] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *Proc. of IEEE ICCV*. IEEE, 2011, pp. 1036–1043.

[4] M. A. Arbib, "Perceptual structures and distributed motor control," *Comprehensive Physiology*, 1981.

[5] H. Arie, T. Arakaki, S. Sugano, and J. Tani, "Imitating others by composition of primitive actions: A neuro-dynamic model," *Robotics and Autonomous Systems*, vol. 60, no. 5, pp. 729–741, 2012.

[6] W. Ding, K. Liu, F. Cheng, and J. Zhang, "Learning hierarchical spatio-temporal pattern for human activity prediction," *Journal of Visual Communication and Image Representation*, vol. 35, pp. 103–111, 2016.

[7] C. S. Soon, M. Brass, H.-J. Heinze, and J.-D. Haynes, "Unconscious determinants of free decisions in the human brain," *Nature neuroscience*, vol. 11, no. 5, pp. 543–545, 2008.

[8] L. Fogassi, P. F. Ferrari, B. Gesierich, S. Rozzi, F. Chersi, and G. Rizzolatti, "Parietal lobe: from action organization to intention understanding," *Science*, vol. 308, no. 5722, pp. 662–667, 2005.

[9] F. Chersi, P. F. Ferrari, and L. Fogassi, "Neuronal chains for actions in the parietal lobe: a computational model," *PloS one*, vol. 6, no. 11, p. e27652, 2011.

[10] F. Chersi, M. Ferro, G. Pezzulo, and V. Pirrelli, "Topological self-organization and prediction learning support both action and lexical chains in the brain," *Topics in Cognitive Science*, vol. 6, no. 3, pp. 476–491, 2014.

[11] L. Mici, G. I. Parisi, and S. Wermter, "Recognition of transitive actions with hierarchical neural network learning," in *Proc. of ICANN*. Springer International Publishing, 2016, pp. 472–479.

[12] L. Mici, G. I. Parisi, and S. Wermter, "A self-organizing neural network architecture for learning human-object interactions," *ArXiv preprint arXiv:1710.01916*, Oct. 2017.

[13] S. Marsland, J. Shapiro, and U. Nehmzow, "A self-organising network that grows when required," *Neural Networks*, vol. 15, no. 8, pp. 1041–1058, 2002.

[14] T. Kohonen, *Self-organization and associative memory*. Berlin: Springer Science & Business Media, 1993.

[15] F. Fleischer, V. Caggiano, P. Thier, and M. A. Giese, "Physiologically inspired model for the visual recognition of transitive hand actions," *The Journal of Neuroscience*, vol. 33, no. 15, pp. 6563–6580, 2013.

[16] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1775–1789, Oct 2009.

[17] B. Yao and L. Fei-Fei, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1691–1703, 2012.

[18] H. Kjellström, J. Romero, and D. Kragić, "Visual object-action recognition: Inferring object affordances from human demonstration," *Computer Vision and Image Understanding*, vol. 115, no. 1, pp. 81–90, 2011.

[19] M. Shimozaki and Y. Kuniyoshi, "Integration of spatial and temporal contexts for action recognition by self organizing neural networks," in *Proc. of IEEE IROS*, vol. 3. IEEE, 2003, pp. 2385–2391.

[20] C. Lea, A. Reiter, R. Vidal, and G. D. Hager, "Segmental spatiotemporal cnns for fine-grained action segmentation," in *Proc. of ECCV*. Springer, 2016, pp. 36–52.

[21] C.-Y. Ma, A. Kadav, I. Melvin, Z. Kira, G. AlRegib, and H. P. Graf, "Attend and Interact: Higher-Order Object Interactions for Video Understanding," *ArXiv preprint arXiv:1711.06330*, Nov. 2017.

[22] Y. Cao, D. Barrett, A. Barbu, S. Narayanaswamy, H. Yu, A. Michaux, Y. Lin, S. Dickinson, J. Mark Siskind, and S. Wang, "Recognize human activities from partially observed videos," in *Proc. of IEEE CVPR*, 2013, pp. 2658–2665.

[23] T. Lan, T.-C. Chen, and S. Savarese, "A hierarchical representation for future action prediction," in *Proc. of ECCV*. Springer, 2014, pp. 689–704.

[24] M. Okada, D. Nakamura, and Y. Nakamura, "Self-organizing symbol acquisition and motion generation based on dynamics-based information processing system," in *Proc. of the second International Workshop on Man-machine Symbiotic Systems*, 2004, pp. 219–229.

[25] A. F. Araujo and G. A. Barreto, "Context in temporal sequence processing: A self-organizing approach and its application to robotics," *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 45–57, 2002.

[26] G. I. Parisi, J. Tani, C. Weber, and S. Wermter, "Emergence of multimodal action representations from neural network self-organization," *Cognitive Systems Research*, vol. 43, pp. 208–221, 2017.

[27] S. Wermter, *Hybrid neural systems*. Springer Science & Business Media, 2000, no. 1778.

[28] T. Taniguchi, T. Nagai, T. Nakamura, N. Iwahashi, T. Ogata, and H. Asoh, "Symbol emergence in robotics: a survey," *Advanced Robotics*, vol. 30, no. 11-12, pp. 706–728, 2016.

[29] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 14–29, 2016.

[30] F. Takens *et al.*, "Detecting strange attractors in turbulence," *Lecture notes in mathematics*, vol. 898, no. 1, pp. 366–381, 1981.

[31] P. Mineiro and D. Zipser, "Analysis of direction selectivity arising from recurrent cortical interactions," *Neural Computation*, vol. 10, no. 2, pp. 353–371, 1998.

[32] L. Mici, G. I. Parisi, and S. Wermter, "An Incremental Self-Organizing Architecture for Sensorimotor Learning and Prediction," *ArXiv preprint arXiv:1712.08521*, Dec. 2017.

[33] J. C. Stapel, S. Hunnius, and H. Bekkering, "Fifteen-month-old infants use velocity information to predict others action targets," *Frontiers in Psychology*, vol. 6, p. 1092, 2015.