A self-organizing neural network architecture for learning human-object interactions

Luiza Mici*, German I. Parisi, Stefan Wermter
Knowledge Technology, Department of Informatics, University of Hamburg, Germany

Abstract

The visual recognition of transitive actions comprising human-object interactions is a key component enabling artificial systems to operate in natural environments. This challenging task requires, in addition to the recognition of articulated body actions, the extraction of semantic elements from the scene such as the identity of the manipulated objects. In this paper, we present a self-organizing neural network for the recognition of human-object interactions from RGB-D videos. Our model consists of a hierarchy of Grow When Required (GWR) networks which learn prototypical representations of body motion patterns and objects, also accounting for the development of action-object mappings in an unsupervised fashion. To demonstrate this ability, we report experimental results on a dataset of daily activities collected for the purpose of this study as well as on a publicly available benchmark dataset. In line with neurophysiological studies, our self-organizing architecture shows higher neural activation for congruent action-object pairs learned during training sessions with respect to artificially created incongruent ones. We show that our model achieves good classification accuracy on the benchmark dataset in an unsupervised fashion, showing competitive performance with respect to strictly supervised state-of-the-art approaches.

Keywords: self-organization, hierarchical learning, action recognition, object recognition, human-object interaction.

1. Introduction

The recognition of transitive actions, i.e. actions that involve the interaction with an object, represents a key function of the human visual system for learning, goal inference, and social communication. The study of transitive actions such as grasping and holding has often been the focus of research in neuroscience and psychology [1, 2]. Nevertheless, this task has remained an open challenge for computational models of action recognition.

The capability of computational approaches to reliably recognize human-object interactions can establish an effective cooperation between assistive systems and people in real-world scenarios and can promote learning from demonstration in robotic systems [3, 4]. Given the outstanding capability of humans to infer the goal of actions from the interaction with objects, the biological visual system represents a source of inspiration for developing computational models. From the computational perspective, an important question arises regarding the potential links between representations of body postures and manipulated objects and, in particular, how these two representations can be integrated.

In the visual system, the information about body pose and objects are processed separately [5] and reside in distinct subcortical areas [6, 7]. Neuroscientists have widely studied object and action perception, with a focus on where and how the visual cortex constructs invariant object representations [8] and how neurons in the superior temporal sulcus (STS) area encode actions in terms of patterns of body posture and motion [9, 10]. It has been shown that the identity of the objects plays a crucial role for the complete understanding of human-object interactions [11] and modulates the response of specific action-selective neurons [12, 13, 14]. Yet, little is known about the underlying neural mechanisms for the integration of actions and objects.

In this paper, we present a neural network architecture that learns to recognize human-object interactions from RGB-D videos containing scenes of daily activities. The design of the proposed architecture relies on the following assumptions: (i) visual features of body pose and man-made objects are represented in two distinct areas of the brain [6, 7, 5], (ii) input-driven self-organization defines the topological structure of specific areas in brain [15], (iii) the representation of objects and concepts is based on prototypical examples [16], and (iv) the identity of the objects is crucial for the understanding of actions performed by other individuals [11, 12].

We develop a hierarchical architecture with the use of growing self-organizing networks, namely the Grow When Required (GWR) network [17], to learn prototypical representations of actions and objects and the resulting action-object mappings in an unsupervised fashion. Growing self-organizing networks have been an effective model for clus-
tering human motion patterns in terms of multi-dimensional flow vectors \cite{18, 19} as well as for learning object representations in an unsupervised fashion \cite{20}. The generative properties of this topology of networks make them particularly suitable for our task when considering a possible generalization of unseen action-object pairs.

The proposed architecture consists of two network streams processing separately feature representations of body postures and manipulated objects. A second layer, where the two streams are integrated, combines the information for the development of action-object mappings in a self-organized manner. On the basis of previously reported results in Mici et al. \cite{21}, this work contributes to improve the architecture design and provides a more in-depth analysis for an extended number of experiments. Unlike our previous work, we use the GWR network for all layers including the object recognition module for which we employed a self-organizing map (SOM) \cite{22}. The reason for this is the considerable impact on the resulting input data mappings of the predefined topological structure of the SOM, especially when having as input high-dimensional complex data distributions like perceptual representations of objects. In our previous model, an additional network was used for learning prototypes of temporal activation trajectories for the body pose processing stream before the integration phase. However, the impact on the overall classification accuracy of the network was minor while introducing more computational complexity.

We evaluate our architecture with a dataset of RGB-D videos containing daily actions acquired for the purpose of this study as well as with a publicly available action benchmark dataset CAD-120 \cite{23}. We present and discuss our results on both datasets. In particular, we look into the role of the objects’ identity as a contextual information for unambiguously distinguishing between different activities, the classification capabilities of our architecture in terms of recognition of human-object interaction activities, and the response of the network when given an input with incongruent action-object pairs.

2. Related work

One important goal of human activity recognition in machine learning and computer vision is to automatically detect and analyze human activities from the information acquired from visual sensing devices such as RGB cameras and range sensors. The literature suggests a conceptual categorization of human activities into four different levels depending on the complexity: gestures, actions, interactions, and group activities \cite{24, 25, 26}. Gestures are elementary movements of a person’s body part and are the atomic components describing the meaningful motion of a person, e.g. stretching an arm or raising a leg. Actions are single-person activities that may be composed of multiple gestures such as walking and waving. Interactions are human activities that involve a person and one (or more) objects. For instance, a person making a phone call is a human-object interaction. Finally, group activities are the activities performed by groups composed of multiple persons or objects, e.g. a group having a meeting.

Understanding human-object interactions requires the integration of complex relationships between features of human body action and object identity. From a computational perspective, it is not clear how to link architectures specialized in object recognition and motion recognition, e.g., how to bind different types of objects and hand/arm movements. Recently, Fleischer et al. \cite{1} proposed a physiologically inspired model for the recognition of transitive hand-actions such as grasping, placing, and holding. Nevertheless, this model works with visual data acquired in a constrained environment, i.e. videos showing a hand grasping balls of different sizes with a uniform background, with the role of the identity of the object in transitive action recognition being unclear. Similar models have been tested in robotics, accomplishing the recognition of grip apertures, affordances, or hand action classification \cite{3, 4}.

There is a number of techniques applied to the recognition of human-object interactions. The most typical approaches are those that do not explicitly model the interplay between object recognition and body pose estimation \cite{27, 28, 29}. Typically, first, objects are recognized and activities involving them are subsequently recognized, by analyzing the objects’ motion trajectories \cite{30}. Yang et al. \cite{31} proposed a method for learning actions comprising object manipulation from demonstrating videos. Their model is able to distinguish among different power and precision grasps as well as recognize objects by using a deep neural network architecture. Nevertheless, the human action is simply inferred as the action with the maximum log-likelihood ratio computed over all possible trigrams \texttt{<Object1, Action, Object2>} extracted from the sentences in the English Gigaword corpus.

Probabilistic approaches have been extensively used for reasoning upon relationships and dependencies among objects, motion, and human activities. Gupta et al. \cite{32, 33} proposed a Bayesian network model for integrating the appearance of manipulated objects, human motion, and reactions of objects. They estimate reach and manipulation motion by using hand trajectories as well as hidden Markov models (HMMs). The Bayesian network integrates all of this information and makes a final decision to recognize objects and human activities. Following a similar probabilistic integration approach, Ryoo and Aggarwal \cite{34} proposed a framework for the recognition of high-level activities. They introduced an additional semantic layer providing feedback to the modules for object identification and motion estimation leading to an improvement of object recognition rates and better motion estimation. Nevertheless, the subjects’ articulated body pose was not considered as input data, leading to applications in a restricted task-specific domain such as airport video surveillance. Other research studies have modeled the mutual context between objects and human pose through graphical models such as Conditional Ran-
The proposed architecture consists of two main network streams processing separately visual representations of the body postures and of the objects being manipulated. The information coming from the two streams is then combined for developing action-object mappings. The building block of our architecture is the GWR network [17], which is a growing extension of the self-organizing networks with competitive learning. An overview of the architecture is depicted in Fig. 1.

The body pose cue is processed under the assumption that action-selective pose cues are sensitive to the temporal order of prototypical patterns. Therefore, the output of the body pose processing stream is computed by concatenating consecutively activated neurons of GWR, following a sliding time window technique. The object appearance cue is processed in order to have topological arrangements in GWR where different 2D views of 3D objects as well as different instances of the same object category are mapped to close-by neurons in the prototypes domain. The advantage of having such topological arrangements consists in its capability to map any unseen view of a known object into the corresponding training views. This capability resembles, to some extent, biological mechanisms for learning the three-dimensional objects in the human brain [12, 13, 17]. Moreover, prototype-based learning approaches are supported by psychological studies claiming that semantic categories in the brain are represented by a set of most typical examples of these categories [16]. For evaluating the architecture in terms of classification of human-object interaction activities, semantic labels are assigned to GWR prototype neurons by extending the GWR algorithm with a labeling strategy.

3. Methodology

Self-organization is an unsupervised mechanism that allows us to represent the input probability distribution through a finite set of prototype vectors. Unlike traditional vector quantization (VQ) methods, self-organizing neural networks such as SOMs [22], neural gas (NG) [11] as well as their growing extensions, e.g., growing neural gas (GNG) [15] and the GWR algorithm [17] associate these prototype vectors with neurons that adaptively form topological preserving maps of the input space in an unsupervised fashion, i.e. similar inputs are mapped to neurons that are near to each other on the map. This input-driven self-organization and topology preservation capability are motivated by a similar neural mechanism found in specific areas of the human visual cortex [15].

From a computational perspective, the GWR algorithm proposed by Marsland [17] is more advantageous than the other learning approaches due to its ability to learn incrementally and to adapt the neuron connectivity patterns through learning. Unlike the GNG algorithm, the neural growth of the GWR algorithm is not constant but depends on the overall network activation with respect to the input. This leads to a faster convergence and makes the GWR algorithm more suitable for learning representations of non-stationary datasets while being less susceptible to noise.

The GWR network is composed of neurons associated with a weight vector and the edges that link the neurons in order to form neighborhood relationships. The initialization phase sees the network with a set of two neurons
randomly initialized from within the training data. Both neurons and edges can be created and removed during each learning iteration. At each learning iteration, given an input data sample $x(t)$, the index $b$ of the best-matching unit (BMU) is given by:

$$b = \arg \min_{j \in W} ||x(t) - w_j||,$$

where $w_j$ is the weight vector of the $j$th neuron and $W$ is the set of all weight vectors. The activity of the network $a$ is computed as a function of the Euclidean distance between the weight of the best-matching unit $w_b$ and the input data sample $x(t)$ at time step $t$:

$$a = \exp(-||x(t) - w_b||).$$

New neurons are added when the activity of the best-matching unit is lower than a predefined threshold, named insertion threshold $a_T$. This parameter modulates the amount of generalization, i.e. the discrepancy between an incoming stimulus and its best-matching unit. Following the Hebbian learning mechanism [10], edges are created between two neurons with the smallest distance from the input data sample, namely the first and the second best-matching unit. As a consequence, after a number of learning iterations two neurons with an existing edge may end up far from each other, thereby not representing similar perceptions. An edge aging mechanism together with a threshold $a_{max}$ takes care of removing such edges and eliminating unconnected neurons consequently. Moreover, a firing rate mechanism which measures how often each neuron has matched the input guarantees sufficient training before new neurons are created. This firing rate variable is initially set to zero and than decreases every time a neuron and its neighbors are trained in the following way:

$$\Delta h_i = \tau_i \cdot \kappa \cdot (1 - h_i) - \tau_i,$$

where $\tau_i$, and $\kappa$ are the constants controlling the behaviour of the decreasing function curve of the firing counter. Usually, the $\tau$ constant is set higher for the best-matching unit, $\tau_b$, than for its topological neighbors, $\tau_i$, so that the firing counter decreases faster for the BMU. Given an input data sample $x(t)$, if no new neurons are added, the weights of the winner neuron and its neighbors are updated as follows:

$$\Delta w_i = \epsilon_i \cdot h_i \cdot (x(t) - w_i),$$

where $\epsilon_i$ and $h_i$ are the constant learning rate and the firing counter variable. The learning of the GWR algorithm stops when a given criterion is met, e.g., the maximum network size or the maximum number of learning epochs.

3.2. Hierarchical learning

We adopt hierarchical GWR learning [18] for the data processing and subsequent action-object integration. Hierarchical training is carried out layer-wise and in an offline manner by applying a batch-learning strategy. We first extract body pose, $A$, and object features, $O$, from the training image sequences, $T$, as described in Section 3.4. The obtained data is processed by training the first layer of the proposed architecture, i.e. $GWR_b$ is trained with body pose data and $GWR_o$ with objects (Fig. 1). After training is completed, the $GWR_b$ network will have created a set of neurons tuned to prototype body pose configurations, and the $GWR_o$ network will have learned to classify objects appearing in each action sequence.

The next step is to generate a new dataset $T^*$ for the $GWR_a$ network that integrates information coming from
both streams (Fig. 2). In order to encode spatiotemporal dependencies within the body pose prototypes space, we compute trajectories of the \textit{GWR}_b best-matching units when having as input training action sequences. For all body pose frames \( x_i \in A \), the best-matching units are calculated as in Eq. \([1]\) and the corresponding neuron weights are concatenated following a temporal sliding window technique, as follows:

\[
\psi(x_i) = b(x_i) \oplus b(x_{i-1}) \oplus \ldots \oplus b(x_{i-q+1}), i \in [q,m],
\]

where \( \oplus \) denotes the concatenation operation, \( m \) is the total number of training frames, and \( q \) is the width of the time window. We will refer to the computed \( \psi(x_i) \) by the name \textit{action segment}.

The object data \( y \in O \) extracted from each action sequence is provided as input to the \textit{GWR}_a network and the best-matching units \( b(y) \) are calculated as in Eq. \([1]\). Objects are extracted only at the beginning of an action sequence. Therefore, the object representations to be learned contain no temporal information and the computation of neural activation trajectories, reported in Eq. \([5]\) is not performed. The label of the \textit{GWR}_a best-matching unit is represented in the form of 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 \textit{GWR}_a 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, the new dataset \( T^* \) is computed by concatenating each action segment \( \psi(x_i) \) with the label of the corresponding object \( \ell_o(y) \) as follows:

\[
T^* = \{ \phi_u(x_i) \equiv \psi(x_i) \oplus \ell_o(y); x_i \in A, y \in O, u \in [q..m-q] \}.
\]

Each pair \( \phi_u \), which we will refer to as an \textit{action-object} segment, encodes both temporally-ordered body pose sequences and the identity of the object being manipulated during the action sequence. The \textit{GWR}_a network is then trained with the newly computed dataset \( T^* \), thereby learning the provided action-object pairs.

The resulting representative vectors of body pose can have a very high dimension which further increases when concatenating them through the temporal window technique. Methods based on the Euclidean distance metric, as in our case, are shown to have a performance degradation when data lies in high-dimensional space \([17]\). Therefore, we apply the principal component analysis (PCA) dimensionality reduction technique to the neural weights of \textit{GWR}_b. The number of principal components is chosen as a trade-off between accounting for the greatest variance in the set of weights and having a smaller dimensional discrepancy with the object’s label. The new basis is then used to project weights of activated neurons in \textit{GWR}_b before the concatenation of the activation trajectories and the subsequent integration step.

3.3. Classification

We extend the GWR algorithm with a labeling strategy in order to solve classification tasks while keeping the learning process unsupervised. For this purpose, we use a simple method based on the majority vote strategy as in \([18]\). For each neuron \( n_i \), we store information about the category of the data points it has matched during the training phase. Thus, each neuron is associated with a histogram \( \text{hist}(c,n_i) \) counting all cases of seeing a sequence with an assigned specific label \( c \). Additionally, the histograms are normalized by scaling the bins with the corresponding inverse class frequency \( f_c \) and with the inverse neuron activation frequency \( f_{a,n_i} \). In this way, class labels that appear less during training are not penalized, and the vote of the neurons is weighted equally regardless of how often they have fired. When the training phase is complete, each neuron that has fired during training, i.e. BMUs, will be associated with a histogram:

\[
H_i = \frac{1}{f_c \cdot f_{a,n_i}} \cdot \text{hist}(c,n_i).
\]

At recognition time, given a test action sequence with length \( k \), the best-matching units are computed for each frame and the action label \( l \) is given by:

\[
l = \arg \max_c \left( \sum_{i=1}^{k} H_{b_i} \right).
\]

The classification of non-temporal data, e.g. object classification with the \textit{GWR}_a network, is performed by applying majority vote only on the histogram associated to one best-matching unit \( H_{b_m} \). This is a special case of Eq. \([8]\) considering that \( k = 1 \) for non-temporal data.

In our case, action sequences are composed of smaller action-object segments as described in Section \([5.2]\). Thus,
the majority vote labeling technique described so far is applied in the following way. Let us assume we have a set of activity labels \( L_a \) along with our training data, for instance, \textit{drinking} and \textit{eating}. Therefore, each action-object segment \( \phi \in T^* \) will be assigned with one of these labels and one action sequence will have the following form:

\[
\Phi = \{ (\phi_1, l_1^a), \ldots, (\phi_k, l_k^a), l^a \in L_a \},
\]

where \( l^a \) is the activity label and \( k \) is the number of action-object segments included in the sequence. During training of the \( GWR_n \) network on the action sequence \( \Phi \), the label \( l_j^a \) will be added to the histogram of the neurons activated for each of its composing segment \( \phi \). After the training is complete, the action sequence \( \Phi \) will be classified according to the majority vote strategy (see Fig. 2). It should be noted that the association of neurons with symbolic labels does not affect in any way the formation of topological arrangements in the network. Therefore, our approach for the classification of objects and actions remains unsupervised.

3.4. Feature extraction

3.4.1. Body pose features

Visual identification and segmentation of body pose from RGB videos are challenging due to the spatial transformations compromising the appearances, such as translations, the difference in the point of view, changes in ambient illumination, and occlusions. For this reason, we use depth sensor technologies, such as the Asus Xtion camera, which provide us with reliable estimations of three-dimensional articulated body pose and motion even in real-world environments. Moreover, three-dimensional skeletal representations are the most straightforward way of achieving invariance to the subjects’ appearance and body size. We consider only the position of the upper body joints (shoulders, elbows, hands, center of torso, neck and head), given that they carry more significant information (than for instance the feet and knee joints) about the human-object interactions we focus on in this paper.

We extract the \textit{skeletal quad} features \cite{39}, which are invariant with respect to location, viewpoint as well as body-orientation. These features are built upon the concept of geometric hashing and have shown promising results for the recognition of actions and hand gestures. Given a quadruple of body joints \( \{J_1, J_2, J_3, J_4\} \) where \( J_i \in \mathbb{R}^3 \), a local coordinate system is built by making \( J_1 \) the origin and mapping \( J_2 \) onto the vector \( [1, 1, 1]^T \). The position of the other two joints \( J_3 \) and \( J_4 \) are calculated with respect to the new local coordinate system and are concatenated in a 6-dimensional vector \( [J_{3,1}, J_{3,2}, J_{3,3}, J_{4,1}, J_{4,2}, J_{4,3}] \). The latter becomes the compact representation of the four body joints’ position. We empirically select two quadruples of joints: \textit{[center torso, neck, left hand, left elbow]} and \textit{[center torso, neck, right hand, right elbow]}. This means that the positions of the hands and elbows are encoded with respect to the torso center and neck. We choose the neck joints’ position. We empirically select two quadruples of joints: \textit{[center torso, neck, left hand, left elbow]} and \textit{[center torso, neck, right hand, right elbow]}. This means that the positions of the hands and elbows are encoded with respect to the torso center and neck. We choose the neck instead of the head position due to the noisy tracking of the head caused by occlusions during actions such as \textit{eating} and \textit{drinking}.

3.4.2. Object features

The natural variations in RGB images such as variations in size, rotation, and lighting conditions, are usually so wide that objects cannot be compared to each other simply based on the images’ pixel intensities. For this reason, we extract visual features from the object images in the following way. We extract dense SIFT features, which are not more than SIFT descriptors \cite{38} computed at crossing points of fixed grids superimposed on each object image\footnote{Dense SIFT from VLFeat library: \url{http://www.vlfeat.org/}}. SIFT features have been successfully applied to the problem of unsupervised object classification \cite{50} and for learning approaches based on self-organization \cite{31}. Moreover, SIFT descriptors are known to be, to some extent, robust to changes in illumination and image distortion. Multiple descriptors with four different window sizes are computed on every image in order to account for scale invariance between images. The orientation of each of these descriptors is fixed and this relaxes the descriptors’ invariance with respect to the object’s rotation. With this kind of representation, we can train the \( GWR_n \) network and obtain neurons tuned to different object views, yet invariant to translation and scale.

We perform quantization followed by an image encoding step in order to have a fixed-dimensional vectorial representation of each object image. This is necessary since, during training of the \( GWR_n \) network, the objects are compared to each other through a vectorial metric, namely the Euclidean distance. We apply the Vector of Locally Aggregated Descriptors (VLAD) \cite{52} encoding method (Fig. 3) which has shown higher discriminative power than the extensively used Bag of Visual Features (BoF) \cite{53, 54}. The BoF method simply computes a histogram of the local descriptors by hard assignment to a dictionary of visual words, whereas the VLAD method computes and traces
the differences of all local descriptors assigned to each visual word.

3.5. Training

In Table 1, we report the parameters used for training the proposed neural architecture throughout the experiments presented in Section 4. The selection of the range of parameters is made empirically while also considering the GWR algorithm learning factors. The parameters that we fix across all layers are the constants controlling the decrease function of the firing rate variable ($\tau\varepsilon_i$ and $\tau\varepsilon_o$), the learning rates for the weights’ update function ($\epsilon_o$ and $\epsilon_i$) and the threshold for the maximum age of the edges ($a_{\text{max}}$). We set a higher insertion threshold parameter for the data processing layers, i.e. $GWR_b$ and $GWR_o$, than for the integration layer $GWR_a$. The higher value chosen for the $GWR_b$ and $GWR_o$ networks leads to a greater number of neurons created and a better representation of the input data as a result, whereas the slightly lower value for the $GWR_a$ seeks to generate a set of neurons that tolerate more discrepancy in the input and generalize relatively more. The insertion threshold parameters are very close to each other and very close to 1, but their impact is not imperceptible given that the input data are normalized, i.e. take values within the interval $[0,1]$. We train each network for 300 epochs over the whole dataset in order to ensure convergence, during which the response of the networks to the input shows little to no significant modifications.

In addition to the aforementioned parameters, the sliding window mechanism applied to processed body pose data also has an impact on the growth of the $GWR_a$ network. Wider windows lead to the creation of more neurons, albeit the slightly lower number of data samples. This is an understandable consequence of the fact that the more temporal frames included in each time window, the higher the variance of the resulting data and the more prototype neurons created as a consequence. However, this parameter has to be set empirically according to the experimental training data distribution. We report the time window width parameter we set in each of our experiments in the following sections.

Table 1: Training parameters of the $GWR_b$, $GWR_o$ and $GWR_a$ networks of our architecture for the classification of human-object interactions.

| Parameter                  | Value |
|---------------------------|-------|
| Insertion Threshold       | $a_T = \{0.98, 0.98, 0.9\}$ |
| Firing Threshold          | $f_T = 0.1$ |
| Learning rates            | $\epsilon_o = 0.1, \epsilon_i = 0.01$ |
| Firing rate behavior      | $\tau_o = 0.3, \tau_i = 0.1, \kappa = 1.05$ |
| Maximum edge age          | $a_{\text{max}} = 100$ |
| Training epochs           | 300   |

4. Experimental results

We evaluated the proposed neural architecture both on the transitive actions dataset (Fig. 3) that we have acquired for the purpose of this study and on a publicly available action benchmark dataset provided by the Cornell University, CAD-120 [23]. In this section, we provide details on both datasets, the classification performances obtained on these datasets, a quantitative evaluation of the integration module in the case of incongruent action-object pairs and a comparative evaluation on CAD-120.

4.1. Experiments with the transitive actions dataset

4.1.1. Data collection

We collected a dataset of the following daily activities: picking up (an object), drinking (from a container like a mug or a can), eating (an object like a cookie) and talking on the phone (Fig. 4). The actions were performed by 6 participants that were given no explicit indication of the purpose of the study nor instructions on how to perform the actions. The dataset was collected with an Asus Xtion depth sensor that provides synchronized RGB and depth frames at a frame rate of 30 fps. The distance of each participant from the sensor was not fixed but maintained within the maximum range for the proper functioning of the depth sensor. The tracking of the skeleton joints was provided by the OpenNI framework\footnote{OpenNI/NITE: \url{http://www.openni.org/software}}. To attenuate noise, we computed the median value for each body joint every 3 frames resulting in 10 joint position vectors per second. We added a mirrored version of all action samples to obtain invariance to actions performed with either the right or the left hand. Action labels were then manually annotated.
The manipulated objects were segmented from each video using a point-cloud-based table-top segmentation algorithm\(^3\) which extracts possible clusters on top of a plane surface, e.g., on the table. False positives obtained through the automatic segmentation were then manually deleted. Finally, the obtained images were used as training data for the object recognition module of our architecture.

### 4.1.2. Classification results

We now assess the performance of the proposed neural architecture for the classification of the actions described in Section 4.1.1. In particular, we want to evaluate the importance of the identity of the manipulated object(s) in disambiguating the activity that a subject performs. For this purpose, we conducted two separate experiments, whereby we process body pose cues alone and in combination with recognized objects. Moreover, to further exclude any possible bias towards a particular subject, we followed a leave-one-subject-out strategy. Therefore, six different trials were designed by using video sequences of the first five subjects for training and using the remaining subject for the testing phase. This type of cross-validation is quite challenging since different subjects perform the same action in a different manner and with a different velocity.

We trained each GWR network with the learning parameters reported in Section 3.5. Since this dataset is composed of short temporal sequences, a time window of five frames was chosen for the concatenation of the processed body cues. This led to action-object segments of 0.5 seconds, considering 10 frames per second. When the training of the whole architecture was complete, the number of neurons reached for an input containing \(\approx 6500\) video frames was: 170 neurons for the \(GWR_a\) network, 182 for \(GWR_o\) and for the \(GWR_o\) network the number varied from 90 to 120 across different trials.

A plot showing the neural weights of the \(GWR_o\) network is depicted in Fig. 5. As it can be seen from the plot, the neurons have been topologically organized into clusters composed of different 2D views of the objects as well as different instances of the same object category. This is quite advantageous for our architecture since it allows for generalization towards unseen object views and to some extent towards unseen object instances. The overlap between the \(can\) and \(mug\) clusters suggests that the visual appearance of these object categories is more similar than compared to the others and, as a consequence, can be confused. However, this does not affect the action classification performance, since both of the objects are involved in the same activity, namely drinking.

We report precision, recall, and F1-score \(^5\) for each class of activity, averaged over all six trials in Fig. 6. We obtained values equal to 100% when using the objects’ identity information and lower percentage values when using only body pose. As expected, the increase of the classification performance is more significant for those cases where the sole body pose introduces ambiguity, e.g., drinking, eating, talking on the phone. For the picking up activity, on the other hand, the difference in the classification performance is minor, due to the fact that this action can be performed on all of the objects and the identity of a specific object does not play a decisive role.

### 4.1.3. Experiments with incongruent action-object pairs

In addition to the classification experiments, we carried out a qualitative evaluation of the integration module when given in input test data sequences of incongruent action-object pairs. We consider incongruent pairs to be conceptually unusual combinations of actions with objects, e.g., drinking with a telephone or eating with a can. Interestingly, fMRI studies on human brain have found several regions affected by object-action congruence \(^4\). The neural response in these areas is greater for actions performed on appropriate objects as opposed to unusual actions performed on the same objects. For this experiment, we artificially created a test dataset, for which we replaced the image of the object being manipulated in each video sequence with the image of an incongruent object extracted from a different action video.

\(^3\)Point Cloud Library: http://www.pointclouds.org/

\(^5\)Precision, Recall, F1-score.
We analyzed the activation values of the GWR\(_a\) BMUs, computed as in Eq. 2, on both the original action sequence and the manipulated one. A few examples of the obtained neural activations are illustrated in Fig. 7. We observed that, typically, the activations were relatively low for the incongruent samples. This can be explained by the fact that the GWR\(_a\) prototypes represent the joint distribution of action segments and congruent objects taken from the congruent set. The activation of the network is expected to be lower when the input has been taken from a different data distribution than the one the model has learned to fit. The incongruent samples can have a higher Euclidean distance with the prototype neuron weights, thereby leading to a lower network activation.

We also noticed some exceptions, e.g., the incongruent pair <talking on the phone, can> depicted in Fig. 7(c). In this case, we can observe that the network activation becomes higher for the incongruent input at a certain point of the sequence, i.e. at a certain action-object segment. However, the drop of the network activation on the congruent input indicates that the network has a high quantization error for that particular action-object segment. It should be noted that a small quantization error of the GWR network is not a requirement for a good performance in the action classification task. As described in Section 3.3, the classification of an action sequence is performed by considering the label histograms associated with the activated neurons. We can also notice some cases where the network activation on the incongruent input is not significantly low at the beginning of the sequence, but even slightly higher in the case of <eating, phone> (Fig. 7(b)). A reason for this is the similar motion of the hand holding the object towards the head which may precede both eating and talking on the phone activities. Therefore, exchanging the object biscuit box with phone for the initial action segments has from little to no impact on the network’s response.

### 4.2. Experiments with CAD-120

We evaluated the classification capabilities of our architecture on a publicly available benchmark dataset provided by Cornell University, CAD-120 (Fig. 8). This dataset consists of 120 RGB-D videos of 10 long daily activities: arranging objects, cleaning objects, having meal, making cereal, microwaving food, picking objects, stacking objects, taking food, taking medicine and unstacking objects. These activities are performed by four different subjects (two males, two females and, of these four, one left-handed) repeating each action three to four times. Each video is annotated with the human skeleton tracks and the position of the manipulated objects across frames.

We computed skeletal quad features (described in Section 3.2) for the encoding of the pose of the upper body, based on the three-dimensional position of skeletal joints provided in the dataset. Additionally, we extracted RGB images of manipulated objects from each frame and en-
For the concatenation of the processed body pose cues, a time window of nine frames was chosen. Since we down-sample the activity video frames to a rate of 10 fps, this leads to an action-object segment having a temporal duration of 0.9 seconds. After training the whole architecture with an input data of $\approx 18,000$ frames, the number of neurons reached in each GWR network was 460 for $\text{GWR}_b$, 410 for $\text{GWR}_c$, while for $\text{GWR}_a$ the number varied from $\approx 3200$ to $\approx 3700$ across different trials of the cross-validation.

In Fig. 9, we show the confusion matrix for the 10 high-level activities of this dataset. We inspected that the activities interchanged by our model were the ones including the same category of objects and similar body motions, e.g., \textit{stacking objects} and \textit{unstacking objects}, \textit{microwaving food} and \textit{taking food}. Also, the activity of \textit{picking objects} was often confused with \textit{arranging objects}, due to the fact that body pose segments of the first are similar to the ones preceding the activity of \textit{arranging objects}. In Table 2 we show a comparison of our results with the state of the art on the CAD-120 dataset with accuracy, precision, and recall as evaluation metrics. We obtained 79% of accuracy, 80.5% of precision, and 78.5% of recall.

We reported only the approaches that do not make use of the ground-truth temporal segmentation of the activities into smaller atomic actions, called sub-activities. The fact that we classify the video of the whole activity without relying on the recognition of these sub-activities places our approach within this group of approaches for this dataset. Our results are comparable with the work from Rybok et al.\cite{55}. Similar to our work, their method considers objects’ appearance as contextual information which is then concatenated with body motion features represented as a bag of words. On the other hand, best state-of-the-art results from Koppula et al.\cite{56} reported 83.1% of accuracy, 87% of precision and 82.7% of recall. In their work, spatiotemporal dependencies between actions and objects are modelled by a Conditional Random Field (CRF) which combines and learns relationships between a number of different features such as the coordinates of the object’s centroid, the total displacement and the total distance moved by the object’s centroid in each temporal segment, the difference in ($x, y, z$) coordinates of the object and skeleton joint locations and their distances. After the generation of the graph which models spatiotemporal relations, they use a Support Vector Machine (SVM) for classifying action sequences. Unlike in our work, they do not perform object classification but rely on manually annotated ground truth labels.

We assume that the tracking of the objects’ position in the scene as well as the objects’ distance from the subject’s hand provides additional information that might improve our classification results and is considered part of our future work.

5. Discussion

5.1. Summary

In this paper, we presented a self-organizing neural network architecture that learns to recognize human-object interaction activities from RGB-D video sequences. Our architecture consists of two pathways of GWR networks processing respectively body pose and appearances of objects and an integration layer merging incoming information in order to recognize human-object interaction activities. The prototype-based learning mechanism of the GWR algorithm allows for input noise attenuation and for generalization towards unseen data samples. For classification purposes, we extended the GWR algorithm with a labeling technique based on majority vote.

The evaluation of our approach has shown good results on a dataset of human-object interactions collected specifically for the study of the importance of the identity of objects. The analysis of the neural response of the integration module showed an overall lower network activation

![Confusion Matrix](image)

Table 2: Results on the CAD-120 dataset for the recognition of 10 high-level activities. Reported are accuracy, precision and recall (in percentage) averaged over 4-fold cross-validation experiments. For comparison, we have included which of the reported methods is unsupervised (U), performs object recognition for the classification of the activities (O.Rec.) or relies on object tracking (O.Tr.).
when given incongruent action-object pairs as input compared to the congruent combinations. Furthermore, the classification accuracy of our architecture on a publicly available action benchmark dataset is similar to the state of the art. However, unlike the other related approaches, we used action labels only to evaluate the overall classification accuracy while leaving the learning unsupervised.

5.2. Self-organizing neural learning and analogies with neuroscience

Prototype-based generative approaches based on self-organization have the ability to learn input probability distribution through a finite set of reference vectors associated with neurons. Moreover, they are capable of reflecting the topological relationships of the input space through the neurons’ organization. Growing extensions of such approaches, such as GNG [45] and GWR networks [17] have a dynamic topological structure able to adapt toward the input data space through the mechanism of the competitive Hebbian learning [46]. Unlike the GNG, where the network grows at a constant rate, the GWR algorithm is equipped with a learning mechanism that creates new neurons whenever the current input representation is not sufficient.

We adapted the original implementation of the GWR algorithm, which processes input data vectors in the spatial domain, to the processing of temporal data by the mechanism of the temporal sliding window [10]. The temporally ordered neural activations obtained through this technique resemble the motion pattern encodings through the snapshot neurons found in the STS area of the brain [10]. There is also neurophysiological evidence that actions are represented by sequences of poses over fixed temporal windows [58]. From the computational perspective, the sliding window technique allows for the extrapolation of spatiotemporal dependencies in the data sequences. The use of prototype-based representations for objects, on the other hand, is also motivated by psychological studies on the nature of human categorization [10]. According to the exemplar-based theory, categories of objects and concepts are typically learned as a set of prototypical examples and the similarity, or the so-called family resemblance, is used for class association.

Finally, the use of the GWR algorithm for integrating information about action and objects produced a behavior resembling the action-selective neural circuits which show sensitivity to the congruence of the action being performed on an object [14].

5.3. Future Work

In this work, we focused on a two-pathway hierarchy for learning human-object interactions represented as a combination of upper body pose configurations and objects’ category labels. However, in order to reduce the computational complexity of the architecture, we have excluded an important component: the motion information. Results from other approaches on recognition of human-object interactions and on the learning of object affordances [36, 23] have shown that tracking objects’ positions and spatial relationship with regards to body limbs can help for better interpretation of the type of interaction. There is evidence also from neuroscience that the observation of tool use in humans activates areas of the lateral temporal cortex which is engaged in perceiving and storing information about motion [5]. Neural mechanisms for the processing of human body motion are also believed to contribute to action discrimination in general [10]. Therefore, a logical next step is to extend our model by including motion and conduct further experiments.

An additional future work direction is the introduction of recurrent connections for the purpose of temporal sequence processing. Recurrence in self-organizing networks has been extensively investigated and applied to temporal data classification [48]. In the current implementation, temporal dependencies are encoded and learned by hard assignments to time windows. However, the concatenation of perceptual feature vectors may lead to very high-dimensional spaces, whereby methods based on a Euclidean distance metric are known to perform worse [47].

In our current work, we used depth information for the efficient extraction of a three-dimensional skeleton model. However, when dealing with more complex activities such as human-object interactions, this type of depth representation may be subject to a number of issues such as the estimated skeleton becoming highly noisy due to body self-occlusions or when the person touches objects in the background. Moreover, skeletal representations lack information about spatial relationships with objects. Therefore, future work should address the limitations of this handcrafted feature extraction with a neural architecture able to extract visual features from raw images, e.g., with the use of deep neural network self-organization [59].

Finally, the results reported in this paper motivate future work towards integration of our learning system into robotic platforms and its evaluation in real-world complex scenarios such as robots learning by imitation or even intelligent systems assisting humans in natural environments.

Acknowledgments

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

References

[1] F. Fleischer, V. Caggiano, P. Thier, M. A. Giese, Physiologically inspired model for the visual recognition of transitive hand actions, The Journal of Neuroscience 33 (15) (2013) 6563–6580. doi:10.1523/JNEUROSCI.4129-12.2013
[41] J. Wang, Z. Liu, Y. Wu, Learning actionlet ensemble for 3D human action recognition, in: Human Action Recognition with Depth Cameras, Springer International Publishing, 2014, pp. 11–40. doi:10.1109/TPAMI.2013.198.

[42] T. Poggio, S. Edelman, A network that learns to recognize three-dimensional objects, Nature 343 (6255) (1990) 263–266. doi:10.1038/343263a0.

[43] D. Perrett, View-dependent coding in the ventral stream and its consequences for recognition, Vision and Movement Mechanisms in the Cerebral Cortex (1996) 142–51.

[44] T. Martinetz, K. Schulten, A “neural-gas” network learns topologies, in: Artificial Neural Networks, Elsevier Science Publisher B.V., 1991, pp. 397–402.

[45] B. Fritzke, A growing neural gas network learns topologies, Advances in Neural Information Processing Systems 7 (1995) 625–632.

[46] T. Martinetz, Competitive hebbian learning rule forms perfectly topology preserving maps, in: International Conference on Artificial Neural Networks (ICANN), Springer, 1993, pp. 427–434. doi:10.1007/978-1-4471-2063-6_104.

[47] C. C. Aggarwal, A. Hinneburg, D. A. Keim, On the surprising behavior of distance metrics in high dimensional space, in: International Conference on Database Theory (ICDT), Springer, 2001, pp. 420–434. doi:10.1007/3-540-44503-X_27.

[48] G. Evangelidis, G. Singh, R. Horan, Skeletal quads: Human action recognition using joint quadruples, in: International Conference on Pattern Recognition (ICPR), 2014, pp. 4513–4518. doi:10.1109/ICPR.2014.772.

[49] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, C. Schmid, Aggregating local image descriptors into compact codes, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (9) (2012) 1704–1716. doi:10.1109/TPAMI.2011.235.

[50] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, International Journal of Computer Vision 88 (2) (2010) 303–338. doi:10.1007/s11263-009-0275-4.

[51] G. I. Parisi, S. Wermter, Lifelong learning of action representations with deep neural self-organization, Tech. rep., AAAI Spring Symposium Series, 608-612 (Mar 2017).