Neurocomputing 307 (2018) 14–24

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

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

ARTICLE INFO

Article history:
Received 6 December 2016
Revised 1 February 2018
Accepted 16 April 2018
Available online 8 May 2018

Communicated by Dr. Xu Zhao

Keywords:
Self-organization
Hierarchical learning
Action recognition
Object recognition
Human-object interaction

ABSTRACT

The visual recognition of transitive actions comprising human-object interactions is a key component for artificial systems operating in natural environments. This challenging task requires jointly the recognition of articulated body actions as well as 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 that learn prototypical representations of body motion patterns and objects, accounting for the development of action-object mappings in an unsupervised fashion. 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 exhibits higher neural activation for congruent action-object pairs learned during training sessions with respect to synthetically created incongruent ones. We show that our unsupervised model shows competitive classification results on the benchmark dataset with respect to strictly supervised approaches.

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 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 ability of computational approaches to reliably recognize human-object interactions can establish an effective cooperation between assistive systems and people in real-world scenarios, promoting 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 the representations of body postures and manipulated objects and, in particular, how these two representations interact and can be integrated.

In the visual system, the information about body pose and objects are processed separately and reside in distinct subcortical areas [5–7]. 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 [12–14]. Yet, little is known about the exact neural mechanisms underlying the integration of actions and objects.

In this paper, we present a self-organizing neural architecture that learns to recognize human-object interactions from RGB-D videos. The design of the proposed architecture relies on the following assumptions: (i) the visual features of body pose and man-made objects are represented in two distinct areas of the brain [5–7], (ii) input-driven self-organization defines the topological structure of specific visual areas in brain [15], (iii) the representation of objects and concepts is based on prototypical examples

* Corresponding author.
E-mail addresses: mici@informatik.uni-hamburg.de (L. Mici), parisi@informatik.uni-hamburg.de (G.I. Parisi), wermter@informatik.uni-hamburg.de (S. Wermter).

https://doi.org/10.1016/j.neucom.2018.04.015

[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 clustering human motion patterns in terms of multi-dimensional flow vectors [18,19] as well as for learning object representations without supervision [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. [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) [22]. The reason for this is the considerable impact on using a predefined topological structure [23], especially when having as input high-dimensional complex data distributions like perceptual representations of objects. In our previous model, an additional network was used to learn prototypes of temporal activation trajectories of body poses before the integration phase. However, the impact on the overall classification accuracy of the network was marginal, while it introduces 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 [24]. 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 performance of our architecture in terms of recognition of human-object interaction activities, and the response of the network when fed with congruent and 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 [25–27]. 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 *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. [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 [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 [28–30]. Typically, first, objects are recognized and activities involving them are subsequently recognized, by analyzing the objects' motion trajectories [31]. Yang et al. [32] 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 $<$*Object1, Action, Object2*$>$ extracted from the sentences in the English Gigaword corpus. Pieropan et al. [33] proposed including action-related audio cues in addition to the spatial relation among objects in order to learn object manipulations for the purpose of robot learning by imitation. However, important descriptive visual features like body motion or fine-grained cues like the hand pose during manipulation were not considered.

Probabilistic approaches have been extensively used for reasoning upon relationships and dependencies among objects, motion, and human activities. Gupta et al. [34,35] 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 [36] 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 Random Fields (CRF) [24,37,38]. These types of models suffer from high computational complexity and require a fine-grained segmentation of the action sequences.

Motivated by the fact that the visual recognition of complex human poses and the identification of objects in realistic scenes are extremely hard tasks, additional methods rely on extracting novel low-level visual features. Yao and Fei-Fei [39] proposed a set of sophisticated visual features called *Grouplet* which captures spatial organization of image patches encoded through SIFT descriptors [40]. Their method is able to distinguish between interactions or just co-occurrences of humans and objects in an image, but no applications on video data have been reported. Aksoy et al. [41] proposed the *semantic event chains* (SEC): a matrix whose entries represent the spatial relation between extracted image segments for every video frame. Action classification is obtained in an unsupervised way through maximal similarity. While this method is suitable for teaching object manipulation commands to robots, the representation of the visual stimuli does not allow for reasoning upon semantic aspects such as the congruence of the action being performed on a certain object.

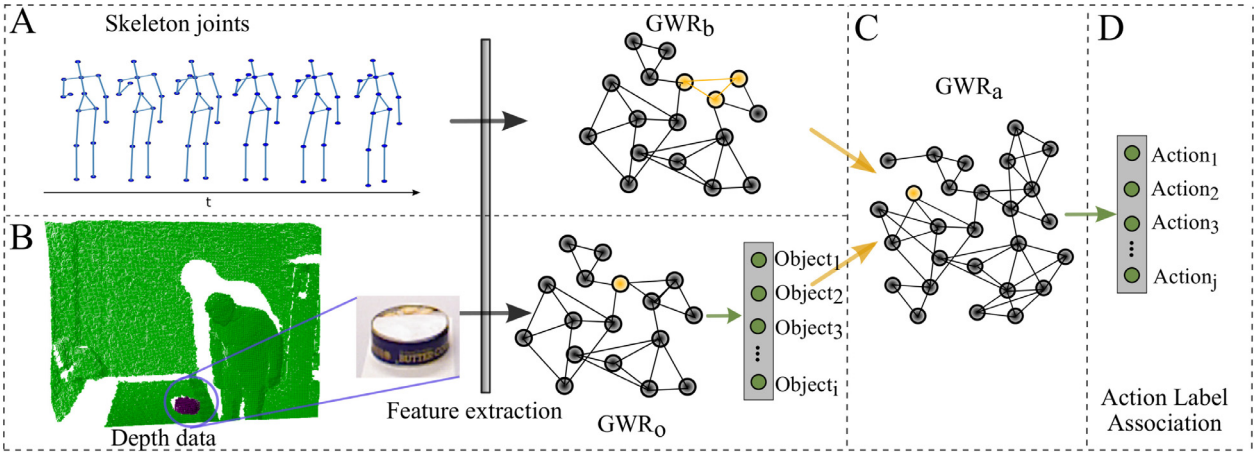

Fig. 1. Overview of the proposed architecture. (A) Processing of the body postures. A set of local features that encode the posture of upper body limbs is extracted and fed to the GWR_b networks. (B) The input for the object recognition module is the RGB image of the manipulated object. The region of interest is automatically extracted through a point-cloud-based table-top segmentation. The object is represented as a compact feature vector and is fed to the GWR_o network which classifies the object. (C) The last network, GWR_a, learns the combinations of body postures and the object(s) involved in an action. (D) Action labels are associated with each neuron in the GWR_a network in order to evaluate the architecture's action classification performance.

Early attempts to apply neural networks for the problem of understanding human-object interactions from visual perception yielded promising results. Shimozaki and Kuniyoshi [42] 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 [43,44].

Systems for the estimation of articulated human body pose from 2D image sequences struggle through a great number of challenges such as changes in ambient illumination, occlusion of body parts and the enduring problem of segmentation. The combination of RGB with depth information, provided by low-cost depth sensing devices such as Microsoft Kinect and Asus Xtion cameras, has shown computational efficiency in sensory data processing and has boosted a number of vision-based applications [45]. This sensor technology provides depth measurements which are used to obtain reliable estimations of 3D human body pose and tracking of body limbs in cluttered environments. Applications of this type of technology have led to the successful classification of full-body actions and recognition of hand gestures [27]. However, a limitation of skeletal features is the lack of information about surrounding objects. Wang et al. [46] proposed a new 3D appearance feature called *local occupancy pattern* (LOP) describing the depth appearance in the neighborhood of a 3D joint, and thus capturing the relations between the human body parts, e.g. hands, and the environmental objects that the person is interacting with. Although their method produces state-of-the-art results, the identity of the objects is completely ignored, and the discriminative power of such features is unclear when the objects being manipulated are small or partially occluded.

3. Methodology

The proposed architecture consists of two main network streams processing separately visual representations of the body postures and of the manipulated objects. The information 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 neurons 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_b, with a sliding time window technique. The object appearance cue is processed in order to have topological arrangements in GWR_o where different 2D views of 3D objects as well as different instances of the same object category are mapped to proximal neurons in the prototypes domain. The advantage of having such a topological arrangement consists in mapping any unseen view of a known object into the corresponding views learned during the training. This capability resembles, to some extent, biological mechanisms for learning the three-dimensional objects in the human brain [7,47,48]. 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 prototype neurons in GWR_a by extending the GWR algorithm with a labeling strategy.

3.1. Learning with the GWR algorithm

Input-driven self-organization is an unsupervised mechanism that learns the input probability distribution through a finite set of prototype vectors. Unlike traditional vector quantization (VQ) methods, self-organizing neural networks such as the SOM [22], the neural gas (NG) [49] as well as their growing extensions such as the growing neural gas (GNG) [50] and the GWR algorithm [17], associate these prototype vectors with neurons that adaptively form topology 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 process of topology preservation is motivated by similar neural mechanisms found in multiple cortical areas of the brainx [15].

Growing self-organizing networks learn incrementally and can add (or remove) neurons according to different criteria. Unlike the GNG algorithm, the neural growth of the GWR algorithm is not constant but rather 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 a set of neurons associated with a weight vector and a set of edges that link the neurons forming neighborhood relationships. The network starts with a set of two neurons randomly initialized and, during the learning iterations, both neurons and edges can be created, updated, or removed. Given an input data sample $\mathbf{x}(t)$, the index b of the best-matching unit (BMU) is given by:

$$b = arg\min_{j \in W} ||\mathbf{x}(t) - \mathbf{w}_j||, \tag{1}$$

where $\mathbf{w}_j$ is the weight vector of the neuron j and W is the set of all neurons. The activity of the network a is computed as a function of the Euclidean distance between the BMU, $\mathbf{w}_b$, and the input data sample $\mathbf{x}(t)$ at time step t:

$$a = \exp(-||\mathbf{x}(t) - \mathbf{w}_b||). \tag{2}$$

New neurons are added when the activity of the BMU is lower than an insertion threshold a_T. This parameter modulates the amount of generalization, i.e., the largest discrepancy between an incoming stimulus and its BMU. Edges are created between two neurons with the smallest distance from the input data sample, namely the first and the second BMUs. Rarely activated edges (after a_{max} iterations) and neurons without edges are removed. Moreover, a firing rate mechanism that measures how often each neuron has been activated by the input leads to a sufficient training before new neurons are created. The firing rate is initially set to zero and than decreases every time a neuron and its neighbors are activated in the following way:

$$\Delta h_i = \tau_i \cdot \kappa \cdot (1 - h_i) - \tau_i, \tag{3}$$

where τ_i, and κ are the constants controlling the behaviour of the decreasing function curve of the firing counter. Typically, the τ constant is set higher for the BMU (τ_b) than for its topological neighbors (τ_i). Given an input data sample $\mathbf{x}(t)$, if no new neurons are added, the weights of the winner neuron and its neighbors are updated as follows:

$$\Delta \mathbf{w}_i = \epsilon_i \cdot h_i \cdot (\mathbf{x}(t) - \mathbf{w}_i), \tag{4}$$

where ϵ_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., a maximum network size or a 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 with batch learning. We first extract body pose, A, and object features, O, from the training image sequences, T (see 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 GWR_b best-matching units when having as input training action sequences. For all body pose frames $\mathbf{x}_i \in A$, the best-matching units are calculated as in Eq. (1) and the corresponding neuron weights

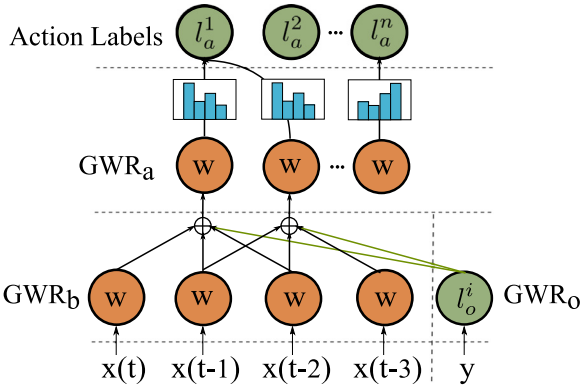

Fig. 2. Schematic description of the hierarchical learning and of the association of action labels (not all neurons and connections are shown). At each time step t, the input data sample $\mathbf{x}(t)$ is represented by the weight $\mathbf{w}$ of the winner neuron which is then concatenated with the 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 associated with a histogram of action categories, and the most frequently matched class will be the recognized action.

are concatenated following a temporal sliding window technique, as follows:

$$\psi(\mathbf{x}_i) = b(\mathbf{x}_i) \oplus b(\mathbf{x}_{i-1}) \oplus ... \oplus b(\mathbf{x}_{i-q+1}), i \in [q, m], \tag{5}$$

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 $\psi(\mathbf{x}_i)$ by the name *action segment*.

The object data $\mathbf{y} \in O$ extracted from each action sequence is provided as input to the GWR_o network and the best-matching units $b(\mathbf{y})$ 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 GWR_o 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 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, the new dataset T^* is computed by concatenating each action segment $\psi(\mathbf{x}_i)$ with the label of the corresponding object $l_o(\mathbf{y})$ as follows:

$$T^* = \{\phi_u(\mathbf{x}_i) \equiv \psi(\mathbf{x}_i) \oplus l_o(\mathbf{y}); \mathbf{x}_i \in A, \mathbf{y} \in O, u \in [q, m-q]\}. \tag{6}$$

Each pair ϕ_u, which we will refer to as an *action-object* segment, encodes both temporally-ordered body pose sequences and the identity of the object being manipulated during the action sequence. The 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 [51]. Therefore, we apply the principal component analysis (*PCA*) dimensionality reduction technique to the neural weights of 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 neu-

rons in 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 for classification tasks while keeping the learning process unsupervised. We use a simple method based on the majority vote strategy as in [52]. 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 $hist(c, n_i)$ 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 weighed 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:

$$H_i = \frac{1}{f_c \cdot f_{a,n_i}} \cdot hist(c, n_i). \tag{7}$$

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:

$$l = arg\max_c \left(\sum_{i=1}^{k} H_{b_i} \right). \tag{8}$$

The classification of non-temporal data, e.g. object classification with the GWR_o network, is performed by applying majority vote only on the histogram associated to one best-matching unit H_{bmu}. 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 3.2. Thus, the majority vote labeling technique described so far is applied as follows. Let us assume we have a set of activity labels L_a along with our training data, for instance, *drinking* and *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:

$$\Phi = \{(\phi_1, l_a^j), \ldots, (\phi_k, l_a^j), l_a^j \in L_a\}, \tag{9}$$

where l_a^j is the activity label and k is the number of action-object segments included in the sequence. During training of the GWR_a network on the action sequence Φ, the label l_j^a will be added to the histogram of the neurons activated for each of its composing segment ϕ. After the training is complete, the action sequence Φ 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 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. However, neither the number of considered joints nor the dimensionality of the input data limits the application of our architecture for the recognition of human-object interactions.

We extract the *skeletal quad* features [53], 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 local coordinate system and are concatenated in a 6-dimensional vector $[\hat{J}_{3,1}, \hat{J}_{3,2}, \hat{J}_{3,3}, \hat{J}_{4,1}, \hat{J}_{4,2}, \hat{J}_{4,3}]$. The latter becomes the compact representation of the four body joints' position. We empirically select two quadruples of joints: [*center torso, neck, left hand, left elbow*] and [*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 *eating* and *drinking*.

Composing such holistic body pose vectors, i.e., concatenations of joint positions, is quite convenient when employing a GWR network for the learning. In the case of missing joints in a data frame, due to, for example, noise or body occlusion, the best-matching unit for that input vector can be computed omitting the missing parts of the body pose vector. Self organizing networks, such as SOMs and the GWR networks as their growing extension, are able to operate robustly in the case of missing values [54].

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 [40] computed at crossing points of fixed grids superimposed on each object image[1]. SIFT features have been successfully applied to the problem of unsupervised object classification [55] and for learning approaches based on self-organization [56]. 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_o 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_o 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) [57] encoding method

[1] Dense SIFT from VLFeat library: http://www.vlfeat.org/.

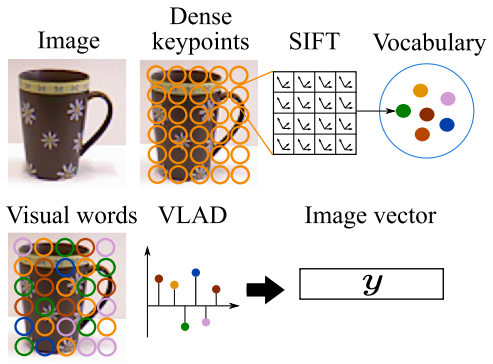

Fig. 3. Illustration of the steps for encoding object images with the VLAD encoding method.

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	$\epsilon_b = 0.1, \epsilon_i = 0.01$
Firing rate behavior	$\tau_b = 0.3, \tau_i = 0.1, \kappa = 1.05$
Maximum edge age	$a_{max} = 100$
Training epochs	300

(Fig. 3) which has shown higher discriminative power than the extensively used Bag of Visual Features (BoF) [58,59]. 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 (τ_b, τ_i and κ), the learning rates for the weights' update function (ϵ_b and ϵ_i) and the threshold for the maximum age of the edges (a_{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.

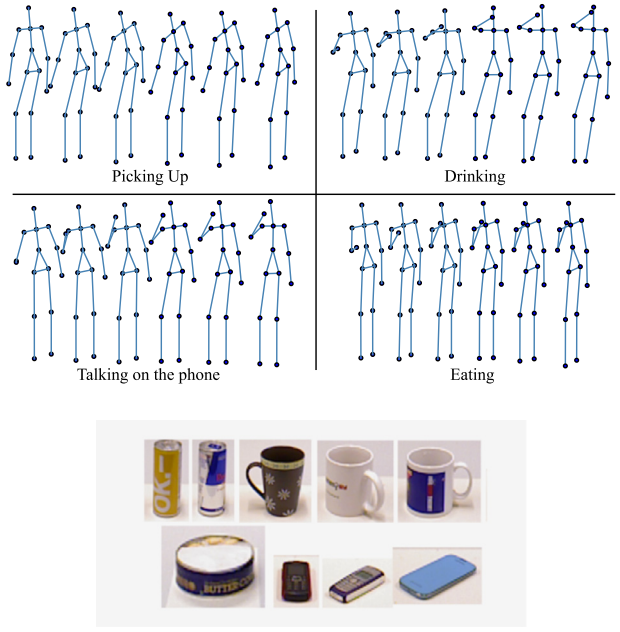

Fig. 4. Examples of sequences of skeleton joints and objects taken from the transitive actions dataset. The object category labels are: *can, mug, biscuit box* and *phone*.

4. Experimental results

We evaluated the proposed neural architecture both on the *transitive actions dataset* (Fig. 4) 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 [24]. 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[2]. 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

[2] OpenNI/NITE: http://www.openni.org/software.
[3] Point Cloud Library: http://www.pointclouds.org/.

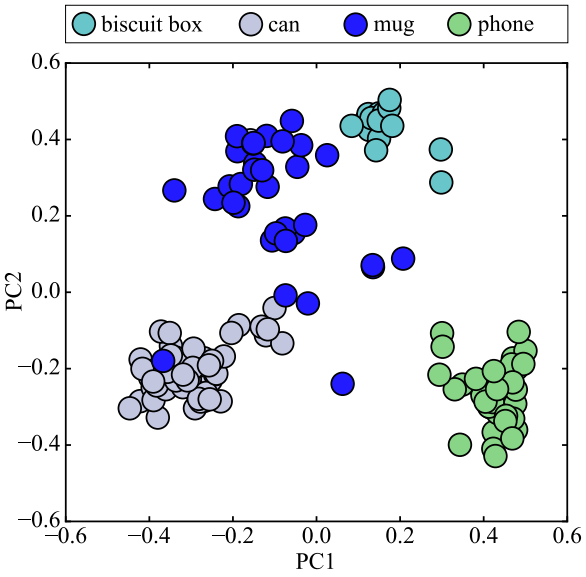

Fig. 5. Neural weights of the GWR_o network after having been trained with the objects from the transitive actions dataset. The first two principal components have been chosen for the visualization in two dimensions.

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_b network, 182 for GWR_o and for the GWR_a 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. Given that the neural weights have a high dimensionality, i.e., the dimensionality of the VLAD descriptors, for illustration purposes we performed principal component analysis (PCA) and show the first two principal components. As it can be seen from the plot, the neurons are 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*.

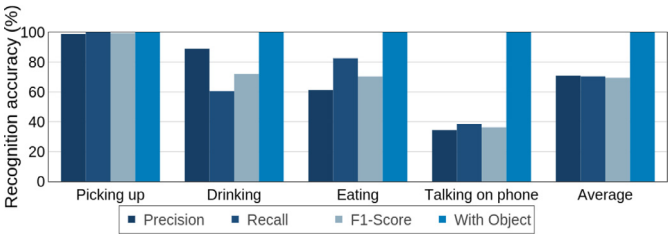

Fig. 6. Classification results on the transitive actions dataset. Illustrated are precision, recall, and F1-score, averaged over 6 trials of cross-validation when using only the body pose information. When using the manipulated object identity, given by the object recognition module GWR_o, we obtained a value of 100% for the reported classification performance measures.

We report precision, recall, and F1-score [60] 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 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 marginal 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 unusual or functionally irrelevant 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 [14]. 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.

We analyzed the activation values of the GWR_a BMUs (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 the activations were typically 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 yield a higher discrepancy with respect to the prototype neurons, thereby leading to a lower network activation. However, 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. Nevertheless, a decreased network activation on the congruent input indicates that the network has a high quantization error for that particular action-object segment.

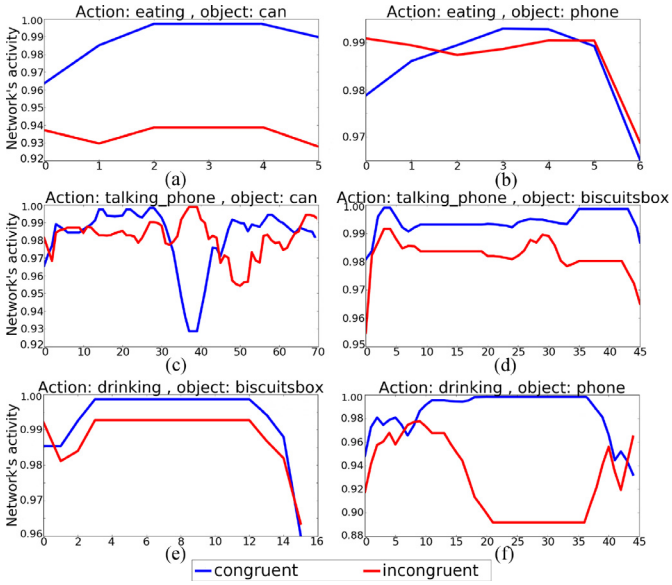

Fig. 7. Comparison of the GWR_a network activations when having as input an action sequence combined with an incongruent object (in red) and one combined with the congruent one (in blue). The y axis represents the activation values, with 1 being the highest, and the x axis represents the number of frames of the illustrated data sequences. The number of frames can vary among different actions, e.g. the action *eating* is typically shorter than *talking on the phone* and *drinking*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 8. Examples of high-level activities from CAD-120 dataset [24], ordered in each row: *microwaving food, taking food, stacking objects, unstacking objects, arranging objects, picking objects, cleaning objects, taking medicine*.

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 performance of our architecture on the publicly available benchmark dataset 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 encoded them through VLAD encoding technique as described in Section 3.2. 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 s. After training the whole architecture with an input data of

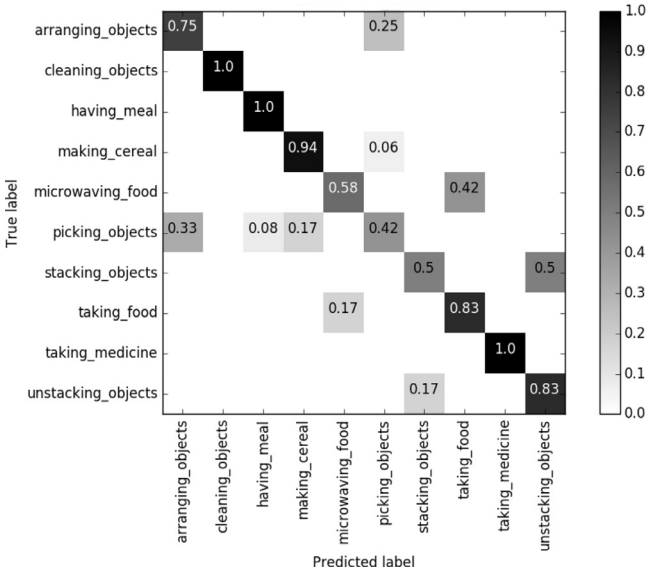

Fig. 9. Confusion matrix for the 10 high-level activities of CAD-120 dataset.

$\approx$18.000 frames, the number of neurons reached in each GWR network was 460 for GWR_b , 410 for GWR_o, while for 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., *stacking objects* and *unstacking objects, microwaving food* and *taking food*. Also, the activity of *picking objects* was often confused with *arranging objects*, due to the fact that body pose segments of the first are similar to the ones

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.).

Algorithm	U	O. Rec.	O. Tr.	Acc.	Prec.	Rec.
Koppula et al. [64], (*CRF*, *SVM*)	–	–	✓	83.1	87.0	82.7
Koppula et al. [24], (*CRF*, *SVM*)	–	–	✓	80.6	81.8	80.0
Our approach, (GWR)	✓	✓	–	**79.0**	**80.5**	**78.5**
Rybok et al. [63], (*SVM*)	–	✓	–	78.2	–	–
Tayyub et al. [65], (*SVM*)	–	–	✓	75.8	77.9	75.4

preceding the activity of *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 use ground-truth temporal segmentation of the activities into smaller atomic actions or sub-activities [61,62]. Our results are comparable with Rybok et al. [63]. 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. The best results were obtained by Koppula and Saxena [64], reporting 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 the relationship 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 our work, they do not perform object classification but rely on manually annotated 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 actions comprising human-object interaction from RGB-D videos. Our architecture consists of two pathways of GWR networks processing respectively body pose and object appearance and identity, with a subsequent integration layer learning action-object mappings. The prototype-based learning mechanism of the GWR allows to attenuate input noise and to generalize towards unseen data samples. For the purpose of classification, we extended the GWR 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 layer showed an overall lower network activation when given incongruent action-object pairs compared to the congruent pairs. Furthermore, the classification accuracy of our unsupervised architecture on a publicly available action benchmark dataset is competitive with respect to the supervised state-of-the-art approaches.

5.2. Self-organizing neural learning and analogies with neuroscience

Generative approaches based on self-organization learn input probability distribution through a finite set of reference vectors associated with neurons. Moreover, they resemble the topological relationships of the input space through the neurons' organization. Growing self-organizing approaches such as the GNG [50] and the GWR networks [17] are characterized by a dynamic topological structure able to adapt toward the input data space through the mechanism of the competitive Hebbian learning [66]. 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 is not well represented by the prototype neurons.

We extended 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 [19]. 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]. 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 is motivated by psychological studies on the nature of human categorization [16]. 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 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 [24,38] have shown that tracking the object's position and spatial relationship with respect to the body can help for a better interpretation of this type of interaction. There is evidence from neuroscience that the observation of using a tool activates areas of the lateral temporal cortex in the human brain 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 possible next step is to extend our model by including motion information.

An additional future work direction is the introduction of recurrent connections in the GWR networks for the purpose of temporal sequence processing. Recurrence in self-organizing networks has been extensively investigated and applied to temporal data classification [52,67]. 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 [51].

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 a highly noisy skeleton due to body self-occlusions when manipulating an object. Therefore, future work will address the limitations of this hand-crafted 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 [67].

Finally, the results reported in this paper motivate future work towards the integration of our learning system into a robotic platform and its evaluation in real-world scenarios such as learning by imitation tasks or human-robot assistance in natural environments.

Acknowledgments

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), 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, J. Neurosci. 33 (15) (2013) 6563–6580, doi:10.1523/JNEUROSCI.4129-12.2013.
[2] K. Nelissen, W. Vanduffel, G.A. Orban, Charting the lower superior temporal region, a new motion-sensitive region in monkey superior temporal sulcus, J. Neurosci. 26 (22) (2006) 5929–5947, doi:10.1523/JNEUROSCI.0824-06.2006.
[3] R. Prevete, G. Tessitore, M. Santoro, E. Catanzariti, A connectionist architecture for view-independent grip-aperture computation, Brain Res. 1225 (2008) 133–145, doi:10.1016/j.brainres.2008.04.076.
[4] G. Tessitore, R. Prevete, E. Catanzariti, G. Tamburrini, From motor to sensory processing in mirror neuron computational modelling, Biol. Cybern. 103 (6) (2010) 471–485, doi:10.1007/s00422-010-0415-5.
[5] M.S. Beauchamp, K.E. Lee, J.V. Haxby, A. Martin, Parallel visual motion processing streams for manipulable objects and human movements, Neuron 34 (1) (2002) 149–159, doi:10.1016/S0896-6273(02)00642-6.
[6] P.E. Downing, M.V. Peelen, The role of occipitotemporal body-selective regions in person perception, Cogn. Neurosci. 2 (3–4) (2011) 186–203, doi:10.1080/17588928.2011.582945.
[7] K. Grill-Spector, Representation of objects, The Oxford Handbook of Cognitive Neuroscience, Volume 2: The Cutting Edges 2 (2013).
[8] D.H. Hubel, T. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, J. Physiol. 160 (1) (1962) 106–154.
[9] E.D. Grossman, R. Blake, Brain areas active during visual perception of biological motion, Neuron 35 (6) (2002) 1167–1175, doi:10.1016/S0896-6273(02)00897-8.
[10] M.A. Giese, T. Poggio, Neural mechanisms for the recognition of biological movements, Nat. Rev. Neurosci. 4 (3) (2003) 179–192, doi:10.1038/nrn1057.
[11] R. Saxe, S. Carey, N. Kanwisher, Understanding other minds: linking developmental psychology and functional neuroimaging, Annu. Rev. Psychol. 55 (2004) 87–124, doi:10.1146/annurev.psych.55.090902.142044.
[12] V. Gallese, L. Fadiga, L. Fogassi, G. Rizzolatti, Action recognition in the premotor cortex, Brain 119 (2) (1996) 593–609, doi:10.1093/brain/119.2.593.
[13] K. Nelissen, G. Luppino, W. Vanduffel, G. Rizzolatti, G.A. Orban, Observing others: multiple action representation in the frontal lobe, Science 310 (5746) (2005) 332–336, doi:10.1126/science.1115593.
[14] E.Y. Yoon, G.W. Humphreys, S. Kumar, P. Rotshtein, The neural selection and integration of actions and objects: an fMRI study, J. Cogn. Neurosci. 24 (11) (2012) 2268–2279, doi:10.1162/jocn_a_00256.
[15] R. Miikkulainen, J.A. Bednar, Y. Choe, J. Sirosh, Computational Maps in the Visual Cortex, Springer Science & Business Media, 2006.
[16] E. Rosch, C.B. Mervis, Family resemblances: Studies in the internal structure of categories, Cogn. Psychol. 7 (4) (1975) 573–605, doi:10.1016/0010-0285(75)90024-9.
[17] S. Marsland, J. Shapiro, U. Nehmzow, A self-organising network that grows when required, Neural Netw. 15 (8) (2002) 1041–1058, doi:10.1016/S0893-6080(02)00078-3.
[18] G.I. Parisi, C. Weber, S. Wermter, Self-organizing neural integration of pose-motion features for human action recognition, Front. Neurorobotics 9 (2015), doi:10.3389/fnbot.2015.00003.
[19] G.I. Parisi, C. Weber, S. Wermter, Human action recognition with hierarchical growing neural gas learning, in: Proceedings of the International Conference on Artificial Neural Networks (ICANN), Springer, 2014, pp. 89–96, doi:10.1007/978-3-319-11179-7_12.
[20] G.S. Donatti, O. Lomp, R.P. Würtz, Evolutionary optimization of growing neural gas parameters for object categorization and recognition, in: Proceedings of the International Joint Conference on Neural Networks (IJCNN), IEEE, 2010, pp. 1–8, doi:10.1109/IJCNN.2010.5596682.
[21] L. Mici, G.I. Parisi, S. Wermter, Recognition of transitive actions with hierarchical neural network learning, in: Proceedings of the Artificial Neural Networks and Machine Learning (ICANN), Springer International Publishing, 2016, pp. 472–479, doi:10.1007/978-3-319-44781-0_56.
[22] T. Kohonen, Essentials of the self-organizing map, Neural Netw. 37 (2013) 52–65, doi:10.1016/j.neunet.2012.09.018.
[23] B. Fritzke, Kohonen feature maps and growing cell structures-a performance comparison, in: Proceedings of the Advances in Neural Information Processing Systems 5 (NIPS), Morgan Kaufmann, 1993, pp. 123–130.
[24] H.S. Koppula, R. Gupta, A. Saxena, Learning human activities and object affordances from RGB-D videos, Int. J. Robot. Res. 32 (8) (2013) 951–970, doi:10.1177/0278364913478446.
[25] J.K. Aggarwal, M.S. Ryoo, Human activity analysis: a review, ACM Comput. Surv. (CSUR) 43 (3) (2011), doi:10.1145/1922649.1922653.
[26] M. Ziaeefard, R. Bergevin, Semantic human activity recognition: a literature review, Pattern Recognit. 48 (8) (2015) 2329–2345, doi:10.1016/j.patcog.2015.03.006.
[27] J.K. Aggarwal, L. Xia, Human activity recognition from 3D data: a review, Pattern Recognit. Lett. 48 (2014) 70–80, doi:10.1016/j.patrec.2014.04.011.
[28] E. Cippitelli, S. Gasparrini, E. Gambi, S. Spinsante, A human activity recognition system using skeleton data from RGBD sensors, Comput. Intell. Neurosci. 2016 (2016), doi:10.1155/2016/4351435.
[29] X. Yang, Y. Tian, Effective 3D action recognition using eigenjoints, J. Vis. Commun. Image Represent. 25 (1) (2014) 2–11, doi:10.1016/j.jvcir.2013.03.001.
[30] J. Sung, C. Ponce, B. Selman, A. Saxena, Unstructured human activity detection from RGBD images, in: Proceedings of the International Conference on Robotics and Automation, (ICRA), IEEE, 2012, pp. 842–849, doi:10.1109/ICRA.2012.6224591.
[31] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, J.M. Rehg, A scalable approach to activity recognition based on object use, in: Proceedings of the International Conference on Computer Vision, (ICCV), IEEE, 2007, pp. 1–8.
[32] Y. Yang, Y. Li, C. Fermüller, Y. Aloimonos, Robot learning manipulation action plans by "watching" unconstrained videos from the world wide web, in: Proceedings of the Association for the Advancement of Artificial Intelligence, (AAAI), 2015, pp. 3686–3693.
[33] A. Pieropan, G. Salvi, K. Pauwels, H. Kjellström, Audio-visual classification and detection of human manipulation actions, in: Proceedings of the IEEE International Conference On Intelligent Robots and Systems, (IROS), IEEE, 2014, pp. 3045–3052.
[34] A. Gupta, L.S. Davis, Objects in action: an approach for combining action understanding and object perception, in: Proceedings of the Computer Vision and Pattern Recognition, (CVPR), IEEE, 2007, pp. 1–8, doi:10.1109/CVPR.2007.383331.
[35] A. Gupta, A. Kembhavi, L.S. Davis, Observing human-object interactions: using spatial and functional compatibility for recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (10) (2009) 1775–1789, doi:10.1109/TPAMI.2009.83.
[36] M.S. Ryoo, J. Aggarwal, Hierarchical recognition of human activities interacting with objects, in: Proceedings of the Computer Vision and Pattern Recognition, (CVPR), IEEE, 2007, pp. 1–8, doi:10.1109/CVPR.2007.383487.
[37] B. Yao, L. Fei-Fei, Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses, IEEE Trans. Pattern Anal. Mach. Intell. 34 (9) (2012) 1691–1703, doi:10.1109/TPAMI.2012.67.
[38] H. Kjellström, J. Romero, D. Kragić, Visual object-action recognition: inferring object affordances from human demonstration, Comput. Vis. Image Understanding 115 (1) (2011) 81–90, doi:10.1016/j.cviu.2010.08.002.
[39] B. Yao, L. Fei-Fei, Grouplet: a structured image representation for recognizing human and object interactions, in: Proceedings of the Computer Vision and Pattern Recognition, (CVPR), IEEE, 2010, pp. 9–16, doi:10.1109/CVPR.2010.5540234.
[40] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110, doi:10.1023/B:VISI.0000029664.99615.94.
[41] E.E. Aksoy, A. Abramov, J. Dörr, K. Ning, B. Dellen, F. Wörgötter, Learning the semantics of object–action relations by observation, Int. J. Robot. Res. (2011) 1229–1249, doi:10.1177/0278364911410459.
[42] M. Shimozaki, Y. Kuniyoshi, Integration of spatial and temporal contexts for action recognition by self organizing neural networks, in: Proceedings of the IEEE International Conference On Intelligent Robots and Systems, (IROS), 3, IEEE, 2003, pp. 2385–2391.
[43] C. Lea, A. Reiter, R. Vidal, G.D. Hager, Segmental spatiotemporal CNNS for fine–grained action segmentation, in: Proceedings of the European Conference on Computer Vision, (ECCV), Springer, 2016, pp. 36–52.
[44] CY. Ma, A. Kadav, I. Melvin, Z. Kira, G. AlRegib, H.P. Graf, Attend and Interact: Higher-Order Object Interactions for Video Understanding, arXiv preprint arXiv:1711.06330 (2017).
[45] J. Han, L. Shao, D. Xu, J. Shotton, Enhanced computer vision with microsoft kinect sensor: a review, IEEE Trans. Cybern. 43 (5) (2013) 1318–1334, doi:10.1109/TCYB.2013.2265378.
[46] J. Wang, Z. Liu, Y. Wu, Learning actionlet ensemble for 3D human action recognition, in: Proceedings of the Human Action Recognition with Depth Cameras, Springer International Publishing, 2014, pp. 11–40, doi:10.1109/TPAMI.2013.198.
[47] T. Poggio, S. Edelman, A network that learns to recognize three-dimensional objects, Nature 343 (6255) (1990) 263–266, doi:10.1038/343263a0.
[48] D. Perrett, View-dependent coding in the ventral stream and its consequences for recognition, Vision Movement Mechanisms in the Cerebral Cortex (1996) 142–151.
[49] T. Martinetz, K. Schulten, A "neural-gas" network learns topologies, in: Proceedings of the Artificial Neural Networks, Elsevier Science Publisher B.V., 1991, pp. 397–402.

[50] B. Fritzke, A growing neural gas network learns topologies, Adv. Neural Inf. Process. Syst. 7 (1995) 625–632.
[51] C.C. Aggarwal, A. Hinneburg, D.A. Keim, On the surprising behavior of distance metrics in high dimensional space, in: Proceedings of the International Conference on Database Theory, (ICDT), Springer, 2001, pp. 420–434, doi:10.1007/3-540-44503-X_27.
[52] M. Strickert, B. Hammer, Merge SOM for temporal data, Neurocomputing 64 (2005) 39–71, doi:10.1016/j.neucom.2004.11.014.
[53] G. Evangelidis, G. Singh, R. Horaud, Skeletal quads: human action recognition using joint quadruples, in: Proceedings of the International Conference on Pattern Recognition (ICPR), 2014, pp. 4513–4518, doi:10.1109/ICPR.2014.772.
[54] T. Vatanen, M. Osmala, T. Raiko, K. Lagus, M. Sysi-Aho, M. Orešič, T. Honkela, H. Lähdesmäki, Self-organization and missing values in SOM and GTM, Neurocomputing 147 (2015) 60–70.
[55] T. Tuytelaars, C.H. Lampert, M.B. Blaschko, W. Buntine, Unsupervised object discovery: a comparison, Int. J. Comput. Vis. 88 (2) (2010) 284–302, doi:10.1007/s11263-009-0271-8.
[56] T. Kinnunen, J.-K. Kamarainen, L. Lensu, H. Kälviäinen, Unsupervised object discovery via self-organisation, Pattern Recognit. Lett. 33 (16) (2012) 2102–2112, doi:10.1016/j.patrec.2012.07.013.
[57] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, C. Schmid, Aggregating local image descriptors into compact codes, IEEE Trans. Pattern Anal. Mach. Intell. 34 (9) (2012) 1704–1716, doi:10.1109/TPAMI.2011.235.
[58] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338, doi:10.1007/s11263-009-0275-4.
[59] R. Szeliski, Computer vision: algorithms and applications, Springer Science & Business Media, 2010.
[60] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. Manag. 45 (4) (2009) 427–437, doi:10.1016/j.ipm.2009.03.002.
[61] N. Hu, G. Englebienne, Z. Lou, B. Kröse, Learning latent structure for activity recognition, in: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)„ IEEE, 2014, pp. 1048–1053.
[62] A. Taha, H.H. Zayed, M. Khalifa, E.-S.M. El-Horbaty, Skeleton-based human activity recognition for video surveillance, Int. J. Sci. Eng. Res. 6 (1) (2015) 993–1004.
[63] L. Rybok, B. Schauerte, Z. Al-Halah, R. Stiefelhagen, "Important stuff, everywhere!" Activity recognition with salient proto-objects as context, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, (WACV), IEEE, 2014, pp. 646–651, doi:10.1109/WACV.2014.6836041.
[64] H.S. Koppula, A. Saxena, Learning spatio-temporal structure from RGB-D videos for human activity detection and anticipation., in: Proceedings of the IEEE International Conference on Machine Learning (ICML), 2013, pp. 792–800.
[65] A. Tayyub JawadTavanai, Y. Gatsoulis, A.G. Cohn, D.C. Hogg, Qualitative and quantitative spatio-temporal relations in daily living activity recognition, in: Proceedings of the IEEE Computer Vision–ACCV 2014, Springer International Publishing, Cham, 2015, pp. 115–130.
[66] T. Martinetz, Competitive Hebbian learning rule forms perfectly topology preserving maps, in: Proceedings of the IEEE International Conference on Artificial Nerual networks, (ICANN), Springer, 1993, pp. 427–434, doi:10.1007/978-1-4471-2063-6_104.
[67] G.I. Parisi, J. Tani, C. Weber, S. Wermter, Lifelong learning of human actions with deep neural network self-organization, Neural Netw. 96 (2017) 137–149.

Luiza Mici received her Bachelor's and Master's degree in Computer Engineering from the University of Siena, Italy. Since 2015, she is a research associate and Ph.D. candidate in the Knowledge Technology Group at the University of Hamburg, Germany, where she was part of the research project CML (Crossmodal Learning). Her main research interests include perception and learning, neural network self-organization, and bio-inspired action recognition.

German I. Parisi received his Bachelor's and Master's degree in Computer Science from the University of Milano-Bicocca, Italy. In 2017 he received his Ph.D. in Computer Science from the University of Hamburg, Germany, where he was part of the research project CASY (Cognitive Assistive Systems) and the international Ph.D. research training group CINACS (Cross-Modal Interaction in Natural and Artificial Cognitive Systems). In 2015 he was a visiting researcher at the Cognitive Neuro-Robotics Lab at the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. Since 2016 he is a research associate of international project Transregio TRR 169 on Crossmodal Learning in the Knowledge Technology Institute at the University of Hamburg, Germany. His main research interests include neurocognitive systems for human-robot assistance, computational models for multisensory integration, neural network self-organization, and deep learning.

Stefan Wermter is Full Professor at the University of Hamburg and Director of the Knowledge Technology institute. He holds an M.Sc. from the University of Massachusetts in Computer Science, and a Ph.D. and Habilitation in Computer Science from the University of Hamburg. He has been a research scientist at the International Computer Science Institute in Berkeley, US before leading the Chair in Intelligent Systems at the University of Sunderland, UK. His main research interests are in the fields of neural networks, hybrid systems, neuroscience-inspired computing, cognitive robotics and natural communication. He has been general chair for the International Conference on Artificial Neural Networks 2014. He is an associate editor of the journals æTransactions of Neural Networks and Learning Systemsg, "Connection Science", "International Journal for Hybrid Intelligent Systems" and "Knowledge and Information Systems" and he is on the editorial board of the journals "Cognitive Systems Research", "Cognitive Computation" and "Journal of Computational Intelligence". Currently he serves as Co-coordinator of the DFG-funded SFB/Transregio International Collaborative Research Centre on "Crossmodal Learning" and is coordinator of the European Training Network SECURE on Safe Robots.