In: Proceedings of the 25th International Conference on Artificial Neural Networks (ICANN 2016), Part II, LNCS 9887, pp. 472–479, Barcelona, Spain (2016)

Recognition of Transitive Actions with Hierarchical Neural Network Learning

Luiza Mici, German I. Parisi, and Stefan Wermter

University of Hamburg - Department of Informatics
Vogt-Koelln-Strasse 30, D-22527 Hamburg - Germany
{mici,parisi,wermter}@informatik.uni-hamburg.de
http://www.informatik.uni-hamburg.de/WTM/

Abstract. The recognition of actions that involve the use of objects has remained a challenging task. In this paper, we present a hierarchical self-organizing neural architecture for learning to recognize transitive actions from RGB-D videos. We process separately body poses extracted from depth map sequences and object features from RGB images. These cues are subsequently integrated to learn action–object mappings in a self-organized manner in order to overcome the visual ambiguities introduced by the processing of body postures alone. Experimental results on a dataset of daily actions show that the integration of action–object pairs significantly increases classification performance.

Keywords: action recognition, self-organization, hierarchical learning.

1 Introduction

The ability to understand others' actions represents a crucial feature of the human visual system that fosters learning and social interactions in natural environments. In particular, the recognition of transitive actions (actions that involve the interaction with a target object) is an important part of human daily activities. Therefore, computational approaches for the recognition of transitive actions are a desirable feature of assistive systems able to interact with people in real-world scenarios. While humans possess an outstanding capability to easily extract and reason about abstract concepts such as the goal of actions and the interaction with objects, this capability has remained an open challenge for computational models of action recognition.

The study of transitive actions such as grasping and holding has often been the focus of research in neuroscience and psychology [1–3], especially after the discovery of the mirror neuron system [3]. It has been shown that a specific set of neurons in the mammalian brain shows selective tuning during the observation of actions for which an internal motor representation is present in the nervous system. Moreover, the response of these neurons differs in case the action is mimicked, i.e. the target object is absent. Neurophysiological studies suggest that only when information about the object identity is added to the semantic

In: Proceedings of the 25th International Conference on Artificial Neural Networks (ICANN 2016), Part II, LNCS 9887, pp. 472–479, Barcelona, Spain (2016)

information about the action, then the actions of other individuals can be completely understood [4]. Together, these results provide an interesting framework that has motivated research work in the field of artificial vision systems and machine learning towards the recognition of action–object mappings (e.g., [5–8]). From the computational perspective, an important question can be posed on the potential links between representations of body postures and manipulated objects involved in the learning of transitive actions and, in particular, on the way these two representations can be integrated.

In this paper, we present a hierarchical, self-organizing neural architecture that learns to recognize transitive actions from RGB-D videos containing daily activities. Unlike our previous work [9], we use self-organizing neural networks motivated by the fact that specific areas of the visual system organize according to the distribution of the inputs [12]. Furthermore, extended models of hierarchical self-organization enable the learning of inherent spatio-temporal dependencies of time-varying input such as body motion sequences [10]. The proposed architecture consists of two main network streams processing separately feature representations of body postures and manipulated objects. The last layer, where the two streams are integrated, combines the information for developing action–object mappings in a self-organized manner. We evaluate our architecture with a dataset of RGB-D videos containing daily actions. We present and discuss our results on this dataset showing that the identity of objects plays a fundamental role for the effective recognition of actions.

2 Neural architecture

The proposed architecture is based on self-organizing neural networks that are capable of learning inherent topological relations of the input space in an unsupervised fashion. An overview of the architecture is depicted in Fig. 1.

2.1 Self-Organizing Maps

Self-organizing maps are neural networks inspired by biological input-driven self-organization [11] and they have been successfully applied to a number of learning tasks [12]. It consists of a 2-dimensional grid of units (neurons), each associated with a weight vector of the same dimension of the input space. The learning is performed by adapting these weights to better encode a submanifold of the input space. Given an input vector $\mathbf{x}_i$, this is done by calculating a best-matching unit $b \in A$, where A is the set of map nodes:

$$b = arg \min_{n \in A} ||\mathbf{x} - \mathbf{w}_n||. \tag{1}$$

Then, the weight vector $\mathbf{w}_b$ is moved closer to the input by a fraction that decreases over time, as are nodes that are in the neighborhood of the winner:

$$\mathbf{w}_b(t+1) = \mathbf{w}_b(t) + \eta(t) \cdot h_b(t) \cdot [\mathbf{x}(t) - \mathbf{w}_b(t)], \tag{2}$$

In: Proceedings of the 25th International Conference on Artificial Neural Networks (ICANN 2016), Part II, LNCS 9887, pp. 472–479, Barcelona, Spain (2016)

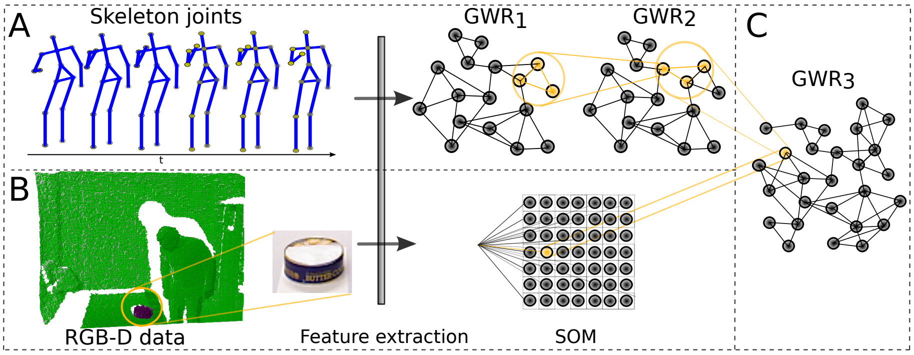

Fig. 1. Overview of the proposed architecture. (A) Processing for the body postures: a set of local features that encode the posture of upper body limbs are extracted and fed to the 2-layered neural architecture with GWR networks. (B) The input for the object recognition module is the RGB image of the object: the region of interest is automatically extracted through a point cloud-based table top segmentation. Objects are represented as compact feature vectors and are fed to a SOM network. (C) The last layer learns the combinations of body postures and objects involved in an action.

where $h_b(t)$ is the neighborhood function that defines the spatial neighbors of the winning neuron and $\eta(t)$ is a decreasing learning rate. In this way, the neurons in the map are organized preserving the topological properties of the input, i.e. similar inputs are mapped to neurons that are near to each other in the map.

The presence of noise in terms of outliers in the input data can have a negative influence on the formation of topological representations using SOMs. Such an issue is better addressed by growing models of self-organizing networks.

2.2 Growing When Required Networks

The Growing When Required network (GWR) [16] is a growing extension of self-organizing networks with competitive Hebbian learning. The GWR has the ability to create neurons and connections between them to incrementally map the topology of the input data distribution. Unlike the well-known Growing Neural Gas (GNG) [17], where the network grows at a constant rate, the GWR has a growth rate as a function of the overall network activation w.r.t. the input.

The GWR network starts with a set A of two nodes with random weights $\mathbf{w}_1$ and $\mathbf{w}_2$ in the input space. At each iteration, the algorithm is given an input $\mathbf{x}(t)$ and the the two closest neurons b and s in A are found (Eq. 1). If the connection (b,s) does not exist, it is created. The activity of the best-matching neuron is computed as $a = \exp(-||\mathbf{x} - \mathbf{w}_b||)$. If the activity is lower than a pre-defined threshold a_T and the firing counter of the neuron is under the firing threshold h_T, then a new neuron is created with weight $\mathbf{w}_r = (\mathbf{w}_b + \mathbf{x}(t))/2$. The firing rate threshold parameter makes sure that neurons are sufficiently trained before inserting new ones. The edge between b and s is removed and the edges (r, b)

In: Proceedings of the 25th International Conference on Artificial Neural Networks (ICANN 2016), Part II, LNCS 9887, pp. 472–479, Barcelona, Spain (2016)

and (r, s) are created. If a new neuron is not added, the weights of the winning neuron and its neighbours are moved towards the input by a fraction of $\epsilon \cdot h$, with $0 < \epsilon < 1$ and h being the firing counter of the neuron. The firing counters are reduced and the age of the edges are increased. The algorithm stops when a given criterion is met, e.g., a maximum network size. The insertion threshold a_T modulates the amount of generalization i.e. how much discrepancy we want to tolerate between the resulting prototype neurons and the input space. The connection-age mechanism leads to neurons being removed if rarely used.

2.3 Learning Sequences of Body Postures

Our study focuses on articulated motion of the upper body limbs during daily activities such as picking up, drinking, eating, and talking on phone. The set of raw full-body joints positions in real-world coordinates does not supply a significant representation of such actions. Therefore, we compute the relative position of upper limbs w.r.t. the head and body center to obtain translation-invariant coordinates. We use the skeletal quads features that are local features built upon the concept of geometric hashing and have shown promising results for the recognition of actions and hand gestures [13]. Given a quadruple of body joints positions in real-world coordinates $X = [x_1, x_2, x_3, x_4]$ with $x \in R^3$, a local coordinate system is built by making x_1 the origin and mapping x_2 onto the vector $[1, 1, 1]^T$. The position of the other two points x_3 and x_4 calculated w.r.t. the local coordinate system are concatenated in a 6-dimensional vector which is the quadruple compact descriptor. In this way, we obtain a lower-dimensional descriptor which is also invariant to translation, scale and body rotation. We select two quadruple of joints: [*center torso, neck, left hand, left elbow*] and [*center torso, neck, right hand, right elbow*], meaning that the positions of the hands and elbows are encoded with respect to the torso center and neck. The latter is chosen instead of the head position due to noisy tracking of the head caused by occlusions during actions such as eating and drinking.

For the recognition of body motion sequences, we train a hierarchical GWR architecture (Fig. 1.A). This approach has been shown to be more suitable than SOM for learning a set of actions from features based on noisy tracked skeletons [10]. We first train the GWR_1 network with the sequences of body postures. After the training is completed, the GWR_2 network is trained with neural activation trajectories from GWR_1. Thus, for each input sample $\mathbf{x}_i$, the best-matching neuron in GWR_1 network is computed as in Eq. 1. The weights of the neurons activated within a temporal sliding window of length q are concatenated and fed as input to GWR_2. The input data for training GWR_2 is of the form:

$$\psi(\mathbf{x}_i) = \{b(\mathbf{x}_i), b(\mathbf{x}_{i-1}), ..., b(\mathbf{x}_{i-q+1}), i \in [q..m]\}, \tag{3}$$

where m is the number of training samples. While the first network learns a set of prototype body postures, the second network will learn temporally-ordered prototype sequences from q consecutive samples. Therefore, the positive recognition of action segments occurs only when neurons along the hierarchy are activated in the correct order.

In: Proceedings of the 25th International Conference on Artificial Neural Networks (ICANN 2016), Part II, LNCS 9887, pp. 472–479, Barcelona, Spain (2016)

2.4 Object Recognition

For the representation of objects, we use SIFT features [14] that yield invariance to translation, rotation and scaling transformations and, to some extent, robustness to occlusions. For the problem of object category recognition, experimental results have shown that better classification performance is achieved by computing *dense* SIFT descriptors on regular grids across each image. Since objects will be compared to each other through vectorial metrics such as the Euclidean distance, we compute a fixed-dimensional vectorial representation of each image by performing quantization followed by an encoding step. For this purpose, we chose the vector of locally aggregated descriptors (VLAD) [15]. Unlike the bag of features (BoF) approach, these descriptors do not apply hard-assignment of SIFT features from an image to the closest code-vectors, i.e. visual words. Instead, they compute and trace the differences between them, leading to a resulting feature vector with a higher discriminative power.

For learning objects, we train a SOM network on a set of objects extracted from RGB action sequences (Fig. 1.B). We attach symbolic labels to each neuron based on the majority of input samples that have matched with each neuron during the training phase. At recognition time, for each input image the best-matching neuron from the trained network (Eq. 1) will be computed. In this way, the knowledge of the category of objects can be transferred to the higher layer of the architecture in the form of a symbolic label.

2.5 Classification of Transitive Actions

Up to this point, the architecture has learned temporally-ordered prototype body posture sequences and the identity of objects. The highest network in hierarchy GWR_3 should integrate the information from the converging streams and learn action–object mappings (Fig. 1.C). For this purpose, we compute a new dataset by merging the activations trajectories from the preceding GWR_2 network and the object's symbolic label from the SOM. The resulting training data consists of pairs ϕ_u of the following form:

$$\phi_u = \{b(\psi(\mathbf{x}_i)), ..., b(\psi(\mathbf{x}_{i-q-1})), l_b(\mathbf{y}), \mathbf{x}_i \in A, \mathbf{y} \in O, u \in [q..m-q]\}, \qquad (4)$$

where $l_b(\mathbf{y})$ represents the label attached to the best-matching neuron of the object recognition module for the object input $\mathbf{y}$. Furthermore, each neuron is assigned with an action label adopting the same labelling strategy as in SOM, meaning that neurons take the label of the best-matching input samples. After the training of GWR_3 is completed, each neuron will encode a prototype segment of the action in terms of action–object pairs.

3 Experimental Results

3.1 Data Collection

The setup of the experiments and the data collection were planned having in mind the role of the objects' identity in distinguishing the actions, in particular

In: Proceedings of the 25th International Conference on Artificial Neural Networks (ICANN 2016), Part II, LNCS 9887, pp. 472–479, Barcelona, Spain (2016)

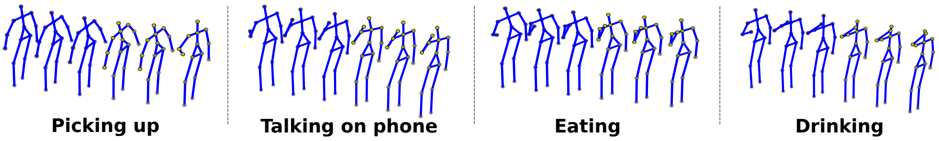

Fig. 2. Examples of sequences of skeleton joints taken from our action dataset.

when the sole body motion information may not be sufficient to unequivocally classify an action. Therefore, we collected a dataset of the following daily activities: *picking up*, *drinking* (from a mug or can), *eating* (cookies) and *talking on a phone*. The variety of style with which the actions were performed across different subjects and their similarities in body posture highlight the importance of the object's identity for their effective classification. The actions were performed by 6 participants that were given no explicit indication on the purpose of the experiments nor an explanation on how to perform the actions in order to avoid biased execution.

The dataset was collected with an Asus Xtion depth sensor that provides a synchronized RGB-D image (color and depth map). The tracking of skeleton joints was computed with the OpenNI framework (Fig. 2). Action labels were manually annotated from ground truth of sequence frames and were cross checked by two different individuals. We added a mirrored version of all action samples to obtain invariance to actions performed with either the right or the left hand. The depth sensor was also used for acquiring the objects dataset. Since object recognition should be reliable regardless of objects' perspective, RGB images were acquired with the camera positioned in two different heights and from objects in different views with respect to the sensor. Object labels were manually annotated for the training sequences, and the labels output from the object recognition module were used for the test sequences.

3.2 Training and Evaluation

In order to evaluate the generalization capabilities of our architecture, we conducted experiments with 10-fold cross-validation, meaning that data was split into 10 random subdivisions of 60% for training and 40% for testing. The results reported in this paper have been averaged over the 10 folds.

We determined empirically the following GWR training parameters: learning step sizes $\epsilon_b = 0.1$, $\epsilon_n = 0.01$, firing threshold $h_T = 0.1$, insertion thresholds $a_T = \{0.5, 0.4, 0.3\}$ (for each network respectively), maximum age $a_{max} = 100$, initial strength $h_0 = 1$, $\tau_b = 0.3$ and $\tau_n = 0.1$ as constants controlling the behaviour of the curve reducing the winning nodes' firing counter. Each GWR network was trained for 50 epochs over the whole actions dataset. The number of neurons reached in each GWR network given a training set with ≈ 18.600 frames were ≈ 480 for GWR_1, ≈ 600 for GWR_2, while for GWR_3 the number varied from ≈ 700 to ≈ 1000 depending on the inclusion or exclusion of the objects (as explained in Fig. 3). For the SOM training we used a 20 x 20 map of

In: Proceedings of the 25th International Conference on Artificial Neural Networks (ICANN 2016), Part II, LNCS 9887, pp. 472–479, Barcelona, Spain (2016)

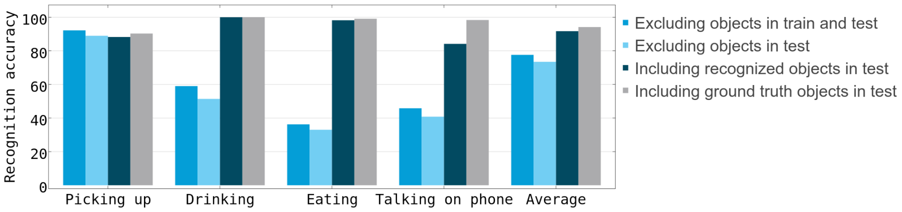

Fig. 3. Evaluation of the recognition accuracy on the test data set under the conditions indicated in the legend.

units organized in a hexagonal topology, a Gaussian neighbouring function and batch training of 50 epochs over the objects dataset.

We evaluated the recognition accuracy of the architecture under three conditions: (1) completely excluding the object identity in both training and testing, (2) including the objects in training while excluding them in testing phase, and (3) no exclusion in both phases. In the third case the label given by the SOM-based object classifier was used during testing. Further experiments were run using the objects' ground-truth labels for comparison. The results are reported in Fig. 3, where it is possible to see a significant improvement of the action classification performance for the third condition. When the objects *can* and *mug* are interchanged by the objects' classifier, the final classification accuracy of the action *drinking* is not affected – this is a desirable generalization capability of our architecture. Furthermore, the relatively low recognition rates in the second condition suggest that the identity of the object is crucial for distinguishing between the actions *drinking*, *eating* and *talking on phone*, while for the action *picking up* the situation does not vary drastically in either case.

4 Conclusions and future work

We presented a hierarchical self-organizing architecture for the learning of action–object mappings from RGB-D videos. The architecture consists of two separate pathways that process body action features and object features in parallel and subsequently it integrates prototypes of actions and the identity of objects being used. A GWR-based learning algorithm is used to learn action sequences, since it can deal better with the presence of noise in the tracked skeleton data. Experimental results have shown that the proposed integration of body actions and objects significantly increases the classification accuracy of action sequences.

The obtained results motivate the evaluation of our framework on a wider number of actions and a more complex scenario, e.g. requiring the use of the same object across different actions. Furthermore, we are working on the extension of the proposed approach for robot experiments towards the recognition of goal-oriented actions and intentions based on the interaction with the environment.

In: Proceedings of the 25th International Conference on Artificial Neural Networks (ICANN 2016), Part II, LNCS 9887, pp. 472–479, Barcelona, Spain (2016)

Acknowledgments. This research was partially supported by the Transregio TRR169 on Crossmodal Learning, by the DAAD German Academic Exchange Service for the Cognitive Assistive Systems project (Kz:A/13/94748), and the Hamburg Landesforschungsförderung.

References

1. Fleischer, F., Caggiano, V., Thier, P., Giese, M.A.: Physiologically inspired model for the visual recognition of transitive hand actions. The Journal of Neuroscience 33(15), 6563–6580 (2013)
2. Nelissen, K., Luppino, G., Vanduffel, W., Rizzolatti, G., Orban, G.: Observing others: Multiple action representation in frontal lobe. Science 310, 332–336 (2005)
3. Gallese, V., Fadiga, L., Fogassi, L., Rizzolatti, G.: Action recognition in premotor cortex. Brain 2, 593–609 (1996)
4. Saxe, R., Carey, S., Kanwisher, N.: Understanding other minds: linking developmental psychology and functional neuroimaging. Annual Review of Psychology 55, 87–124 (2004)
5. Gupta, A., Davis, L.S.: Objects in action: An approach for combining action understanding and object perception. In: IEEE CVPR'07, pp. 1–8 (2007)
6. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from rgb-d videos. The International Journal of Robotics Research 32(8), 951–970 (2013)
7. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: IEEE CVPR'10, pp 17–24 (2010)
8. Kjellström, H., Romero, J., Kragíc, D.: Visual object-action recognition: Inferring object affordances from human demonstration. Computer Vision and Image Understanding 115(1), 81–90 (2011)
9. Mici, L., Hinaut, X., Wermter, S.: Activity recognition with echo state networks using 3D body joints and objects category. In: ESANN, pp. 465–470. Belgium (2016)
10. Parisi, G.I., Weber, C., Wermter, S.: Self-organizing neural integration of pose-motion features for human action recognition. Frontiers in Neurorobotics 9(3) (2015)
11. Kohonen, T.: The self-organizing map. Proc. of the IEEE 78(9), 1464–1480 (1990)
12. Miikkulainen, R., Bednar, J.A., Choe, Y., Sirosh, J.: Computational maps in the visual cortex. Springer Science & Business Media (2006)
13. Evangelidis, G., Gurkirt, S., Radu, H.: Skeletal quads: Human action recognition using joint quadruples. In: ICPR (2014)
14. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60.2, 91–110 (2004)
15. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: IEEE CVPR'10, pp. 3304–3311 (2010)
16. Marsland, S., Shapiro, J., Nehmzow, U.: A self-organising network that grows when required. Neural Networks 15(8), 1041–1058 (2002)
17. Fritzke, B.: A growing neural gas network learns topologies. Advances in Neural Information Processing Systems 7, 625–632 (1995)