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.