Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research 39 (2016) 33–41

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

Hierarchies of Self-Organizing Maps for action recognition

Miriam Buonamente [a], Haris Dindo [a,*], Magnus Johnsson [b]

[a] RoboticsLab, DICGIM, Polytechnic School, University of Palermo, Viale delle Scienze, Ed. 6, 90128 Palermo, Italy
[b] Lund University Cognitive Science, Helgonavägen 3, 221 00 Lund, Sweden

Received 9 April 2015; received in revised form 20 September 2015; accepted 16 December 2015
Available online 21 January 2016

Abstract

We propose a hierarchical neural architecture able to recognise observed human actions. Each layer in the architecture represents increasingly complex human activity features. The first layer consists of a SOM which performs dimensionality reduction and clustering of the feature space. It represents the dynamics of the stream of posture frames in action sequences as activity trajectories over time. The second layer in the hierarchy consists of another SOM which clusters the activity trajectories of the first-layer SOM and learns to represent action prototypes. The third- and last-layer of the hierarchy consists of a neural network that learns to label action prototypes of the second-layer SOM and is independent – to certain extent – of the camera's angle and relative distance to the actor. The experiments were carried out with encouraging results with action movies taken from the INRIA 4D repository. In terms of representational accuracy, measured as the recognition rate over the training set, the architecture exhibits 100% accuracy indicating that actions with overlapping patterns of activity can be correctly discriminated. On the other hand, the architecture exhibits 53% recognition rate when presented with the same actions interpreted and performed by a different actor. Experiments on actions captured from different view points revealed a robustness of our system to camera rotation. Indeed, recognition accuracy was comparable to the single viewpoint case. To further assess the performance of the system we have also devised a behavioural experiments in which humans were asked to recognise the same set of actions, captured from different points of view. Results form such a behavioural study let us argue that our architecture is a good candidate as cognitive model of human action recognition, as architectural results are comparable to those observed in humans.

Keywords: Self-Organizing Map; Neural network; Action recognition; Hierarchical models; Intention understanding

1. Introduction

Recognition of human intentions is becoming increasingly demanded due to its potential application in a variety of domains such as assisted living and ambient intelligence, video and visual surveillance, human–computer interfaces, gaming and gesture-based control. Typically, an intention recognition system is focused on a sequence of observed actions performed by the agent whose intention is being recognised. A necessary condition for a successful intention understanding is certainly a prompt recognition of the current (i.e. proximal) action, together with environmental (contextual) cues and *a priori* knowledge (Dindo, Donnarumma, Chersi, & Pezzulo, 2015).

Many challenges make the action recognition task extremely difficult to achieve in artificial systems, as people

* Corresponding author.
E-mail addresses: miriam.buonamente@unipa.it (M. Buonamente), haris.dindo@unipa.it (H. Dindo), magnus.johnsson@lucs.lu.se (M. Johnsson).

http://dx.doi.org/10.1016/j.cogsys.2015.12.009
1389-0417/

differ in terms of height, weight, shape of the human body and gender. Another difficulties arise from the choice of input sensor. For instance, when dealing with cameras, variations in the viewing impact the accuracy of the action recognition performance (Chella, Dindo, & Infantino, 2005). Multi-camera setups have been employed to implement view independent methods (Ahmad & Lee, 2008, 2006; Weinland, Ronfard, & Boyer, 2006). These methods are based on the observation of the human body from various angles, and building a view-invariant representation.

Dealing with action recognition, it is important to give a brief definition of what do we mean by action. We adopt the following action hierarchy: *actions* and *activities*. The term action is used for simple motion patterns typically executed by a single human. An example of an action is that of crossing arms. A sequence of actions represents an activity, such as the activity of dancing. Activities usually involve coordination amongst persons, objects and environments. In this paper, we focus only on the recognition of actions, where actions can be viewed as *sequences* of body postures.

An important question is how to implement the human action recognition ability in an artificial agent. In our previous work, we have focused on the representational part of the same problem: we endowed an artificial agent with the ability to internally represent action patterns (Buonamente, Dindo, & Johnsson, 2013) using Associative Self-Organizing Map (Johnsson, Balkenius, & Hesslow, 2009), a variant of the Self-Organizing Map (SOM) (Kohonen, 1988).

In this paper, we present a novel cognitive model able to represent and classify others' behaviour. In order to get a more complete classification system we adopt a hierarchical neural approach. The first level in the system is a SOM that learns to represent postures – or posture changes – depending on the input to the system. The second level is another SOM that represents the superimposed activity trace in the first level SOM during the action, i.e. it learns to represent actions. The third level is a supervised artificial neural network that learns to label the action.

In our previous paper (Buonamente, Dindo, & Johnsson, 2013) we showed that we could get discriminable activity traces using an A-SOM, which corresponds to the first level SOM in the current system. The system was able to *simulate* the likely continuation of the recognised action. Due to this ability, the A-SOM could receive an *incomplete* input pattern (e.g. an initial part of the input sequence only) and continue to elicit the most likely evolution of the action, i.e. to carry out sequence completion of perceptual activity over time. In the present system, instead, we focus on the problem of robust action representation and recognition, given the whole (noisy) input sequence. We are currently working towards an integration of the two approaches.

We have tested the ability of our architecture to recognise observed actions on movies taken from the "INRIA 4D repository",[1] a publicly available dataset of movies representing 13 common actions: check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, point, pick up, and throw (see Fig. 1). In addition, we have validated our system in a human behavioural study.

The implementation of all code for the experiments presented in this paper was done in C++ using the neural modelling framework "Ikaros" (Balkenius, Morén, Johansson, & Johnsson, 2010).

This paper is organised as follows: A short presentation of the proposed architecture is given in Section 2; Section 3 presents the experiment for evaluating the model; and finally conclusions are outlined in Section 4.

2. Proposed architecture

The architecture presented in this paper is composed of three layers of neural networks, see Fig. 3. The first and the second layers consist of SOM networks whereas the third layer consists of a custom made supervised neural network. The first layer SOM receives sequences of vectors representing preprocessed sequences of posture images. The activity trajectories, Fig. 2, elicited during the time actions last are superimposed and vectorized into a new representation before entering the layer two SOM as input. This superimposition process can be imagined as the projection of the matrices representing the activity in the grid of neurons in the SOM for all the iterations an action lasts onto a new matrix of the same dimensionality, followed by a vectorization process. The second layer SOM thus clusters the activity trajectories and learns to represent action prototypes independent of how long the activity trajectories in the first layer SOM last. Thus the second layer SOM provides a kind of time independent representation of the action prototypes. The activity of the second layer SOM is conveyed to a third level neural network that learns to label the action prototypes of the second layer SOM independent of the camera's capturing angle and distance to the actor.

2.1. Action representation: Low layers

The first and the second layers of the architecture consist of SOMs. The SOM is one of the most popular neural networks and has been successfully applied in pattern recognition and image analysis. The SOM is trained using unsupervised learning to produce a smaller discretized representation of its input space. In a sense it resembles the functioning of the brain in pattern recognition tasks. When

[1] The repository is available at http://4drepository.inrialpes.fr. It offers several movies representing sequences of actions. Each video is captured from 5 different cameras. For the experiments in this paper we chose for training the movie "Andreas2" captured from two different frontal camera views "cam0" and "cam1", for testing "Hedlena2" with frontal camera view "cam0".