Applied Soft Computing 59 (2017) 574–585

Contents lists available at ScienceDirect

Applied Soft Computing

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

First and second order dynamics in a hierarchical SOM system for action recognition

Zahra Gharaee*, Peter Gärdenfors, Magnus Johnsson

Lund University Cognitive Science, Helgonavägen 3, 221 00 Lund, Sweden

ARTICLE INFO

Article history:
Received 23 January 2016
Received in revised form 30 April 2017
Accepted 4 June 2017
Available online 10 June 2017

Keywords:
Self-Organizing Maps
Conceptual spaces
Neural networks
Action recognition
Hierarchical models
Attention
Dynamics

ABSTRACT

Human recognition of the actions of other humans is very efficient and is based on patterns of movements. Our theoretical starting point is that the dynamics of the joint movements is important to action categorization. On the basis of this theory, we present a novel action recognition system that employs a hierarchy of Self-Organizing Maps together with a custom supervised neural network that learns to categorize actions. The system preprocesses the input from a Kinect like 3D camera to exploit the information not only about joint positions, but also their first and second order dynamics. We evaluate our system in two experiments with publicly available datasets, and compare its performance to the performance with less sophisticated preprocessing of the input. The results show that including the dynamics of the actions improves the performance. We also apply an attention mechanism that focuses on the parts of the body that are the most involved in performing the actions.

1. Introduction

The success of human–robot interaction depends on the development of robust methods that enable robots to recognize and predict goals and intentions of other agents. Humans do this, to a large extent, by interpreting and categorizing the actions they perceive. Hence, it is central to develop methods for action categorization that can be employed in robotic systems. This involves an analysis of on-going events from visual data captured by cameras to track movements of humans and to use this analysis to identify actions. One crucial question is to know what kind of information should be extracted from observations for an artificial action recognition system.

Our ambition is to develop an action categorization method that, at large, works like the human system. We present a theory of action categorization due to Gärdenfors and Warglien [9] (see also [7,8]) that builds on Gärdenfor's [6] theory of conceptual spaces. The central idea is that actions are represented by the underlying force patterns. Such patterns can be derived from the second order dynamics of the input data. We present experimental data on how humans categorize action that supports the model. A goal of this article is to show that if the dynamics of actions is considered, the performance of our Self-Organizing Maps (SOMs) [20] based action recognition system can be improved when categorizing actions based on 3D camera input.

The architecture of our action recognition system is composed of a hierarchy of three neural network layers. These layers have been implemented in different versions. The first layer consists of a SOM, which is used to represent preprocessed input frames (e.g. posture frames) from input sequences and to extract their motion patterns. This means that the SOM reduces the data dimensionality of the input and the actions in this layer are represented as activity patterns over time.

The second layer of the architecture consists of a second SOM. It receives the superimposed activities in the first layer for complete actions. The superimposition of all the activity in the first layer SOM provides a mechanism that makes the system time invariant. This is because similar movements carried out at different speed elicit similar sequences of activity in the first layer SOM. Thus the second layer SOM represents and clusters complete actions. The third layer consists of a custom made supervised neural network that labels the different clusters in the second layer SOM with the corresponding action.

We have previously studied the ability of SOMs to learn discriminable representations of actions [2], and we have developed a hierarchical SOM based action recognition architecture. This archi-

* Corresponding author.
E-mail addresses: zahra.gharaee@lucs.lu.se (Z. Gharaee), peter.gardenfors@lucs.lu.se (P. Gärdenfors), magnus@magnusjohnsson.se (M. Johnsson).

http://dx.doi.org/10.1016/j.asoc.2017.06.007
1568-4946/

tecture has previously been tested using video input from human actions in a study that also included a behavioural comparison between the architecture and humans [3], and using extracted joint positions from a Kinect like 3D camera as input with good results [12].

This article presents results that suggest that the performance of our action recognition architecture can be improved by exploiting not only the joint positions extracted from a Kinect like 3D camera, but also simultaneously the information present in their first and second order dynamics.

Apart from analysing the dynamics of the data, we implement an attention mechanism that is inspired by how human attention works. We model attention by reducing the input data to those parts of the body that contribute the most in performing the various actions. Adding such an attention mechanism improves the performance of the system.

The rest of the paper is organized as follows: First we present the theoretical background from cognitive science in Section 2. The action recognition architecture is described in detail in Section 3. Section 4 presents two experiments to evaluate the performance of the architecture employing new kinds of preprocessing to enable additional dynamic information as additional input. Section 5 concludes the paper.

2. Theoretical background

When investigating action recognition in the context of human–robot interaction, it should first be mentioned that human languages contain two types of verbs describing actions [22,40]. The first type is manner verbs that describe how an action is performed. In English, some examples are 'run', 'swipe', 'wave', 'push', and 'punch'. The second type is result verbs that describe the result of actions. In English, some examples are 'move', 'heat', 'clean', 'enter', and 'reach'.

In the context of robotics, research has focused on how result verbs can be modelled (e.g. [4,19,21,5]). However, when it comes to human–robot interaction, the robot should also be able to recognize human actions by the manner they are performed. This is often called recognition of biological motion [16]. Recognizing manner action is important in particular if the robot is supposed to model the intentions of a human. In the literature, there are some systems for categorizing human actions described by manner verbs, e.g. [15,14]. However, these systems have not been developed with the aim of supporting human–robot interaction. Our aim in this article is to present a system that recognizes a set of manner actions. Our future aim is, however, to integrate this with a system for recognizing results verbs that can be used in linguistic interactions between a human and an robot (see [4,27] for examples of such linguistic systems).

Results from the cognitive sciences indicate that the human brain performs a substantial information reduction when categorizing human manner actions. Johansson [18] has shown that the kinematics of a movement contain sufficient information to identify the underlying dynamic patterns. He attached light bulbs to the joints of actors who were dressed in black and moved in a black room. The actors were filmed performing actions such as walking, running, and dancing. Watching the films – in which only the dots of light could be seen – subjects recognized the actions within tenths of a second. Further experiments by Runesson and Frykholm [31], see also [30], show that subjects extract subtle details of the actions performed, such as the gender of the person walking or the weight of objects lifted (where the objects themselves cannot be seen).

One lesson to learn from the experiments by Johansson and his followers is that the kinematics of a movement contains sufficient information to identify the underlying dynamic force patterns. Runesson [30] claims that people can directly perceive the forces that control different kinds of motion. He formulates the following thesis:

Kinematic specification of dynamics: The kinematics of a movement contains sufficient information to identify the underlying dynamic force patterns.

From this perspective, the information that the senses – primarily vision – receive about the movements of an object or individual is sufficient for the brain to extract, with great precision, the underlying forces. Furthermore, the process is automatic: one cannot help but perceiving the forces.

Given these results from perceptual psychology, the central problem for human–robot interaction now becomes how to construct a model of action recognition that can be implemented in a robotic system. One idea for such a model comes from [26,33], who extend Marr and Nishihara's [25] cylinder models of objects to an analysis of actions. In Marr's and Vaina's model, an action is described via differential equations for movements of the body parts of, for example, a walking human. What we find useful in this model is that a cylinder figure can be described as a vector with a limited number of dimensions. Each cylinder can be described by two dimensions: length and radius. Each joining point in the figure can be described by a small number of coordinates for point of contact and angle of joining cylinder. This means that, at a particular moment, the entire figure can be written as a (hierarchical) vector of a fairly small number of dimensions. An action then consists of a sequence of such vectors. In this way, the model involves a considerable reduction of dimensionality in comparison to the original visual data. Further reduction of dimensionality is achieved in a skeleton model.

It is clear that, using Newtonian mechanics, one can derive the differential equations from the forces applied to the legs, arms, and other moving parts of the body. For example, the pattern of forces involved in the movements of a person running is different from the pattern of forces of a person walking; likewise, the pattern of forces for saluting is different from the pattern of forces for throwing [34].

The human cognitive apparatus is not exactly evolved for Newtonian mechanics. Nevertheless, Gärdenfors [7] (see also [40,8]) proposed that the brain extracts the forces that lie behind different kinds of movements and other actions:

Representation of actions: An action is represented by the pattern of forces that generates it.

We speak of a pattern of forces since, for bodily motions, several body parts are involved; and thus, several force vectors are interacting (by analogy with Marr's and Vaina's differential equations). Support for this hypothesis will be presented below. One can represent these patterns of forces in principally the same way as the patterns of shapes described in [8], Section 6.3. In analogy with shapes, force patterns also have meronomic structure. For example, a dog with short legs moves in a different way than a dog with long legs.

This representation fits well into the general format of conceptual spaces presented by Gärdenfors [6,7]. In order to identify the structure of the action domain, similarities between actions should be investigated. This can be accomplished by basically the same methods used for investigating similarities between objects. Just as there, the dynamic properties of actions can be judged with respect to similarities: for example, walking is more similar to running than to waving. Very little is known about the geometric structure of the action domain, except for a few recent studies that we will present below. We assume that the notion of betweenness is meaningful in the action domain, allowing us to formulate the following thesis in analogy to the thesis about properties (see [6,7,9]):

Thesis about action concepts: An action concept is represented as a convex region in the action domain.