Biologically Inspired Cognitive Architectures xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Biologically Inspired Cognitive Architectures

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

Online recognition of actions involving objects

Zahra Gharaee[a,*], Peter Gärdenfors[a,b], Magnus Johnsson[a,c]

[a] Lund University Cognitive Science, Helgonavägen 3, 221 00 Lund, Sweden
[b] Social Robotics Studio, Centre of Quantum Computation and Intelligent Systems, University of Technology, Sydney, Australia
[c] Department of Intelligent Cybernetic Systems, NRNU MEPhI, Moscow, Russia

ARTICLE INFO

Keywords:
Hierarchical models
Self-organizing maps
Action recognition
Object detection

ABSTRACT

We present an online system for real time recognition of actions involving objects working in online mode. The system merges two streams of information processing running in parallel. One is carried out by a hierarchical self-organizing map (SOM) system that recognizes the performed actions by analysing the spatial trajectories of the agent's movements. It consists of two layers of SOMs and a custom made supervised neural network. The activation sequences in the first layer SOM represent the sequences of significant postures of the agent during the performance of actions. These activation sequences are subsequently recoded and clustered in the second layer SOM, and then labeled by the activity in the third layer custom made supervised neural network. The second information processing stream is carried out by a second system that determines which object among several in the agent's vicinity the action is applied to. This is achieved by applying a proximity measure. The presented method combines the two information processing streams to determine what action the agent performed and on what object. The action recognition system has been tested with excellent performance.

Introduction

Action recognition plays an important role in interactions between any agents whether they are humans, animals or robots. Johansson (1973) showed by using a patch light technique that humans can identify an action after only about two hundred milliseconds. Such an efficient mechanism for interpreting and categorizing a perceived action is an important factor behind smooth interaction and cooperation between humans. His studies opened up the field of biological motion within psychology.

Later studies of human categorizations of actions have shown a number of features that are relevant also for robotic models. Firstly, categorizations of actions exhibit the same prototype effects as categorizations of objects (Hemeren, 2008). Secondly, actions can be categorized in terms of the force patterns involved (Gärdenfors & Warglien, 2012; Gärdenfors, 2014; Runesson, 1994). In other words, the dynamics of an action may be more characteristic than its kinematics. Thirdly, human judgments concerning the segmentation of actions show large agreements (Radvansky & Zacks, 2014).

Given the efficiency of the human action recognition system, it should serve as an inspiration when developing fast and robust methods for action recognition that can be employed in social robotic systems that are interacting with humans online. The general task for such a robotic system is to use online visual data from cameras to track movements of humans and to use this information to categorize actions and then generate an appropriate response, linguistic or non-linguistic. It should be noted, however, that online automatic action recognition is not only useful for human-robot interaction, but also for areas such as video surveillance, human–computer interaction, video retrieval, sign language recognition, medical health care and sport.

In this article we present a biologically inspired system for online action recognition that merges the information analyses from two subsystems running in parallel. To some extent, our architecture is inspired by the two-streams hypothesis about how the brains processes visual information (Goodale & Milner, 1992). This hypothesis distinguishes between a ventral stream (the what pathway) and a dorsal stream (the where or how pathway). Our two subsystems can be seen as corresponding to these two streams. The first subsystem determines which object the agent acts on by applying a proximity measure (our system, however, takes a shortcut when identifying objects). The second subsystem recognizes what action is performed by using a hierarchical self-organizing map system that analyses the spatial trajectories of the agent's movements.

The fact that the dorsal pathway is called both the how system and the where system reflects that two perspectives can be used when categorizing an action. The first focuses on the manner in which an action is performed (how), for example, whether an object is pushed or pulled. The second perspective focuses on the result of the action, for example,

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

http://dx.doi.org/10.1016/j.bica.2017.09.007
Received 23 August 2017; Received in revised form 20 September 2017; Accepted 20 September 2017

whether an object moves (where) or changes some property. In parallel with this distinction, natural languages contain two types of verbs describing actions (Levin & Rappaport Hovav, 2005; Warglien, Gärdenfors, & Westera, 2012). 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. Manner verbs express causes and result verbs express effects of actions. In our previous experiments (Gharaee, Gärdendors, & Johnsson, 2016, 2017b, 2017a) actions without objects have been studied and the verbs describing the output have been manner verbs. In the present study that include objects, the output contains both manner and result verbs.

In human-robot applications it is important to collaborate about objects, so it is necessary to develop a system that can categorize actions involving objects as well as pure manner actions. Within robotics, action recognition systems have, until recently, been based on the result perspective, focusing on how result verbs can be modeled, e.g. (Cangelosi et al., 2008; Demiris & Khadhouri, 2006; Kalkan, Dag, Yürüten, Borghi, & Sahin, 2014; Lallee, Madden, Hoen, & Ford Dominey, 2010). From this perspective, it is sufficient to know the pre-state and post-state of the environment before and after performing an action in order to categorize an action. In the method proposed in Lallee et al. (2010), the robot learns four actions including objects as “cover”, “uncover”, “give” and “take” through linguistic interactions with human agents and as a result generates spoken language that represents its perceptions of the performed action. A human-robot communication system that includes both manner and result verbs has been developed in Mealier, Gärdenfors, and Pointeau (2016). In this study, the action comprehension and object detection occurs through visual perception (observations) and human-robot spoken interactions (expression of causes and effects of the actions).

In the literature one finds several systems that can categorize different sets of human actions. In the past, research focused on categorizing actions based on image sequences from ordinary visible light cameras (Niebles, Wang, & Li, 2008). Unfortunately such cameras have severe limitations such as sensitivity to color and illumination variations, occlusions, and background clutters. As a consequence, Kinect and other depth cameras are often used instead since they provide 3D information about the scene, which offers more discerning information of the human postures involved in the actions that are studied. The depth camera can also operate in total darkness which is a benefit for applications such as continuous patient/animal monitoring systems. The skeletons estimated from depth images are quite accurate, but the algorithm still has limitations. It gives inaccurate results when the human body is partly occluded, and the estimation is not reliable when the person touches the background or when the person is not in an upright position (Xia & Aggarwal, 2013).

In Li, Zhang, and Liu (2010), a data set of 20 actions, each performed by 10 actors in 2 or 3 different events, has been collected from sequences of depth maps obtained by a depth camera. An action graph is used to model the dynamics of the actions, and a collection of 3D points is used to characterize a set of salient postures corresponding to the nodes in the action graph. The same data set, often called the MSR Action 3D data set, has been studied by many other researchers. Here we briefly present some of the methods that have been used.

In Xia, Chen, and Aggarwal (2012), a method applied to the histogram of 3D joint locations as a compact representation of postures is introduced. It uses Linear Discrimant Analysis to project the histogram of 3D joint locations extracted from the action depth sequences and then clusters them into k posture visual words (the prototypical action poses). Another method for activity recognition from videos gained by a depth sensor is represented in Oreifej and Liu (2013). It builds a histogram to capture the distribution of the surface normal orientation in the 4D space of time, depth, and spatial coordinates by creating 4D projectors, which quantize the 4D space and represent the possible directions for the 4D normal. The method presented in Yang and Tian (2012), also uses body joints extracted from sequences of depth maps. It applies features based on position differences of joints (eigen joints) that combine action information including static posture, motion, and offset and then uses the Naive Bayes Nearest Neighbour classifier to classify actions. To recognize human actions from depth maps, Yang, Zhang, and Tian (2012) use depth maps that are projected onto three orthogonal planes and global activities through entire video sequences that are accumulated to generate a Depth Motion Map. Then the histograms of Oriented Gradients are extracted from the Depth Motion Map as the representation of an action video.

A pose-based action recognition system is introduced in Wang, Wang, and Yuille (2013) that extends the method in Yang and Ramanan (2011) to estimate human poses from action videos. It infers the best poses by best-K pose estimation for each frame by incorporating segmentation and temporal constraints for all frames in the video. A visual representation for 3D action recognition from sequences of depth maps, called Space–Time Occupancy Patterns, is used in Vieira, Nascimento, Oliveira, Liu, and Campos (2012). In the proposed feature descriptor method, a 4D grid for each depth map sequence is produced by dividing space and time axes into multiple segments to preserve spatial and temporal information between space–time cells. In Wang, Liu, Chorowski, Chen, and Wu (2012), semi-local features called Random Occupancy Pattern features are extracted and a sparse coding approach is used to encode these features. In Wang, Liu, Wu, and Yuan (2012), an actionlet ensemble model which learns to represent each action and to capture the intra-class variance is introduced. The model proposes features of depth data that are capable of characterizing human motion and human-object interactions. The use of local spatio-temporal interest points (STIPs) and the resulting features from RGB videos is the base of the activity recognition method presented in Xia and Aggarwal (2013). In the method, first the STIPs are extracted from depth videos (called DSTIP), and then 3D local cuboid in depth videos by Depth Cuboid Similarity Feature (DCSF) are described. Using DSTIP and DCSF to recognize activities from depth videos have no dependence on the skeletal joints information. A non-parametric Moving Pose framework for low-latency human activity recognition is proposed in Zanfir, Leordeanu, and Sminchisescu (2013), which is a descriptor that considers the pose information together with the speed and acceleration of the skeleton joints. The descriptor works with a modified k-nearest neighbours classifier, which employs both the temporal location of a particular frame within the action sequence as well as the discrimination power of its moving pose descriptor compared to other frames in the training set.

Common to all the systems presented here is that they use a prerecorded data set of actions, typically the MSR data (Li et al., 2010). The movies for actions are edited so that they only cover one of the actions that will be categorized and not the intermediate intervals. The systems are then trained to classify the actions. The experiments are mostly performed on one specific way of data assortment. This means that the systems are in general not tested on movies outside the data set and it is unknown to what extent they can generalize to new movies. Moreover, these systems are tested in offline experiments while in human-robot interaction scenarios, the system are supposed to identify actions online in real time.

Among related studies that propose an online implementation of action recognition, Ellis, Masood, Tappen, Laviola, and Sukthankar (2013) proposes an online action classification method based on the canonical body poses, and a feature descriptor based method for classifying actions from depth sequences introduced in Vieira et al. (2012). In both studies, the actions used in online experiments do not involve objects, only body movements which form the agent’s spatial trajectories.

Variants of our system presented below have also been trained and tested with the prerecorded MSR data set in the research studies presented in (Gharaee, Gärdendors, & Johnsson, 2017b, 2017a). In addition to that, our system is also used in online experiments with new