J. Vis. Commun. Image R. 35 (2016) 103–111

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

# Learning hierarchical spatio-temporal pattern for human activity prediction ☆

Wenwen Ding, Kai Liu *, Fei Cheng, Jin Zhang

*School of Computer Science and Technology, Xidian University, Xi'an, China*

ARTICLE INFO

*Article history:*
Received 7 September 2015
Accepted 9 December 2015
Available online 18 December 2015

*Keywords:*
Skeleton joints
3D action feature representation
Self-organizing map
Hebbian learning
Variable order Markov model
Probabilistic suffix tree
RGB-D dataset
3D trajectory segmentation

ABSTRACT

Human activity prediction has become increasingly valuable in many applications. This paper, initially from the perspective of cognition science, presents a novel approach to learning a hierarchical spatio-temporal pattern of human activities to predict ongoing activities from videos that contain only the onsets of the activities. Spatio-temporal pattern can be learned by a Hierarchical Self-Organizing Map (HSOM), which consists of two self-organizing maps (i.e., action map and actionlet map) connected via associative links trained by Hebbian learning. Ongoing activities can be predicted by Variable order Markov Model (VMM), which provides the means for capturing both large and small order Markov dependencies based on the training actionlet sequences. Experiments of the proposed method on four challenging 3D action datasets captured by commodity depth cameras show promising results.



## 1. Introduction

Human action recognition, the automatic recognition of ongoing actions performed by humans, is an active research topic in computer vision. It has a variety of real-world applications, including video surveillance, video retrieval, and even health-care. Over the past few decades, research has primarily concentrated on processes of learning and recognizing actions from video sequences [1,2]. In contrast, less attention has been paid to early detection of unfinished activities from video streams where early prediction of ongoing activity is extremely valuable [3]. For instance, in a supermarket, it would be beneficial to equip a surveillance system that can provide real-time surveillance, detect suspicious activities, and raise the alarm for theft before it happens.

Neurobiological studies [4] have concluded that the human brain can perceive actions by observing only a few actionlets[1] and component action units obtained from temporal decomposition during action execution. In the neuropsychological perspective, Friston [7] has linked the hierarchical theory of action with the organization of the brain and also described how action representations are selected, maintained, and inhibited at multiple levels of abstraction and how layers are mediated by effective connectivity.

In this vein, this paper uses a Hierarchical Self-Organizing Map (HSOM) [8] to generate model whose structure can capture the natural hierarchy which can be layered as actionlet and action from a small granularity to a large, thus make it easier to comprehension and decomposition activities at varying levels of abstraction present in human activity. Furthermore, a worthwhile approach is proposed to describe actions as sequences of consecutive actionlets and recognize action depend on a little actionlets extracted from the beginning of this action.

This paper proposes a novel framework, shown in Fig. 1, for human activity recognition from partially observed videos that use sequences of 3D skeleton joint positions as input. To obtain meaningful action units, we first learn superior segmentation points $\mathcal{S} = \{s_1, \ldots s_i, \ldots, s_j, \ldots, s_m\}(1 < i < j < m)$ to segment 3D trajectory of an action, as shown in Fig. 2a. Then decompose action, using motion velocities, the direction of motion, and the curvatures of trajectories, into a sequence of actionlets, as shown in Fig. 2b. The detailed process can be viewed clearly in our previous work [9]. The features of actions $\xi_a$ and actionlets $\xi_{al}$ are extracted from these segmented trajectories. Two Self-Organizing Maps (SOMs)

☆ This paper has been recommended for acceptance by M.T. Sun.

* Corresponding author.
*E-mail addresses:* dww2048@163.com (W. Ding), kailiu@mail.xidian.edu.cn (K. Liu), chengfei8582@163.com (F. Cheng), jinzhang.cv@gmail.com (J. Zhang).

[1] In this paper, we use actionlet to refer to meaningful atomic actions obtained from spatio-temporal decomposition using motion velocities, the direction of motion, and the curvatures of trajectories. It should be noted that the same term *actionlet* has also been used in the recent work, which refers to action components based on a spatial segmentation [5] and component action units obtained from temporal decomposition [6].

http://dx.doi.org/10.1016/j.jvcir.2015.12.006
1047-3203/

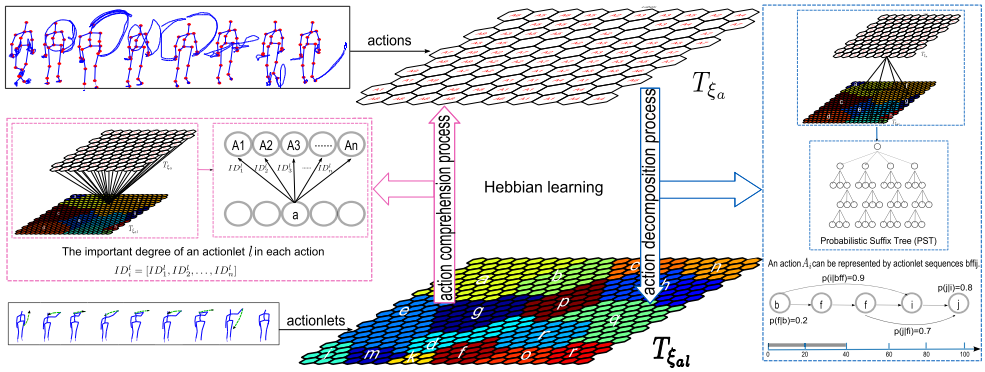


**Fig. 1.** The general framework of the proposed approach.

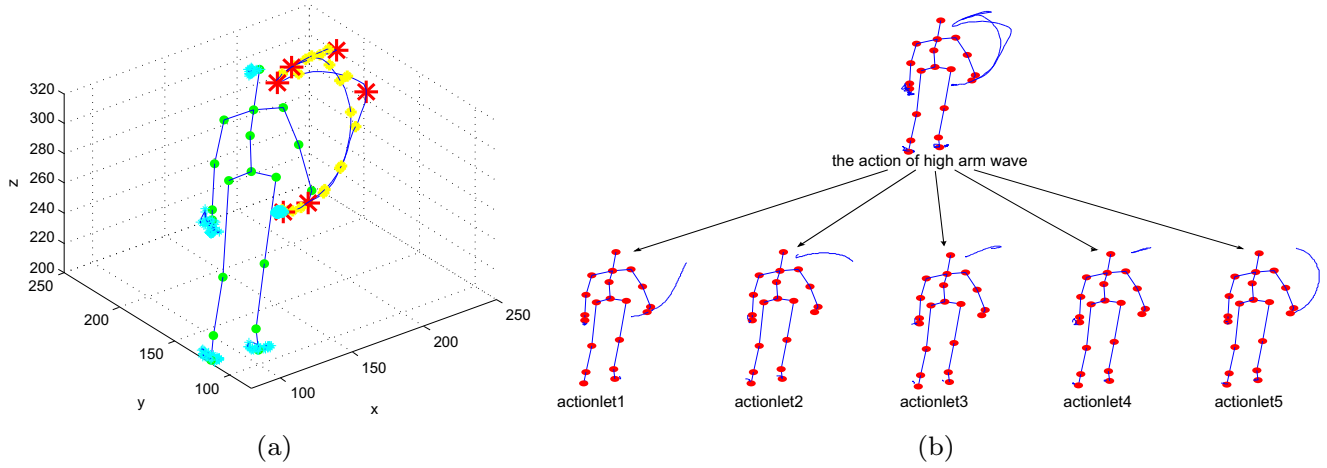


**Fig. 2.** (a) The trajectory of an action of *high hand wave* is segmented by *red stars*. (b) The action *high hand wave* can be decomposed by five actionlets through motion velocities, the direction of motion, and the curvatures of trajectories. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

constitute the Hierarchical SOM. $\xi_a$ is mapped to one SOM as $T_{\xi_a}$, and $\xi_{al}$ is mapped to the other SOM as $T_{\xi_{al}}$. Thus, numerically similar adjacent features of actions and actionlets can be mapped to a single representative vector $\boldsymbol{m}$ (model vector) on a SOM, which itself is a form of clustering process. Unlike actions that have labels (for example, high hand wave) that can be acquired from human natural language, actionlets do not have such labels. Therefore, $T_{\xi_{al}}$ can be scattered in plots named with English alphabets referred as the labels of actionlets according to the Davies–Bouldin index value [10], which is a good candidate for map unit clusters. The associative weights between $T_{\xi_a}$ and $T_{\xi_{al}}$ can be obtained through Hebbian learning [11], which can measure the important degree (ID) of an actionlet in each action. Thus, complex actions can be represented by sequences of the English letters, which are seen as context. With the help of the context and a Variable order Markov Model (VMM) [12], the probability of the next possible actionlet or the whole action can be predicted.

From the spatio-temporal perspective, an action can be characterized by the spatio-temporal information of actionlets, which can be effectively used in the learning and predicting processes. Suppose an action and actionlet in context are regarded as a word and a letter, respectively. These processes can be described as person A predicting the meaning of a sentence written by person B. For example, *eat apple* and *eat banana* represent two different meaning of English words. When person B is writing, letters are shown one by one, such as *eat na*. The word *eat banana* can be predicted by person A because the ID of letters *a*, *n* and *e* is high in word *eat banana* (especially *a*), and the causality between *na* and *eat banana* exists. However, if the next letter extracted is *p*, the sequence of letters becomes *eat nap*. Thus, the word *eat apple* may be predicted because the ID of letters *a* and *p* is not low in the word *eat apple* and *ap* has more direct causality with *eat apple* than *eat banana*.

The major contributions in this paper include: (1) HSOM is proposed to systematically exhibit the intrinsic hierarchical structure of human activity accordance with human cognition and perception from global to local as well as coarse to fine, thus making it easier to comprehension and decomposition activities at varying levels of abstraction present in human activity; and (2) Hebbian learning between actions and actionlets is modeled that allows for the representation of the important degree of actionlets in each action.

The rest of the paper is organized as follows, Section 2 presents the related work; Section 3 elaborates on the proposed method of action and actionlet representation, mapping, clustering, learning and prediction; Section 4 presents our experimental results and discussion; and Section 5 concludes this paper.

## 2. Related work

**Action recognition.** There has been tremendous amount of work on human action recognition from static images and 2D video sequences. Wang et al. [13] learns multiple features from a small number of labeled videos, and automatically utilizes data distributions between labeled and unlabeled data to boost the recognition performance. Sadanand and Corso [14] present the conception of