Signal Processing 000 (2017) 1–7

Contents lists available at ScienceDirect

Signal Processing

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

One-shot learning based pattern transition map for action early recognition

Yanli Ji[a,*], Yang Yang[b], Xing Xu[b], Heng Tao Shen[b,*]

[a] School of Automation Engineering, Center for Future Media, University of Electronic Science and Technology of China, Chengdu, China
[b] Center for Future Media, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China

ARTICLE INFO

Article history:
Received 15 March 2017
Revised 31 May 2017
Accepted 1 June 2017
Available online xxx

Keywords:
Human action recognition
Reinforcement learning
Early recognition
Pattern transition map

ABSTRACT

In a natural and accessible Human Robot Interaction (HRI), it is required to understand human activities instantly. In this paper, we present a novel approach for early recognition of human actions. Using reinforcement learning, we separate human action to several patterns and learn pattern transition maps which include temporal ordered patterns and their transition relationships in action sequences. Due to the difficulty of pattern separation and definition in large quantity of action sequences for training, we adopt one-shot learning to automatically define patterns. Moreover, we propose a pattern transition map based soft-regression approach for early recognition. We evaluate the proposed approach on the MSR Action Pairs Database and the SYSU 3DHOI database. Experiments show that our approach recognizes ongoing sequences with high accuracies. Compared with state-of-the-art approaches, our proposed approach also obtain encouraging results for full action sequence recognition.

1. Introduction

In recent years, the robotics industry is booming in our world. Advances in robotics research bring robots closer to real world applications. Although robots have become increasingly capable, productive interaction is still restricted to specialists in the field. A major challenge in designing robots for real-world applications is to enable natural and accessible interaction between robots and nontechnical users, while ensuring long-term, robust performance in complex environments without the direct control of a human operator. For the end, robots are required to understand and predict next-step actions of users, and give a suitable response to users. Therefore, early recognition of human activities are very crucial in HRI. However, it is still a challenge problem.

In traditional researches for action recognition, recognitions are performed after fully observing the entire video sequences [1–3]. Features in one video are integrated to one vector without considering their orders for action representation. For early recognition, Ryoo et al. [4] presented a dynamic bag-of-words approach to generate integral histograms of spatio-temporal features for modeling how feature distributions change over time. Barnachona et al. [5] also used the integral histogram for ongoing action representation, and a Dynamic Programming (DP) algorithm was used to compare sub-actions and to compute the recognition score between multiple human action instances. Though these approaches achieved encouraging recognition results, the temporal relationships of action sequences were ignored. Fig. 1 shows us two frame sequences. The top row expresses an action sequence of taking up a cup, and the bottom row is an action of putting down a cup. However, the two action sequences are composed of same frames, but in inverse temporal orders. It indicates that temporal relationships are crucial for human action representation and recognition.

Temporal relationships are frequently represented by pose sequences [6] and integrated silhouette [7] in action recognition. Moreover, local features in a spatio-temporal cuboid [8–10] are also integrated to represent temporal motion of actions. Comparing with local features consisting of hidden temporal information, inspired by the utility of event transition information in facial Action Units (AU) detection [11], Kim et al. [12] presented a temporal segmentation and classification method that accounted for transition patterns between events of interest. They defined event peak patterns manually to segment event patterns and to calculate transition probabilities of connected patterns. Here, action pattern refers to a short frame/pose sequence which represents an action state in action meaning level. For instance, we may separate an action of “drinking” to four consistent patterns, “taking a cup”, “moving cup near the mouth”, “drinking” and “moving cup away”. In HRI, robots are respected to understand human actions and to respond to users promptly. Approaches considering pattern transition is more suitable for HRI. Manually defining action patterns pro-

* Corresponding authors.
E-mail address: yanliji@uestc.edu.cn (Y. Ji).

http://dx.doi.org/10.1016/j.sigpro.2017.06.001
0165-1684/

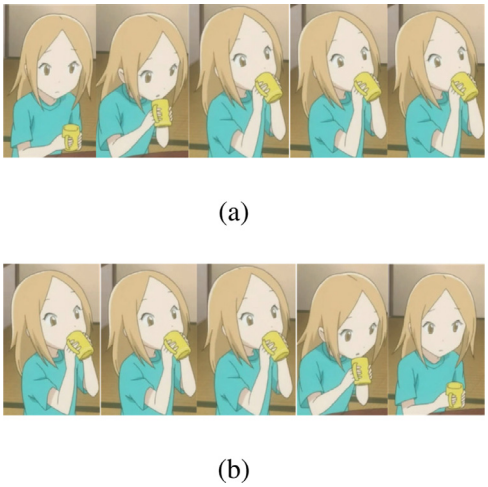

Fig. 1. Are they two actions? (a) Frames in the positive order; (b) Frames in the inverted order.

vides clear transition relationships between patterns in [12]. However, it is impossible to define patterns manually for large quantity of actions. Unsupervised clustering [13–15] and one-shot learning approaches [16,17] provide a suitable solution to define patterns automatically.

Q-learning was first introduced by Watkins [18] in 1989, and Watkins and Dayan [19] presented the convergence proof later. Q-learning can be used to find an optimal action-selection policy for any given (finite) Markov decision process (MDP). It works by learning an action-value function that ultimately gives the expected utility of taking a given action in a given state and following the optimal policy thereafter [20]. Traditionally, Q-learning was applied to solve control problems. A recent application of Q-learning to deep learning, developed by Google DeepMind, has been successful at playing some Atari 2600 games at expert human levels [21]. In this paper, we use the Q-learning to learn temporal relationships of action patterns and represent the relationships with transition rules for action early recognition.

In this paper, we propose a novel approach combining one-shot learning and reinforcement learning for early recognition of human actions. The framework is shown in Fig. 2. To represent the temporal information of human actions, we separate action sequences to several patterns and learn transition maps including temporal ordered patterns and their transition relationships using reinforcement learning algorithm. Since human actions have various categories, varying flexibly, it is difficult to define large quantity of action patterns manually for training. Therefore, different from [12] defining patterns manually, we adopt one-shot learning to automatically separate and define patterns in action sequences. Moreover, we propose a pattern transition map based soft-regression approach for action early recognition. There are two main contributions in this paper. First of all, we adopt one-shot learning to automatically separate and define patterns in action sequences. Secondly, we propose pattern transition maps to represent temporal relationships of pattern sequences, and propose a pattern transition maps based soft-regression approach for action early recognition.

The paper is arranged as following. We explain automatical separation and definition of action patterns using one-shot learning in Section 3. Section 4 expresses how to learn pattern transition maps using Q-learning algorithm. In Section 5, the pattern transition maps based soft-regression approach for action early recognition is explained in detail. Experiment results and discussions are shown in Section 6. Finally, we give conclusion in Section 7.

2. Related work

For a natural and accessible interaction between robots and human in real-world applications, robots are required to understand and predict next temporal activity of users, and give a suitable response. Two main research topics referring to human actions in HRI are human action recognition and action prediction. Among them, action recognition includes the after-the-fact recognition and early recognition.

After-the-fact action recognition refers to recognizing human actions after action information being totally observed. It is a traditional processing for action recognition. Most of recent works belong to after-the-fact recognition [3]. According to features used in recognition, recent works can be separated to action recognitions based on RGB features (including optical flow based approaches) [2,22–24], skeleton features [25–27], depth feature based works [28,29] and confused features [30]. According to applications, we can classify action recognitions to surveillance, sports [31,32], daily life [25], cooking assistant [33], children/elder care [34] and HRI [35], etc. After-the-fact action recognition achieves various results. However, since recognition delay exists, it cannot be used in real-time applications.

Early recognition is an important research topic for real-time action recognition in surveillance and HRI applications. With respect to the after-the-fact recognition, it is very common to integrate all features in an action sequence to one vector or one histogram for action sequence representation. Similarly, for early action recognition, integral approaches are also used to represent the observed action sequences. Ryoo et al. [4] developed a methodology named dynamic bag-of-words to generate integral histograms of spatio-temporal features for modeling how feature distributions change over time, and Barnachona et al. [5] used the integral histogram for ongoing sub-section action representation to realize early recognition. Moreover, Ryoo et al. [36] presented a methodology for early recognition of activities from robot-centric videos (i.e., first-person videos) obtained from a robots viewpoint during its interaction with humans. In addition, Nicolescu et al. [35] presented their works on action intention prediction and users need detection for realistic robot assistance.

The other kind of approaches for early recognition are pattern transition based approaches. Kim et al. [12] presented a temporal segmentation and classification method that accounts for transition patterns between events of interest. On a real HRI platform, Koppula et al. [37,38] represent past action sections and possible future actions using an anticipatory temporal conditional random field (ATCRF) that models rich spatial-temporal relations among human and objects. For real-world HRI applications, action patterns or sections based approaches are more suitable for early recognition and prediction because they are able to describe the varying temporal relationships among human poses, objects and surroundings. Therefore, we separate patterns in action sequences and describe the transition relationship of patterns for early recognition in this paper.

3. Automatical pattern definition using one-shot learning

Human actions are composed of a sequence of continue poses. In order to represent temporal information of actions, we also separate human actions to several short pose sequences like Kim et al. did [12]. One short pose sequence corresponds to one action pattern. In [12], patterns are manually segmented and assigned pattern labels. However, pattern annotation manually is a very heavy work. To overcome the problem, we adopt one-shot learning to define action patterns automatically.

We automatically segment and define action patterns relying on the CS4VM (Cost-Sensitive Semi-Supervised Support Vector Ma-

Please cite this article as: Y. Ji et al., One-shot learning based pattern transition map for action early recognition, Signal Processing (2017), http://dx.doi.org/10.1016/j.sigpro.2017.06.001