Image and Vision Computing 44 (2015) 29–43

Contents lists available at ScienceDirect

Image and Vision Computing

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

Hankelet-based dynamical systems modeling for 3D action recognition

Liliana Lo Presti [a,*], Marco La Cascia [a], Stan Sclaroff [b], Octavia Camps [c]

[a] DICGIM — University of Palermo, V.le delle Scienze, Ed. 6 90128 Palermo, Italy
[b] Computer Science Department — Boston University, 111 Cummington Mall, 02215 Boston, MA, USA
[c] Dept. of Electrical and Computer Eng., Northeastern University, 360 Huntington Ave., 02115 Boston, MA, USA

ARTICLE INFO

Article history:
Received 5 February 2015
Received in revised form 27 July 2015
Accepted 2 September 2015
Available online 22 October 2015

Keywords:
Hidden Markov Model
Hankel Matrix
Linear time invariant system
Discriminative learning
Action

ABSTRACT

This paper proposes to model an action as the output of a sequence of atomic Linear Time Invariant (LTI) systems. The sequence of LTI systems generating the action is modeled as a Markov chain, where a Hidden Markov Model (HMM) is used to model the transition from one atomic LTI system to another. In turn, the LTI systems are represented in terms of their Hankel matrices. For classification purposes, the parameters of a set of HMMs (one for each action class) are learned via a discriminative approach. This work proposes a novel method to learn the atomic LTI systems from training data, and analyzes in detail the action representation in terms of a sequence of Hankel matrices. Extensive evaluation of the proposed approach on two publicly available datasets demonstrates that the proposed method attains state-of-the-art accuracy in action classification from the 3D locations of body joints (skeleton).

1. Introduction

In recent years, a large portion of the research in computer vision has focused on the problem of action recognition and modeling. Detection, recognition and analysis of actions are of great interest in several application domains such as surveillance [1–4], human–computer interaction [5], assistive technologies [6], sign language [7–9], and, more recently, computational behavioral science [10,11] and consumer behavior analysis [12].

The wide diffusion of cheap depth cameras, and the seminal work by Shotton, et al. [13] for estimating the locations of the joints of a human body from depth maps, have given new stimulus to the research in 3D action classification both by quickening the development of novel applications, and by providing a setting to test new ideas and frameworks. Therefore, very recently, we have seen a proliferation of works introducing novel body pose representations for action recognition given depth maps and/or skeleton data [14–20].

In this paper, we propose to represent an action as a series of movements to exploit their temporal structure while discriminating among different action classes. As an example, consider the action of handshaking which can be modeled by the following ordered sequence of movements: moving the whole body to approach the other person, raising the arm, and shaking the hand. Furthermore, each of these movements can be represented as a sequence of observations (for example a sequence of body poses) which are characterized by their own dynamics. Therefore, an action can be represented in terms of a "sequence of simpler dynamics".

This reasoning leads to the idea that an action should be modeled by a hierarchical dynamical model, such as a mixture of Hidden Markov Models (HMMs) [21], coordinated mixture of factor analyzers [22] or switching models [23]. However, the burden of learning the model parameters and the size of the required training set may limit the applicability of these methods.

Here, we propose to approximate the abovementioned complex hierarchical dynamical model by adopting a simpler representation for the movements. In particular, we focus our attention on the switching of the dynamics across time. For this purpose, we represent movements using body motion templates. A body motion template may be either an ordered set of trajectories (i.e. trajectories of body parts such as hands, arms, legs, head, torso) or a sequence of frame descriptors (based on bag-of-words, oriented flow, dense trajectories, etc.) within a temporal window. For simplicity, in the remainder of the paper we will assume that a body motion template is an ordered set of trajectories of 3D body joints within a temporal window. However, our framework may be used with other feature representations as long as they have an ordering relation.

Fig. 1 illustrates the basic idea of our approach. An action is a temporal series of body motion templates (movements). Each body motion template is a series of raw observations in a temporal window (eventually of varying duration) which is characterized by a specific dynamic. Thus, we aim at decomposing an action into sub-trajectories that are modeled as the outputs of a sequence of atomic linear time invariant (LTI) systems, using an HMM to model the transitions from one atomic LTI system to another. Furthermore, each body motion template is

* Corresponding author.
E-mail addresses: liliana.lopresti@unipa.it (L. Lo Presti), marco.lacascia@unipa.it (M. La Cascia), sclaroff@bu.edu (S. Sclaroff), camps@coe.neu.edu (O. Camps).

http://dx.doi.org/10.1016/j.imavis.2015.09.007
0262-8856/

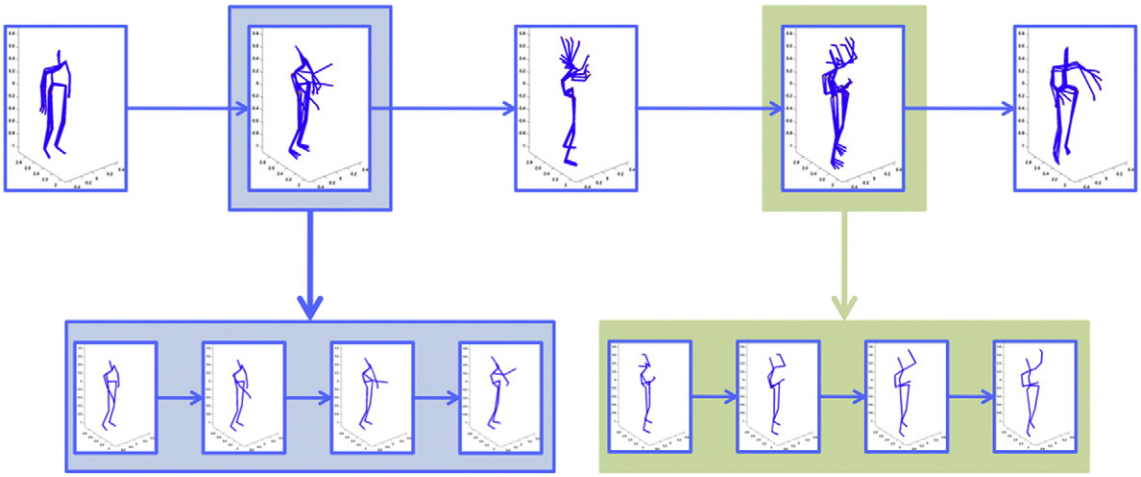

Fig. 1. "Tennis serve" action from the MSRA-3D dataset. An action is a sequence of body motion templates (movements). Each body motion template is, in the case illustrated in this figure, a sequence of 3D Joints Trajectories characterized by their dynamics. The figure shows a sequence of only 5 body motion templates (sub-sampled from the original sequence) and expands only two of them for clarity.

described by means of a truncated Hankel matrix (Hankelet) [24], which embeds the parameters of the LTI system [25]. In summary, an action is modeled by an HMM where the observations are Hankel matrices, computed in a sliding window, and where each hidden state represents an LTI system for which only a Hankelet is known. Finally, for classification purposes, we train a set of HMMs (one for each action class) using a discriminative approach.

Fig. 2 contrasts traditional HMM representations against our approach. Instead of learning the parameters of a switching HMM (Fig. 2(a)), we consider a probabilistic switching LTI system (Fig. 2(b)) where an HMM is used over the Hankel matrices of the systems, avoiding the need of performing any system identification (Fig. 2(c)).

The results presented here are an extension of our preliminary work [26]. In particular, in this paper:

- we account for the learning of the atomic LTI systems via a discriminative method that encourages correct predictions of the HMMs;
- we provide a deeper description of our discriminative learning approach in relation to former models;
- we present an extensive validation of our Hankelet-based action representation for different parameter settings.

The paper is organized as follows: in Section 2 we review previous work on action recognition and modeling, and on discriminative learning of HMM parameters; in Section 3 we present our Hankelet-based action representation; in Section 4 we explain our action model and describe our classification and LTI inference methods; in Section 5 we describe the discriminative learning of the model parameters and of the atomic LTI systems; in Section 6 we present experimental results on publicly available datasets and analyses of the performance of our technique for varying settings and parameters of the Hankelet-based representation. Finally, in Section 7 we present conclusions and outline future research directions.

2. Related work

The literature about action recognition and time series modeling is very extensive. Here, we focus on three main aspects of the methods at the state-of-the-art: action representation, especially for 3D data, modeling of time-varying dynamics, and discriminative learning of parameters. We refer the reader to the following surveys for more general discussions on these topics: [27–31].

2.1. Action representation

Most approaches for human action recognition in still images and RGB video [29] attempt to extract features that may be correlated with the human body pose (human body pose represents the configuration of body parts including head, arms, and legs). Descriptors such as Histogram of Oriented Gradients (HOG) [32], 3D-SIFT [33], Local Binary Pattern (LBP) [34] have been widely used in the literature. Often, a bag-of-words approach is used to compute a histogram of visual words based on a dictionary of local features [35].

Good motion representations can help to discriminate among actions during recognition. Several techniques have tried to combine body pose representation with motion information. Recently, Spatio-Temporal Interest Points (STIPs) [36] and Dense Trajectories (DT) [37, 25], jointly with Motion Boundary (MB) [38], have proved to increase accuracy of action recognition in video sequences.

Since the introduction of depth cameras and the work by Shotton, et al. [13] for estimating the body part locations in depth maps, several researchers have focused on the problem of recognizing actions from depth maps and/or 3D skeletons of the body.

A depth map stores the distance of each point in the scene to the camera. This allows reasoning about body surfaces and shapes across time. Li et al. [39] proposed to use an action graph where each node is a bag of 3D points that encodes the body pose. In Wang et al. [19], a 3D action sequence is treated as a 4D shape and a Random Occupancy Patterns (ROP) feature is extracted. Sparse coding is used to encode only the features that contain information useful for classification purposes. In Vieira et al. [40], space and time axes are divided in cells, and space-time occupancy patterns are computed to represent depth sequences. Oreifej et al. [16] describe the depth sequence as histograms of oriented surface normals (HON4D) captured in the 4D volume, based on depth and spatial coordinates.

The main difficulty of working directly with 3D skeleton data arises from inaccuracy or failures of the skeleton estimation method. Moreover, "one of the biggest challenges of using pose-based features is that semantically similar motions may not necessarily be numerically similar" [41]. Most of the research using only 3D skeleton data tries to extract features to represent the correlation among the locations of the joints. In [15], the body pose is represented by