Neurocomputing 199 (2016) 163–177

Contents lists available at ScienceDirect

Neurocomputing

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

Activity recognition using a supervised non-parametric hierarchical HMM

Natraj Raman *, S.J Maybank

Department of Computer Science and Information Systems, Birkbeck, University of London, United Kingdom

ARTICLE INFO

Article history:
Received 14 August 2015
Received in revised form
5 February 2016
Accepted 8 March 2016
Available online 29 March 2016

Keywords:
Activity classification
Depth image sequences
Hierarchical HMM
HDP
Inference
Multinomial logistic regression

ABSTRACT

The problem of classifying human activities occurring in depth image sequences is addressed. The 3D joint positions of a human skeleton and the local depth image pattern around these joint positions define the features. A two level hierarchical Hidden Markov Model (H-HMM), with independent Markov chains for the joint positions and depth image pattern, is used to model the features. The states corresponding to the H-HMM bottom level characterize the granular poses while the top level characterizes the coarser actions associated with the activities. Further, the H-HMM is based on a Hierarchical Dirichlet Process (HDP), and is fully non-parametric with the number of pose and action states inferred automatically from data. This is a significant advantage over classical HMM and its extensions. In order to perform classification, the relationships between the actions and the activity labels are captured using multinomial logistic regression. The proposed inference procedure ensures alignment of actions from activities with similar labels. Our construction enables information sharing, allows incorporation of unlabelled examples and provides a flexible factorized representation to include multiple data channels. Experiments with multiple real world datasets show the efficacy of our classification approach.

1. Introduction

Activity recognition involves automatic identification of interesting events that occur in a video. It has applications in diverse areas such as video synthesis, smart surveillance and human computer interaction. The recent advent of depth sensing technology that produces depth images in addition to the RGB images has offered opportunities to solve the challenging activity recognition problem. The depth images facilitate robust extraction of the human silhouette and the estimation of a human skeleton's 3D joint positions [1]. High level actions and activities can be inferred from these joint positions.

An activity is typically composed of a set of actions that occur over time. An action in turn is composed of a sequence of skeleton and object poses. The skeleton pose is a particular arrangement of the joint positions and the object pose is a specific representation of an object associated with the action. For example, a rinse-mouth activity may be composed of drink and spit actions. The drink action may involve skeleton poses corresponding to lifting an arm and the object pose may be a representation of a mug. The same pose may be present in different actions and the same action may be present in multiple activities. This composition allows the sharing of data across the activities and the learning of poses and actions from a limited set of examples. Furthermore the activities can now be classified just from the action representations without explicitly taking into account the pose representations. Thus, decomposing an activity into a set of actions and in turn an action into a set of poses enables information sharing and model simplification. The precise definition of the time scale for the actions and activities may depend on the task. In this work, evaluations are performed on fairly simple activities that span less than a minute. The joint positions extracted from a depth image are used for representing the skeleton poses. The object poses are represented using the information in the depth image patches around the joint positions.

A natural way to model a sequence of observations is to use a state-space model such as a Hidden Markov Model (HMM). In an HMM, discrete state variables are linked in a Markov chain by a state transition matrix and observations are drawn independently from a distribution conditioned on the state [2]. A simple HMM is not sufficient in our case, because there are two sequences at different levels – a top level for the coarse action sequence and a bottom level for the granular pose sequence. Intuitively, for a given action state at the top level, we have a sub-HMM conditioned on this state that emits a pose sequence. The hierarchical HMM (H-HMM) captures such a multi-level structure by making each hidden state an autonomous probabilistic model of its own [3].

* Correspondence to: Birkbeck, Malet St, London WC1E7HX, UK.
Tel.: +44 2076316700
E-mail addresses: nraman01@dcs.bbk.ac.uk (N. Raman), sjmaybank@dcs.bbk.ac.uk (S. Maybank).

http://dx.doi.org/10.1016/j.neucom.2016.03.024
0925-2312/

It generates sequences by recursively activating the sub-states of a state. In this context, when an action state is activated, it will use its own probabilistic model to emit a sequence of pose states with a pose state emitting an observation. We can flatten the H-HMM to a standard HMM by introducing a large number of states, however the inference of the activities would become intractable.

In the above H-HMM, the number of action states and pose states must be specified in advance. This is a problem in general with all variants of the classical parametric HMMs where the number of hidden states are fixed a-priori, even though for many applications this number is not known in advance. The usual technique for circumventing this problem is to carry out training using different choices for the number of states and then to apply a model selection criterion. A better approach than this ad hoc procedure is to estimate the correct number of states automatically from the data.

The Dirichlet Process (DP) is a non-parametric Bayesian method used for mixture modeling. It estimates the number of mixture components automatically from data. Its extension, the Hierarchical Dirichlet Process (HDP), is used for modelling groups of data. A mixture model is produced for each group but all the groups share the same mixture components [4]. By drawing parallels from a HMM state to a group in grouped data, the HDP-HMM [5] can be viewed as a non-parametric variant of the classical HMM in which the number of hidden states is inferred from data. This paper uses a non-parametric extension to the H-HMM for modelling activities. The number of action states and the number of pose states are not bounded a priori. During inference these numbers are automatically estimated. The pose states use a factorized representation for the skeleton and object poses.

The above non-parametric H-HMM that models activities cannot be used for classification. A separate H-HMM can be trained for each activity class, but this would prohibit the sharing of actions and poses across the activity classes. In this work, a single H-HMM is trained for all the activities together. In order to perform classification, multinomial logistic regression is used to capture the relationship between the activity labels and the actions. More specifically, the activity labels are regressed on the action states with the regression coefficients learned using a sparsity promoting Laplacian prior [6]. When sampling the action states during inference, the conditional likelihood of actions for a given activity label is incorporated. This ensures that the learnt actions not only explain the observations but also can predict the activity labels.

Our main contribution is the definition of a new factorized non-parametric H-HMM model integrated with multinomial logistic regression. We also propose a tractable inference procedure that is suitable for sequential data and conduct experiments on multiple real world datasets. This proposed model offers the following advantages: (a) the hierarchical composition of actions and poses enables information sharing and model simplification, (b) the non-parametric extension precludes the need for specifying a priori bounds on the number of states, (c) the factorized state representation allows incorporation of multiple data channels and (d) unlabelled examples can be used thus promoting semi-supervised learning. Although our model is generic and can be applied to other hierarchical sequence classification problems, our experiments focus on the activity classification problem. Fig. 1 provides an overview of our approach.

The paper is organized as follows. Section 2 briefly reviews the related work, Section 3 provides relevant background, Section 4 explains the model and Section 5 contains the inference procedure. Experiments conducted on the activity datasets are produced in Section 6 and Section 7 is a conclusion.

2. Related research

Human activity analysis is a very broad research area. The various techniques are reviewed in [7]. We focus specifically on approaches used for activity recognition in depth images. An overview of such approaches can be found in [8–10].

Several activity classification methods rely on computing sophisticated features from the 3D joint positions. In [11], a conjunction of the features for a subset of joints, called an actionlet, is used to represent the interactions between joints. The temporal dynamics are captured using a Fourier temporal pyramid and a multiple kernel learning approach is used for classification. In [12], heterogeneous features are constructed from the skeleton, color and depth patterns of an RGB-D image with the Fourier temporal pyramid used here as well for representing the temporal dynamics. A set of subspaces are then mined from these features in order to perform classification. In [13], a histogram based representation of the joint positions, called HOJ3D, is employed to describe human poses. A discrete HMM is then used to model a low dimensional projection of these features. A similar histogram based representation, but using 2D projections of 3D trajectories for describing displacements, called HOD, is used in [14]. In [15], a spatio-temporal representation, called atomic action template, is composed from key poses that are fed into classification models such as Support Vector Machine (SVM) and Random Forest.

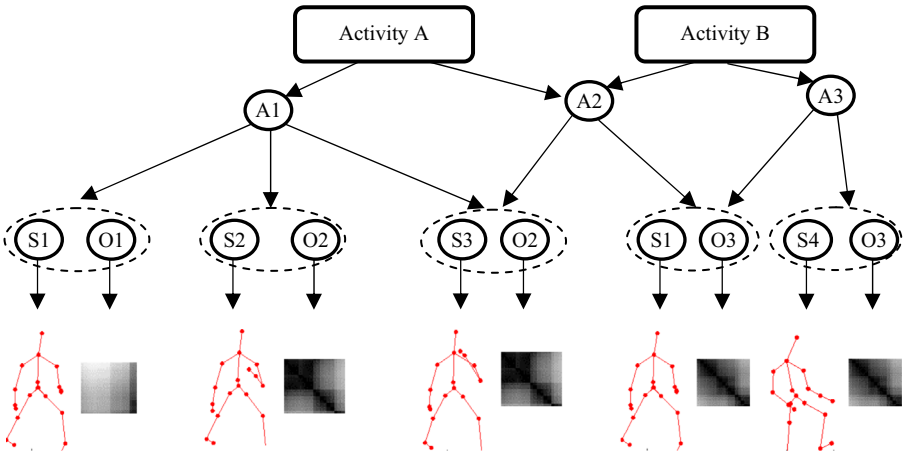

Fig. 1. Activity recognition overview – poses are learnt from observations and actions are learnt from poses. S1 to S4 represent the skeleton pose states, O1 to O3 the object pose states and A1 to A3 the action states. The skeleton poses are based on the 3D joint positions and the object poses are based on the depth image patch around the joint positions. The same pose can be present in multiple actions and different activities can contain the same action. The activities are classified only based on the action states.