Pattern Recognition 81 (2018) 545–561

Contents lists available at ScienceDirect

Pattern Recognition

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

Learning structures of interval-based Bayesian networks in probabilistic generative model for human complex activity recognition

Li Liu[a,b,*], Shu Wang[c,**], Bin Hu[d], Qingyu Qiong[b], Junhao Wen[b], David S. Rosenblum[e]

[a] *Ministry of Education, Key Laboratory of Dependable Service Computing in Cyber Physical Society, Chongqing 400044, China*
[b] *School of Software Engineering, Chongqing University, Chongqing 400044, China*
[c] *Faculty of Materials and Energy, Southwest University, Chongqing 400715, China*
[d] *School of Information Science and Engineering, Lanzhou University, 730000, China*
[e] *School of Computing, National University of Singapore, 117417, Singapore*

ARTICLE INFO

Article history:
Received 19 September 2017
Revised 1 February 2018
Accepted 24 April 2018
Available online 25 April 2018

Keywords:
Complex activity recognition
Structure learning
Bayesian network
Interval
Probabilistic generative model
American Sign Language dataset

ABSTRACT

Complex activity recognition is challenging due to the inherent uncertainty and diversity of performing a complex activity. Normally, each instance of a complex activity has its own configuration of atomic actions and their temporal dependencies. In our previous work, we proposed an atomic action-based Bayesian model that constructs Allen's interval relation networks to characterize complex activities in a probabilistic generative way: By introducing latent variables from the Chinese restaurant process, our approach is able to capture all possible styles of a particular complex activity as a unique set of distributions over atomic actions and relations. However, a major limitation of our previous models is their fixed network structures, which may lead to an overtrained or undertrained model owing to unnecessary or missing links in a network. In this work, we present an improved model that network structures can be automatically learned from empirical data, allowing itself to characterize complex activities with structural varieties. In addition, a new dataset of complex hand activities has been constructed and made publicly available, which is much larger in size than any existing datasets. Empirical evaluations on benchmark datasets as well as our in-house dataset demonstrate the competitiveness of our approach.

1. Introduction

A *complex activity* consists of a set of temporally-composed events of *atomic actions*, which are the lowest-level events that can be directly detected from sensors. In other words, a complex activity is usually composed of multiple atomic actions occurring consecutively and concurrently over a duration of time. Modeling and recognizing complex activities remains an open research question as it faces several challenges: First, understanding complex activities calls for not only the inference of atomic actions, but also the interpretation of their rich temporal dependencies. Second, individuals often possess diverse styles of performing the same complex activity. As a result, a complex activity recognition model should be capable of capturing and propagating the underlying uncertainties over atomic actions and their temporal relationships. Third, a complex activity recognition model should also tolerate errors introduced from atomic action level, due to sensor noise or low-level prediction errors.

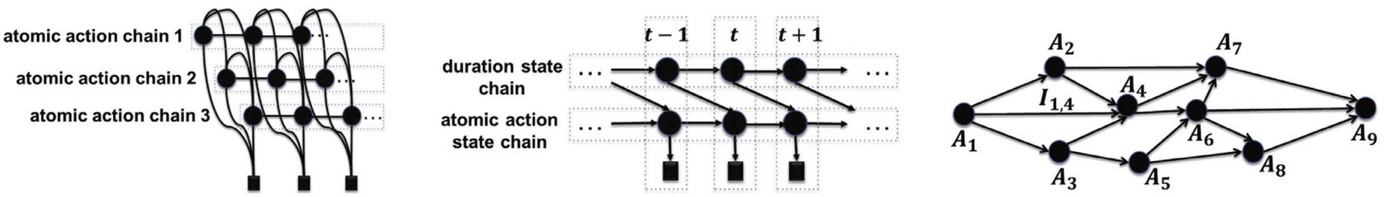
1.1. Related work

Currently, a lot of research focuses on semantic-based complex activity modeling, as using semantic representation has been accredited for its promising performance and desirable ability for human-understandable reasoning [8]. Chang et al. [9] focused on detecting complex events in videos by considering a zero-shot setting where no training data is supplied and evaluating the semantic correlation of each event of interest. Unfortunately, such semantic-based models are capable of representing rich temporal relations, but they often do not have expressive power to capture uncertainties. Many semantic-based models such as context-free grammar (CFG) [38] and Markov logic network (MLN) [22,29] are used to represent complex activities, which can handle rich temporal relations. Yet formulae and their weights in these models (e.g. CFG grammars and MLN structures) need to be manually encoded, which could be rather difficult to scale up and is almost impossible for many practical scenarios where temporal relations among activities are intricate. Although a number of semantic-based ap-

[*] Corresponding author at: School of Software Engineering, Chongqing University, No. 174 Shazhengjie, Chongqing 400044, China.

E-mail addresses: dcsliuli@cqu.edu.cn (L. Liu), shuwang@swu.edu.cn (S. Wang), bh@lzu.edu.cn (B. Hu), xiong03@cqu.edu.cn (Q. Qiong), jhwen@cqu.edu.cn (J. Wen), david@comp.nus.edu.sg (D.S. Rosenblum).

https://doi.org/10.1016/j.patcog.2018.04.022
0031-3203/© 2018 Elsevier Ltd. All rights reserved.



(a) IHMM

(b) DBN

(c) ITBN

Fig. 1. The structures of three graphical models for complex activity recognition. (a) IHMM, where the observations of atomic actions (square-shaped nodes) and several chains of hidden states (round-shaped nodes) are used to handle overlapping; (b) DBN, where duration states and atomic action states are represented as chains of nodes; (c) ITBN, where any atomic-action type (A_1–A_9) is represented by a node and the set of all possible interval relations between any pair of atomic-action types A_i and A_j is represented by a link $I_{i,j}$.

proaches have been proposed for learning temporal relations, such as stochastic context-free grammars [42] and Inductive Logic Programming (ILP) [18], they can only learn formulas that are either true or false, but cannot learn their weights, which hinders them from handling uncertainty.

On the other hand, graphical models become increasingly popular for modeling complex activities because of their capability of managing uncertainties [44]. Unfortunately, most of them can handle three temporal relations only, i.e. equals, follows and precedes. Both Hidden Markov model (HMM) and conditional random field (CRF) are commonly used for recognizing sequential activities, but are limited in managing overlapping activities [12–14,24]. Many variants with complex structures have been proposed to capture more temporal relations among activities, such as interleaved hidden Markov models (IHMM) [31], skip-chain CRF [23] and so on. However, these models are time point-based, and hence with the growth of the number of concurrent activities they are highly computationally intensive [34]. Dynamic Bayesian network (DBN) can learn more temporal dependencies than HMM and CRF by adding activities' duration states, but imposes more computational burden [32]. Moreover, the structures of these graphical models are usually manually specified instead of learned from the data. The interval temporal Bayesian network (ITBN) [44] differs significantly from the previous methods, as being a graphical model that first integrates interval-based Bayesian network with the 13 Allen's relations. Nonetheless, ITBN has several significant drawbacks: First, its directed acyclic Bayesian structure leads ITBN having to ignore some temporal relations to ensure a temporally consistent network. As such, it may result in loss of internal relations. Second, it would be rather computationally expensive to evaluate all possible consistent network structures, especially when the network size is large. Third, neither can ITBN manage the multiple occurrences of the same atomic action, nor can it handle arbitrary network size as it remains unchanged as the count of atomic-action types. Fig. 1 illustrates the graph structures of the three commonly-used graphical models.

It is worth noting that we will focus on complex activity recognition in this paper, and interested readers may consult the excellent reviews [1,5,15–17,25] for further details regarding atomic-level action recognition. Atomic actions are referred to as primitive events that can be inferred from sensors data and images and cannot be further decomposed under application semantics [39]. The interval of a primitive event can also be obtained as the period of time over which the corresponding status remains unchanged. Many excellent approaches have been proposed in the literature to atomic actions (or called events in some papers) which can be inferred from various sources. Chatzis and Kosmopoulos [11] presented a variational Bayesian treatment of multistream fused hidden Markov models, and applied it in the context of active learning based visual workflow recognition for human behavior understanding in video sequences. Simonyan and Zisserman [33,40] built deep convolutional neural networks to estimate the upper body pose of humans in gesture videos, and also presented a two-stream architecture of deep convolutional networks, which incorporates spatial and temporal networks for action recognition in video. Chang et al. [10,19] presented a semantic pooling approach for event detection, recognition, and recounting in videos by defining semantic saliency that assesses the relevance of each shot with the event of interest.

Our model focuses on the representations of a complex activity with diverse combinations of atomic actions and their temporal relations under uncertainty, assuming that atomic actions are already recognized and labeled in advance. In the field of human activity recognition, it is increasingly important to understand how those representations work and what they are capturing [21]. Unlike the complex activity recognition methods that operate directly on raw values such as sensor data and video clips, the atomic action-based methodology provides an intermediate space between low-level raw data and high-level complex activities, thereby freeing it from dependence on a particular source modality. The atomic action-based recognition system may have several benefits: the ability to operate on a range of data sources, allowing the knowledge of activities to be shared across a wide range of possible sensor modalities and camera types [3]; the ability to handle across a range of activities efficiently, attenuating the complexity of recognizing complex activities directly from raw source data; and the ability to be reused without needing a special configuration or retraining, regardless of what and how many sources are involved. In our experiments, we adopt the existing approaches to detect atomic actions from different sources of raw data, such as motion sensors [5] and videos [7,44].

1.2. Our approach

To address the problems in the existing models, we presented the ***g**enerative **p**robabilistic model with **A**llen's interval-based relations* (GPA in short) to explicitly model complex activities, which is achieved by constructing probabilistic interval-based networks with temporal dependencies. In other words, our model considers a probabilistic generative process of constructing interval-based networks to characterize the complex activities of interests. Briefly speaking, a set of latent variables, called *tables*, which are generated from the *Chinese restaurant process* (*CRP*) [35] are introduced to construct the interval-based network structures of a complex activity. Each latent variable characterizes a unique *style* of this complex activity by containing its distinct set of atomic actions and their temporal dependencies based on Allen's interval relations. There are two advantages to using *CRP*: It allows our model to describe a complex activity of arbitrary interval sizes and also to take into account multiple occurrences of the same atomic actions. We further introduce *interval relation constraints* that can guarantee the whole network generation process is globally temporally