Pattern Recognition Letters 43 (2014) 62–70

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Context augmented Dynamic Bayesian Networks for event recognition

Xiaoyang Wang *, Qiang Ji

Dept. of ECSE, Rensselaer Polytechnic Institute, 110 Eighth Street, Troy, NY 12180, USA

ARTICLE INFO

Article history:
Available online 5 August 2013

Communicated by Ken-ichi Maeda

Keywords:
Context model
Event recognition
Dynamic Bayesian Networks
Video surveillance
Probabilistic Graphical Model

ABSTRACT

This paper proposes a new Probabilistic Graphical Model (PGM) to incorporate the scene, event object interaction, and the event temporal contexts into Dynamic Bayesian Networks (DBNs) for event recognition in surveillance videos. We first construct the baseline event DBNs for modeling the events from their own appearance and kinematic observations, and then augment the DBN with contexts to improve its event recognition performance. Unlike the existing context methods, our model incorporates various contexts simultaneously into one unified model. Experiments on real scene surveillance datasets with complex backgrounds show that the contexts can effectively improve the event recognition performance even under great challenges like large intra-class variations and low image resolution.

1. Introduction

The topic of modeling and recognizing events in video surveillance system has attracted growing interest from both academia and industry (Oh et al., 2011). Various graphical, syntactic, and description-based approaches (Turaga et al., 2008) have been introduced for modeling and understanding events. Among those approaches, the time-sliced graphical models, i.e. Hidden Markov Models (HMMs) and Dynamic Bayesian Networks (DBNs), have become popular tools.

However, surveillance video event recognition still faces difficulties even with the well-built models for describing the events. The first difficulty arises from the tremendous intra-class variations in events. The same category of events can have huge variations in their observations due to visual appearances differences, target motion variations, viewpoint change and temporal variability. Also, the low resolution of event targets also affects event recognition. To compensate for such challenges, we propose to capture various contextual knowledge and systematically integrate them with image data using a Probabilistic Graphical Model (PGM) (Koller and Friedman, 2009) to improve the performance of event recognition on challenging surveillance videos.

Contextual knowledge can be regarded as one type of extra information that does not directly describe the recognition task, but it can support the task. As an additional information that can capture certain temporal, spatial or logical relationships with the recognition target, context plays an important role for various visual recognition tasks. Various contexts that are available/retrievable both during training and testing are widely used in many approaches. For example, Yao and Fei-Fei (2010) and Yao and Fei-Fei (2012) propose a context model to make human pose estimation task and object detection task as mutual context to help each other. Also, Ding et al. (2012) use both the local features and the context cues in the neighborhood windows to construct a combined feature level descriptor for pedestrian detection. For action recognition, existing work integrates contexts both as features and as models. Several approaches such as Kovashka et al. (2010) and Wang et al. (2011) integrate contexts into spatial or spatial–temporal features. On the other hand, several approaches (Gupta et al., 2007; Li et al., 2007; Yao and Fei-Fei, 2010; Choi et al., 2011) incorporate contexts into the pattern recognition models to capture the interactions between actions, objects, scene and poses.

Different from the previous approaches integrating contexts into static models, we propose a Probabilistic Graphical Model that simultaneously incorporates various contexts into the dynamic DBN model for event recognition. Inspired by a number of recognition frameworks that exploit the scene context (Russell et al., 2007; Marszalek et al., 2009; Li et al., 2007; Oh et al., 2010) and the object-action interaction context (Gupta et al., 2007; Li et al., 2007; Yao and Fei-Fei, 2010) in different applications, we apply both the scene context and the event-object interaction context into our model. Moreover, we propose to capture the event temporal context, which describes the semantic relationships of events over time. Experiments on real scene surveillance videos show that using either the object interaction context or the scene and event temporal contexts alone can already effectively improve the event recognition performance. Moreover, with the combination of three contexts, the event recognition performance can be significantly improved even under great challenges like large intra-class variations and low image resolution.

* Corresponding author. Tel.: +1 518 276 6040.
E-mail addresses: wangx16@rpi.edu (X. Wang), jiq@rpi.edu (Q. Ji).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.07.015

In summary, the novelty of this paper includes the following: (1) it proposes a PGM model that simultaneously incorporates the scene, event-object interaction, and the event temporal contexts into the baseline DBN model. (2) it introduces the event temporal context which describes the semantic relationships of events over time.

2. Related work

Visual event recognition, which is defined as the recognition of semantic spatio-temporal visual patterns such as "getting into vehicle", and "entering a facility" in Oh et al. (2011), is an important pattern recognition problem in the computer vision application. Much existing related work (Wang et al., 2011; Zhu et al., 2013) has been focusing on recognition of basic human action/activities (like "walking", "turning around" etc.) in clean backgrounds using datasets such as KTH (Schuldt et al., 2004), Weizmann (Gorelick et al., 2007) and HOHA (Laptev et al., 2008). Comparatively, we focus on the event recognition task that generally involves the interaction of persons and objects like vehicles and facilities in the real scene surveillance videos with complex backgrounds as in Oh et al. (2011). Due to the low resolution of the videos and the large intra-class variations of event appearances, the event recognition on these videos are rather challenging. However, with context knowledge augmentation, we can still improve the event recognition performance on these videos.

Different types of DBNs have been built for recognizing different actions/activities. Standard HMM is employed for modeling simple action/activity in Yamato et al. (1992) and Zhang et al. (2005). For modeling more complex activities, different variants of HMM like Parallel HMMs (PaHMMs) (Vogler and Metaxas, 2001), Coupled HMM (CHMM) (Oliver et al., 2000) and dynamic multiply-linked HMM (DML-HMM) (Xiang et al., 2006) are proposed. Since these variants of HMM are still restricted by the specific model structure, more general DBNs are further built for modeling actions/activities. Wu et al. (2007) propose to combine video data and RFID, and formalize a DBN that is essentially a layered HMM with multiple observations. Laxton et al. (2007) build a hierarchical DBN to recognize complex activities. Based on these progresses, we propose to develop a DBN that combines the target appearance and kinematic states with various context information.

Context in recognition problems is generally regarded as extra information that is not the recognition task itself, but it can support the task. Context knowledge has become very important to help object and action recognition problems. A comprehensive review on context based object recognition is given in Galleguillos and Belongie (2010). In object recognition tasks, PGM has shown its power for integrating contexts such as scenes (Russell et al., 2007), co-occurrence objects (Rabinovich et al., 2007), and materials (Heitz et al., 2008). For action, activity and event recognition, contexts are integrated both as features and as models. Work such as Kovashka et al. (2010) and Wang et al. (2011) integrates contexts with spatial or spatial–temporal features. On the other hand, approaches in Gupta et al. (2007), Li et al. (2007), Yao and Fei-Fei (2010) and Choi et al. (2011) incorporate contexts into static models to build the interactions between actions, objects, scene and poses. Inspired by these existing approaches incorporating contexts into the recognition models, we propose a PGM model that can jointly incorporate the scene, event-object interaction, and the event temporal contexts simultaneously into the baseline DBN model. Different from the existing approaches integrating contexts into static models, we simultaneously incorporate various contexts into the dynamic DBN model.

Many different sources of contexts are discussed in the literature (Biederman, 1981; Oliva and Torralba, 2007; Strat, 1993), and have been applied to many applications (Russell et al., 2007; Li et al., 2007). Divvala et al. (2009) give an empirical study about the taxonomy of sources of contexts for object detection, and catalog 10 possible sources of context that could be available to a vision system. As to the scene context, Russell et al. (2007) and Oh et al. (2010) use scene context to impose spatial priors on the locations of objects and activities respectively. Also, Marszalek et al. (2009) propose to utilize the correlation between human actions and particular scene classes to benefit the human action recognition. As to the object-action interaction context, Gupta et al. (2007) present a Bayesian approach for combining action understanding with object perception; Li et al. (2007) introduce a model to classify events in static images by integrating object categories; Filipovych and Ribeiro (2008) propose a probabilistic framework that models the joint probability distributions about actor and object states. Based on the existing research on the scene context and object-action interaction context, we propose a new PGM based context model to simultaneously incorporate these two sources of contexts. Moreover, we introduce a third source of context, i.e. the event temporal context, to capture the semantic relationships of events over time. Our final proposed context model can incorporate the three sources of contexts simultaneously.

We presented a preliminary version of this work in Wang and Ji (2012). The model discussed in this paper is based on Wang and Ji (2012), and differs from Wang and Ji (2012) in the following aspects. (1) We present our overall event recognition approach, and describe more details about each pre-processing part including the feature extraction. (2) We include a detailed discussion about the baseline DBN model learning and inference. (3) We add a complete discussion about the MAP learning of our context model. (4) We incorporate a new experimental comparison with other models using contexts. (5) We add a thorough comparison with the related methods in this section.

3. Overall event recognition approach

In this paper, we propose to use a Probabilistic Graphical Model that incorporates various contexts into the dynamic DBN model for event recognition. Fig. 1 gives the overall framework of our event recognition system. During training, given multiple training clips, target detection and tracking is performed first, and then the system extracts features from the tracks of the training clips. After the AdaBoost (Freund and Schapire, 1995) based feature selection, the baseline event DBN models are learned based on the selected features. Also, we utilize the static object, scene and event temporal relation information in the training data as contextual information to train the context model. During testing, given a query track, we extract the features, and then use the baseline DBNs in the event DBN library to obtain the event measurements. These event measurements are combined with the scene measurement, object measurement and the previous event prediction to jointly infer the current event type using the proposed context model.

The paper is organized as follows: We will briefly discuss feature extraction in Section 4. This is then followed by describing our baseline DBN model in Section 5. Section 6 will focus on discussing the unified context model. We present the experiment evaluations in Section 7, and the conclusion in Section 8.

4. Event feature extraction

The first step of our event recognition system is target detection and tracking, where static objects (e.g. parked vehicles) and the moving targets are respectively detected and tracked through sliding window detection and Kanade–Lucas–Tomasi (KLT) tracker (Lucas et al., 1981). The feature vectors across the track intervals are used as inputs to the models for both training and testing.