Incorporating Contextual Knowledge to Dynamic Bayesian Networks for Event Recognition

Xiaoyang Wang and Qiang Ji
Dept. of ECSE, Rensselaer Polytechnic Institute, USA
{*wangx16, jiq*}*@rpi.edu*

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 event DBNs for modeling the events from their own appearance and kinematic observations, and then extend the DBN to incorporate the contexts for boosting event recognition performance. Unlike the existing context methods, our model incorporates various contexts into one unified model. Experiments on natural scene surveillance videos show that the contexts can effectively improve the event recognition performance even with 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 [12]. Various graphical, syntactic, and description-based approaches [16] 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 such as the visual appearance, speed of motion, viewpoints and temporal variability. Also, the low resolution of event targets also affects event recognition. To compensate such challenges, we propose to incorporate contextual knowledge to our DBN models with a Probabilistic Graphical Model (PGM) [6] to aid the event classification task.

Context in recognition problems can be 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 [2]. In object recognition tasks, PGM has shown its power for integrating contexts such as scenes [14], co-occurrence objects [13], and materials [5]. As to action recognition, contexts are integrated both as features and as models. Works such as [7, 17] integrate contexts with spatial or spatial-temporal features. On the other hand, many works [4, 9, 18, 1] incorporate contexts to model the interactions between actions, objects, scene and poses.

Different from approaches integrating contexts into static models, we propose a PGM model that incorporates various contexts into the dynamic DBN model for event recognition. Inspired by a number of recognition frameworks that exploit the scene context [14, 10, 9, 11] and the object-action interaction context [4, 9, 18] in different applications, we apply both the scene context and the event-object interaction context into our model. Moreover, we proposed the usage of the event temporal context, which describes the semantic relationships of events over time.

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

2 Baseline DBN Event Model

As shown in Figure 1, our baseline DBN model for event recognition consists two layers. The top layer includes two hidden nodes GM and SA respectively. The GM node represents the global motion state, and the SA node represents the shape and appearance state. The bottom layer consists of two measurement nodes OGM and OSA. The OGM node denotes the kinematic fea-

tures extracted from the global motion measurements. The OSA node denotes the HoG and HoF image features extracted from the appearance measurements.

Besides the nodes, there are two types of links in the model: intra-slice links and inter-slice links. The intra-slice links couple different states to encode their dependencies. And the inter-slice links represent the temporal evolution and capture the dynamic relationships between states at different times. The proposed DBN event model is essentially a coupled HMM.

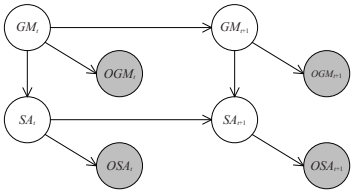

Figure 1: The baseline DBN model for event recognition.

For the parameter learning of DBN models, because of the presence of the hidden nodes (GM and SA), the Expectation Maximization (EM) method is employed to estimate the parameters from the training data. We can obtain parameters of K models for the K events to be recognized. Here, we denote the learned K models to be $\hat{\Theta}_1, \hat{\Theta}_2, ..., \hat{\Theta}_K$.

In DBN inference, suppose the evidence of the testing sequence to be E_T. For the cth model where $c \in [1, K]$, we need to infer the likelihood of an event given the evidence E_T, i.e., $P(E_T|\hat{\Theta}_c)$. These likelihoods are evaluated by the forward propagation of dynamic junction tree.

3 Context for Event Recognition

The context is often defined as the surroundings, circumstances, environment, background or settings which help determine, specify, or clarify the meaning of an event. We proposed to use three contexts for event recognition in surveillance videos: scene context, event-object interaction, and the temporal context. These contexts are combined into one unified model to improve the baseline DBN for event recognition.

- Scene Context
  Events in surveillance videos are frequently constrained by properties of scenes and demonstrate high correlation with scene categories. E.g. events in parking lot are different from events in playground. Knowledge of the scene can provide a prior probability of the events.
- Event-object Interaction
  Static objects provide clues for events. E.g. person "walking" on *sidewalk*, and person "getting out of vehicle" aside *vehicle* shown in Fig. 2a.
- Temporal Context
  Event occurrence is also constrained by the natural temporal causalities of executing events. E.g. event "unloading" *follows* event "getting out of vehicle", while event "walking" *precedes* event "getting into vehicle", as shown in Fig. 2b.

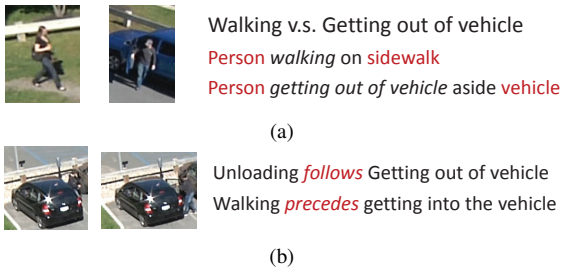

Figure 2: Events with contexts.

3.1 Context Model Formulation

We systematically incorporated the three contexts into one unified model, and use this model to infer the posterior probability of events in different conditions. The model graph is shown in Fig. 3.

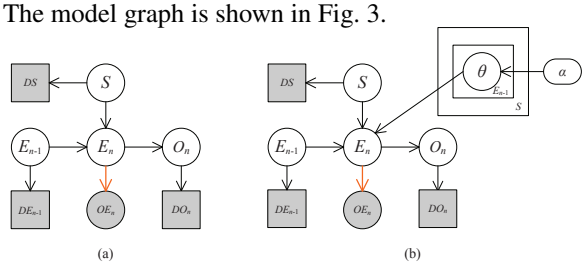

Figure 3: PGM model combining the scene, temporal and object interaction contexts. (a) model without hyper parameter; (b) model with hyper parameter.

As shown in Figure 3, event nodes E (both E_{n-1} and E_n) have K discrete values where K stands for the K different categories of events. The subscripts $n-1$ and n on E nodes stands for the events at two different times, where E_{n-1} stands for the previous event, and E_n stands for the current event. The link between E_{n-1} and E_n captures the temporal dependency, i.e. the temporal context. The S node stands for the scene. The link between S and E_n captures the influence of the scene context on event. The O_n nodes stands for the contextual object for current event clip. The link between E_n and O_n captures the event and object interaction context. During testing, all these four nodes E_n, E_{n-1}, S and O_n are latent, with their corresponding measurement nodes OE_n, DE_{n-1}, DS and DO_n to indicate their states.

The circular OE_n node is a continuous vector representing the appearance observation for the current event; the link from E_n (with a node value $c \in [1, K]$) to OE_n is captured by the probability $P(OE_n|E_n = c)$,

which we take from the output of the DBN model, i.e., $P(E_T|\hat{\Theta}_c)$. This allows naturally incorporating the baseline DBN model into the context model. The remaining three measurement nodes DE_{n-1}, DS and DO_n are discrete nodes resulted from the classifier detections of the corresponding contexts. The parameters of the context model in Fig. 3(a) are estimated using the maximum likelihood method.

In Fig.3(b), a conjugate prior is further added to the node E_n to handle the cases of limited training samples for learning the event transitions $P(E_n|E_{n-1}, S)$ in different scenes. Hence, the parameter learning of this model is performed in a MAP process with the conjugate priors added as in Fig.3(b).

3.2 Event Recognition with Context Model

In the usual case where all three contexts are available, we would need to infer the marginal probability $P(E_n|OE_n, DO_n, DS, DE_{n-1})$. Its factorization can be written as

$$P(E_n|OE_n, DO_n, DS, DE_{n-1}) \propto P(OE_n|E_n) \cdot \sum_{O_n}\{P(O_n|E_n)P(DO_n|O_n)\} \cdot \sum_{S,E_{n-1}} \{P(E_n|S, E_{n-1}) \cdot P(S)P(DS|S)P(E_{n-1})P(DE_{n-1}|E_{n-1})\} \quad (1)$$

In practice, such inference can be solved with variable elimination [6]. The classification finds the class label c^* that maximizes the the posterior probability as

$$c^* = \arg\max_c P(E_n = c|OE_n, DO_n, DS, DE_{n-1}) \quad (2)$$

Moreover, we observe that, when measurements for certain contexts are missing, the inference using the model in Fig. 3 would naturally degrade to the model combining the baseline DBN with only existing context(s). This can be proved by marginalizing the joint probability distribution over the missing context measurements.

4 Experiments

We use two surveillance datasets to verify our proposed methods. The first dataset is the VIRAT aerial dataset (ApHill) [12]. We choose to recognize eight events from this dataset. They are: *Loading a Vehicle* (LAV), *Unloading a Vehicle* (UAV), *Opening a Trunk* (OAT), *Closing a Trunk* (CAT), *Getting into a Vehicle* (GIV), *Getting out of a Vehicle* (GOV), *Entering a Facility* (EAF), *Exiting a Facility* (XAF). Computed tracks with event category labels are used for both training and testing with five fold cross validations.

The other dataset is the VIRAT public 1.0 dataset [12]. There are in total six types of events which form a subset of selected events in VIRAT aerial dataset. The events are LAV, UAV, OAT, CAT, GIV and GOV respectively. Currently, the event category labels of the VIRAT public 1.0 testing dataset are not released. Therefore, in our experiments, we only use its training dataset with five fold cross validations.

These two datasets are very challenging compared to other event/activity/action datasets like KTH [15], Weizmann [3], and HOHA [8]. The two datasets are collected in real natural scenes with low resolution surveillance videos. Also, they focus on complex events which include the interactions between persons and vehicles. These complex events are more difficult to recognize than the simple events like walking or running.

4.1 DBN with Object Interaction Context

In this experiment, we show that event recognition is improved by object context using our context model. The VIRAT aerial dataset is used, and vehicle is chosen as the object that person interacts with, where the vehicle detector receives 62.43% recall rate and 33.87% false alarm rate on this dataset. The event recognition performance on each event is given as area under ROC curve shown in Table 1.

Table 1: ROC area comparison for event recognition with object context in VIRAT aerial dataset.

Events	Baseline DBN	DBN with Vehicle Context
LAV	0.5510	**0.5686**
UAV	**0.7419**	0.7050
OAT	0.4763	**0.6003**
CAT	0.6225	**0.7190**
GIV	0.6300	**0.7386**
GOV	0.6844	**0.6939**
EAF	**0.7685**	0.7561
XAF	0.7251	**0.7389**
Average	0.6499	**0.6900**

With vehicle context, the event recognition performance in VIRAT aerial dataset improves on six of the eight events, with a 4% improvement on average area of ROC curves. Beside the ROC curves, we also evaluate the overall recognition rate on these eight events. The recognition rate improves from 32.34% to 37.13% with the help of vehicle context.

4.2 DBN with Scene and Temporal Context

The training dataset in VIRAT public 1.0 dataset contains three different scenes of parking lots. There are respectively 5, 28 and 19 video sequences in each scene. The event sequences in each of these videos are closely related temporally. Thus, we use such dataset to verify both the scene and the temporal context. We apply the scene context, the temporal context and their

combinations. For scene context, our global scene classifier reaches 95.74% recognition rate on recognizing these three scenes. The area under ROC curve performance comparisons are shown in Table 2.

In Table 2, the global scene context can slightly improve the DBN overall performance by over 1%. The temporal context, which describes the event temporal relationships, can improve the DBN overall performance by close to 4%. If we incorporate both contexts with the DBN model, we can receive above 6% improvement over the baseline DBN on average area of ROC curves. In addition, we evaluated the overall recognition rate on these six events. With the help of both contexts, the recognition rate improves from 26.09% to 35.64%.

Table 2: ROC area comparison with scene and temporal contexts in VIRAT public 1.0 dataset.

Events	DBN	DBN + Scene	DBN + Temporal	DBN + Temporal + Scene
LAV	0.5501	0.5819	0.5876	**0.6507**
UAV	0.6192	**0.6715**	0.4757	0.4760
OAT	0.6157	0.4618	0.7186	**0.7291**
CAT	0.5911	0.6559	**0.7446**	0.7409
GIV	0.5416	0.5420	0.5207	**0.5674**
GOV	0.5364	0.6486	0.6316	**0.6601**
Average	0.5757	0.5936	0.6131	**0.6374**

4.3 DBN with Three Contexts Combined

We show the results of utilizing the event-object interaction context, the scene context, and the temporal context simultaneously in classifying the test sequences. We set the VIRAT aerial data set happened in the fourth scene, and then combine the VIRAT aerial data set with the VIRAT public 1.0 dataset. With the combined dataset, the task becomes even more challenging with more variations brought in. In Fig. 4, we present the overall recognition rates with five fold cross validations. An overall 9% improvement can be realized utilizing all contexts.

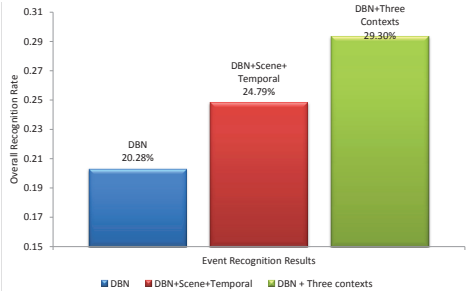

Figure 4: Event recognition accuracy with different combined contexts on combined data.

5 Conclusion

In this work, we focus on event recognition in surveillance videos. The events in surveillance videos are challenging due to large intra-class variation, low image resolution, error and missed tracking etc. To handle such challenges, we propose a probabilistic context model that systemically combines a baseline DBN event model with three types of contexts: the object, scene and event temporal contexts. Experiments on real surveillance videos show that the contexts can effectively improve the event recognition performance.

References

[1] W. Choi, K. Shahid, and S. Savarese. Learning context for collective activity recognition. In *CVPR*, 2011.

[2] C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6):712 – 722, 2010.

[3] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Trans. on PAMI*, 29(12):2247 –2253, 2007.

[4] A. Gupta and L. Davis. Objects in action: An approach for combining action understanding and object perception. In *CVPR*, 2007.

[5] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *ECCV*. 2008.

[6] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

[7] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *CVPR*, 2010.

[8] I. Laptev, M. Marszalek, and etc. Learning realistic human actions from movies. In *CVPR*, 2008.

[9] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *ICCV*, 2007.

[10] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009.

[11] S. Oh and A. Hoogs. Unsupervised learning of activities in video using scene context. In *ICPR*, 2010.

[12] S. Oh, A. Hoogs, and etc. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR*, 2011.

[13] A. Rabinovich, A. Vedaldi, and etc. Objects in context. In *CVPR*, 2007.

[14] B. Russell, A. Torralba, and etc. Object recognition by scene alignment. In *NIPS*. 2007.

[15] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR*, 2004.

[16] P. Turaga, R. Chellappa, and etc. Machine recognition of human activities: A survey. *IEEE Trans. on Circ. and Sys. for Video Tech.*, 18(11):1473–1488, 2008.

[17] J. Wang, Z. Chen, and Y. Wu. Action recognition with multiscale spatio-temporal contexts. In *CVPR*, 2011.

[18] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.