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

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).
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.
There are two types of input features, i.e. kinematic features and appearance features. The kinematic features are obtained by first filtering the position measurements of a given track interval through a sliding window Least-Squares filter that has a constant velocity dynamics model. The Least-Squares filter produces filtered state estimates of 2D position and velocity in the ground plane. These estimates are then used to derive six scene independent kinematic features: speed, heading, change-in-heading, range, entropy of change-in-heading, and curvature. These kinematic features reflect the global motion of the tracked target.

We use the HoG (histogram of oriented gradient) and HoF (histogram of optical flow) features (Dalal and Triggs, 2005) as the target appearance features. The histogram votes are accumulated into 8 evenly spaced bins of pixel gradient orientations or optical flow directions for HoG or HoF features respectively. We extract features through the histograms in the scale of 1-by-1, 2-by-2 and 3-by-3 blocks in the image bounding boxes of the tracks.

With the extracted kinematic and appearance features, two AdaBoost (Freund and Schapire, 1995) classifiers are used to select the most discriminative kinematic and appearance features respectively. AdaBoost algorithm is executed on all features in each of the two feature pools, where a decision stump (Iba and Langley, 1992) is used as a weak classifier. Since the decision stump chooses the single most discriminative feature that minimizes the overall training error, a count of the selected features results in a ranking of the most discriminative features upon completion of AdaBoost. The top $N$ ($N=20$) most discriminative kinematic and appearance features are selected respectively on a per event basis and are used as the continuous inputs to the observation nodes of the baseline DBN model.

5. Baseline DBN model

Given the event features, we develop the DBN model for event recognition. The DBN model captures the spatial and temporal pattern of each event.

5.1. DBN model structure

As shown in Fig. 2, 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 features 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 model is essentially a coupled HMM, that captures dynamic interaction between target motion and its appearance over time. In our baseline DBN model, the two hidden nodes $GM$ and $SA$ are discrete. They represent the hidden states for the target motion and appearance respectively. For the baseline DBN model, there are three kinds of parameters: the initial state distribution for the $GM$ node, the intra-slice state conditional probability of $SA$ given $GM$, and the inter-slice state transition probability both from $GM_t$ to $GM_{t+1}$, and from $SA_t$ to $SA_{t+1}$. Our observations $OGM$ and $OSA$ are all continuous observations. Given their corresponding hidden states, we assume the observations of $OGM$ and $OSA$ follow Gaussian distribution.

To model $K$ types of events, we build in total $K$ such DBN models, each corresponding to one type of event.
5.2. Parameter learning of baseline DBN

Suppose the training sequences for event $c$ ($c \in [1, K]$) is $\mathcal{E}_c$, which contains $N_c$ sequences, i.e. $\mathcal{E}_c = (E_1, E_2, \ldots, E_{N_c})$. As DBN usually captures the joint distribution of sequence of variables, it is typically learned by maximizing the log likelihood of all training sequences. Let the learned parameter for the DBN model of event $c$ to be $\Theta_c$, then we can write the general learning principle as:

$$\hat{\Theta}_c = \arg \max_{\Theta_c} \log P(\mathcal{E}_c | \Theta) = \arg \max_{\Theta_c} \sum_{t \in \mathcal{E}_c} \log P(E_t | \Theta)$$

(1)

Because of the presence of hidden nodes (GM and SA), the Expectation Maximization (EM) method is employed to estimate the parameters from the training data. Since the EM method is sensitive to the initialization of the Gaussian observation node (OGM and OSA) parameters, the initialization process uses the k-means clustering algorithm to identify the starting mean vector and covariance matrix for each class. After the parameter initialization step, the DBN parameters are learned iteratively through a DBN EM (Murphy, 2002), where a junction tree inferring engine is used for the expectation step.

With the EM method, we can get the parameters of $K$ DBNs to model the $K$ events. Here, we denote the learned $K$ models to be $\Theta_1, \Theta_2, \ldots, \Theta_K$.

5.3. Inference of baseline DBN

In DBN inference, suppose the evidence of the testing sequence to be $E_c$. For the $c$th model where $c \in [1, K]$, we need to infer the likelihood of an event given the evidence $E_c$, i.e. $P(E_c | \Theta_c)$. These likelihoods are evaluated by forward propagation (Rabiner, 1989) which is a precise and fast inference method suitable for our DBN model. The classification with baseline DBN model can be written as:

$$\hat{c} = \arg \max_{c \in [1, K]} P(E_c | \hat{\Theta}_c)$$

(2)

6. Contexts for event recognition

Context is often defined as the surroundings, circumstances, environment, background or settings which help determine, specify, or clarify the meaning of an event. It is a type of additional information that is available or can be retrieved both during training and testing.

6.1. Event contexts

We propose to use three types of contexts for event recognition in surveillance videos: scene context, event-object interaction context, and the event 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 (would include events like “loading a vehicle”, “getting out of a vehicle” etc.) are different from events in playground (would include events like “jogging”, “running” etc.). Knowledge of the scene categories can hence provide a prior probability of the events when utilized as scene context.

- **Event-object interaction context**
  Static objects that interact with the events would provide valuable clues for recognizing the events. E.g. person “walking” on sidewalk, and person “getting out of vehicle” aside vehicle, as shown in Fig. 3(a). The interacting objects sidewalk and vehicle would provide valuable information to judge whether the event is “walking” or “getting out of vehicle”, even if the kinematic and appearance properties of the target persons are similar. We can utilize these interacting objects as event-object interaction context to help infer the event type.

- **Event temporal context**
  Event occurrence is also constrained by the natural temporal causalities of executing events. E.g. event “closing a trunk” typically follows event “loading/unloading a vehicle”, while event “walking” often precedes event “getting into vehicle”, as shown in Fig. 3(b). We utilize the previous event as event temporal context to provide a clue for the current event prediction.

6.2. Context model formulation

Different from the existing context models, we introduce a unified model to systematically incorporate the three contexts, and use this model to infer the posterior probability of events in different conditions. The model graph is shown in Fig. 4(a).

As shown in Fig. 4(a), 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 event temporal context.

The $S$ node stands for the scene. The link between $S$ and $E_n$ captures the causal 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-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 $O_{E_n}, O_{E_{n-1}}, D_{S}$ and $D_{O_n}$ as estimates of their states.

The circular $O_{E_n}$ node is a continuous vector representing the event observation for the current event; the link from $E_n$ (with a node value $c \in [1, K]$) to $O_{E_n}$ is captured by the probability $P(O_{E_n} | E_n = c)$, which we take from the output of the DBN model, i.e. $P(E_n | \Theta_c)$ in Eq. 2. This allows naturally incorporating the baseline DBN model into the context model. The remaining three measurement nodes $D_{E_{n-1}}, DS$ and $DO_n$ are discrete nodes resulted from the classifier detections of the corresponding contexts. The parameters of the context model in Fig. 4(a) are estimated using the maximum a posteriori (MAP) approach as discussed in Section 6.3.

6.3. Learning the context model

In the learning of our context model, 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 context model is per-
formed in a maximum a posteriori (MAP) manner with the conjugate prior added as shown in Fig. 4(b).

In the MAP learning of the context model, given a training set D containing scene information, event-object interaction information, the previous and current event labels, and their corresponding measurements (shaded nodes in Fig. 4(b)) from the training samples, our goal is to estimate the parameter \( \theta \) with structure \( G \) in Fig. 4(b) by the MAP approach as shown in Eq. 3.

\[
\theta^* = \arg \max_{\theta} P(\theta | D, G) = \arg \max_{\theta} \log P(\theta | D, G) P(\theta)
\]

where \( P(\theta | D, G) \) represents the joint likelihood, and \( P(\theta) \) represents the prior distribution of the node parameters. In our context model, the prior distribution for node \( E_n \) is assumed to be related to the event frequency as discussed in the following part of this section. For the rest nodes, we set uniform priors to the parameters.

As a directed Probabilistic Graphical Model, the parameter learning of the context model can be factorized into estimating the parameters of the distribution for each node given its parents (derived from the \( \log P(D | \theta, G) P(\theta) \) term in Eq. 3). More specifically, the parameter learning of node \( O_{En} \) in the context model is fulfilled by the baseline DBN learning in Section 5.2 since the probability \( P(O_{En} | E_n) \) is captured by the likelihood \( P(\mathbf{F} | \Theta_c) \) of DBN models. Also, without loss of generality, we denote one of the discrete nodes in the context model to be \( X_i \). For the parameter learning of this discrete node \( X_i \), the posterior distribution of parameter \( \theta_k \) for \( X_i \) given its parent \( s \) \( pai(X_i) \) in state \( j \) can be represented by Dirichlet distribution:

\[
P(\theta|D,G) = \text{Dir}(1 + x_{i1} + N_{i1}, \ldots, 1 + x_{ija} + N_{ija})
\]

where \( r_i \) is the number of states for variable \( X_i \). We use \( X_i = k \) to denote that node \( X_i \) is in state \( k \), and use \( pai(X_i) \) = \( j \) to denote the parent \( s \) of node \( X_i \) to be in state \( j \). And, \( N_{ik} \) reflects the number of cases in the training set \( D \) for which \( X_i = k \) and \( pai(X_i) = j \), and \( x_{ik} \) is the hyper-parameter that reflects the prior beliefs about how often the case \( X_i = k \) and \( pai(X_i) = j \) would appear.

With the MAP approach in Eq. 3, the analytical solution for parameter \( \theta_{ik} \) of discrete node \( X_i \) which stands for the probability \( P(X_i = k | pai(X_i) = j) \) would be:

\[
\theta_{ik} = \frac{x_{ik} + N_{ik}}{x_{ij} + N_{ij}}
\]

where \( N_{ij} = \sum_{k=1}^{r_i} N_{ik} \) and \( x_{ij} = \sum_{k=1}^{r_i} x_{ik} \). Thus, the MAP learning for node \( X_i \) will degenerate to the maximum likelihood (MLE) learning approach when using a uniform prior by setting \( z_{ik} = 0 \).

To add conjugate prior to node \( E_n \) in the context model, we define the hyper-parameters \( \xi_{En,k} \) as \( \xi_{En,k} \), which means its hyper-parameters are not related to the states of its parents \( (S \) and \( E_{n-1} \)). Moreover, we set \( \xi_{En,k} \) to be proportional to the event frequency. Hence, once the training samples for learning the event transitions \( P(E_n | E_{n-1}, S) \) in different scenes are limited, the parameters learned by MAP would be close to the event frequency rather than to the parameters learned by MLE.

6.4. Event recognition with context model

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

\[
P(E_n | O_{En}, DO_n, DS, DE_{n-1}) \propto P(O_{En} | E_n) \cdot \sum_{En} \{P(O_{En} | E_n)P(DOn | O_{En}) \} \cdot \sum_{DS} \{P(DS | E_n)P(On | DO_n) \} \cdot \sum_{DE_{n-1}} \{P(DE_{n-1} | E_n, En-1) \}
\]

where \( P(O_{En} | E_n) \) is taken from the DBN inferred likelihood \( P(\mathbf{F} | \Theta_c) \), and the \( DO_n, DS \) and \( DE_{n-1} \) are the discrete context measurements.

In practice, such inference can be solved with variable elimination (Koller and Friedman, 2009). The classification finds the class label \( c^* \) that maximizes the posterior probability as

\[
c^* = \arg \max_{c} P(E_n = c | O_{En}, DO_n, DS, DE_{n-1})
\]

Moreover, we observe that, when the measurements for certain contexts are missing, the inference using the model in Fig. 4(a) would naturally degrade to the model combining the baseline DBN with only existing context measurement \( s \). This can be proved by marginalizing the joint probability distribution over the missing context measurement \( s \).

7. Experiments

We use two surveillance datasets to verify our proposed methods. The first dataset is the VIRAT aerial dataset (ApHill) (Oh et al., 2011) with approximately 4 h of videos. 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). In this dataset, we have 167 sequences of computed tracks (Lucas et al., 1981) that belong to one of the above 8 events. These 167 sequences contain 319.9 s of videos which we can use for event recognition. For each

Fig. 4. PGM model combining the scene, event-object interaction, and event temporal contexts. (a) context model; (b) context model learning with hyper parameters in a MAP process.
event, the sequence numbers are: 20 sequences for LAV, 23 sequences for UAV, 11 sequences for OAT, 15 sequences for CAT, 15 sequences for GIV, 17 sequences for GOV, 28 sequences for EAV, and 38 sequences for XAF. These 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 (Oh et al., 2011) with approximately 8 h of videos. 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. For this set, a total number of 188 sequences belong to each of the six events, resulting in 577.9 s of videos that can be used for event recognition. For each event, the sequence numbers are: 11 sequences for LAV, 16 sequences for UAV, 18 sequences for OAT, 19 sequences for CAT, 61 sequences for GIV, and 63 sequences for GOV. These two datasets are very challenging compared to other event/ activity/ action datasets like KTH (Schuldt et al., 2004), Weizmann (Gorelick et al., 2007), and HOHA (Laptev et al., 2008). As shown in Fig. 5, the two datasets are collected in real scenes with low resolution. 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.

7.1. 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 (Chellappa et al., 2004) that we use receives 62.43% recall rate and 33.87% false alarm rate on the query video sequences of this dataset. The event recognition performance on each event is given as Area Under ROC Curve (AUC) shown in Table 1.

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. Besides the ROC curves, we also evaluate the overall recognition rate on these eight events. The overall recognition rate improves from 32.34% to 37.13% with the help of vehicle context.

7.2. DBN with synthetic object interaction context

The experiment in Section 7.1 shows that even with a vehicle detector (Chellappa et al., 2004) that just receives 62.43% recall rate and 33.87% false alarm rate, our context model can already utilize these vehicle detection results as object interaction context to benefit the event recognition task. But, we are also curious to know in which cases the context information would help the recognition, and in which cases the context information may hurt the recognition instead. To make a thorough study on these issues, we further test our context augmented DBN model using synthetic vehicle detections from a series of synthetic vehicle detectors with different accuracies. If the accuracy of one synthetic vehicle detector is c, we generate the corresponding synthetic vehicle detections to make its recall rate and false alarm rate to be c and 1 – c respectively. This experiment is also performed on VIRAT aerial dataset. We test with 7 synthetic vehicle detectors whose accuracies are 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0 respectively. The overall event recognition rates of both the context augmented DBN and the baseline DBN are shown in Fig. 6.

In Fig. 6, the baseline DBN receives a 32.34% recognition rate which is the same as the performance in Section 7.1. However, the performances of context augmented DBN are different with different synthetic vehicle detectors. When using a synthetic vehicle detector whose accuracy is 0.4 (even below the random chance for “0/1” sequence based vehicle detection), the vehicle context actually hurts the event recognition. On the other hand, with a synthetic vehicle detector that provides 100% accuracy, the overall event recognition rate can be improved to 45.51%. In addition, with a 0.6 accuracy synthetic vehicle detector which performs slightly better than random chance, the proposed model can improve the performance of baseline DBN to 35.33%. Thus, we can empirically say that with the context information that is generally constructive (e.g. with accuracy better than random chance), our context model could utilize this information to help event recognition.

Table 1

<table>
<thead>
<tr>
<th>Events</th>
<th>Baseline DBN</th>
<th>DBN with vehicle context</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAV</td>
<td>0.5510</td>
<td>0.5686</td>
</tr>
<tr>
<td>UAV</td>
<td>0.7419</td>
<td>0.7059</td>
</tr>
<tr>
<td>OAT</td>
<td>0.4763</td>
<td>0.6003</td>
</tr>
<tr>
<td>CAT</td>
<td>0.6225</td>
<td>0.7190</td>
</tr>
<tr>
<td>GIV</td>
<td>0.6300</td>
<td>0.7386</td>
</tr>
<tr>
<td>GOV</td>
<td>0.6844</td>
<td>0.6939</td>
</tr>
<tr>
<td>EAV</td>
<td>0.7685</td>
<td>0.7561</td>
</tr>
<tr>
<td>XAF</td>
<td>0.7231</td>
<td>0.7389</td>
</tr>
<tr>
<td>Average</td>
<td>0.6499</td>
<td>0.6900</td>
</tr>
</tbody>
</table>

Fig. 5. Examples of the eight events we experiment for surveillance video recognition. The exemplar images of (a) through (f) are from the VIRAT public 1.0 dataset, and the images of (g) and (h) are from the VIRAT aerial dataset.

Please cite this article in press as: Wang, X., Ji, Q. Context augmented Dynamic Bayesian Networks for event recognition. Pattern Recognition Lett. (2013), http://dx.doi.org/10.1016/j.patrec.2013.07.015
7.3. DBN with scene and event temporal context

We further explore the usage of the scene context and event temporal context for improving event recognition in this section. 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 context and the event temporal context. We apply the scene context, the event temporal context and their combinations. For scene classification, our global scene classifier which is a support vector machine (SVM) trained and tested on color histograms (Xiao et al., 2010) feature reaches 95.74% recognition rate on recognizing these three scenes. In this experiment, the Area Under ROC Curve (AUC) performance comparisons for event recognition are shown in Table 2.

In Table 2, the global scene context can slightly improve the DBN overall performance by over 1%. The event 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, from Table 2, we can observe that combining both scene and event temporal contexts does not always give the best AUC performances for each event. More specifically, for event Unloading a Vehicle (UAV) and event Closing a Trunk (CAT), combining both contexts does not yield performance better than incorporating scene context or event temporal context alone. This is explainable.

<table>
<thead>
<tr>
<th>Events</th>
<th>DBN</th>
<th>DBN + Scene</th>
<th>DBN + Temporal</th>
<th>DBN + Temporal + Scene</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAV</td>
<td>0.5501</td>
<td>0.5819</td>
<td>0.3876</td>
<td>0.6507</td>
</tr>
<tr>
<td>UAV</td>
<td>0.6192</td>
<td>0.6715</td>
<td>0.4757</td>
<td>0.4790</td>
</tr>
<tr>
<td>CAT</td>
<td>0.5911</td>
<td>0.6559</td>
<td>0.7446</td>
<td>0.7409</td>
</tr>
<tr>
<td>GOV</td>
<td>0.5416</td>
<td>0.5420</td>
<td>0.5207</td>
<td>0.5674</td>
</tr>
<tr>
<td>Average</td>
<td>0.5757</td>
<td>0.5936</td>
<td>0.6131</td>
<td>0.6374</td>
</tr>
</tbody>
</table>

7.4. DBN with three contexts combined

In this section, we show the results of utilizing the event-object interaction context, the scene context, and the event 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 data, the task becomes even more challenging with more variations. We first show the performances of DBN and context models trained with different numbers of training sequences in Section 7.4.1, and then present the performances of DBN and context models trained with all available training samples in Section 7.4.2. In addition, the statistics about the computation time of our models are also included in Section 7.4.2.

7.4.1. Performances with different training sequence number

The training of DBN models usually requires a large amount of data. However, in our application, the training sequence number for each event is relatively limited. In this section, we try to experimentally analyze how the number of training sequences would affect the performance of DBN models, and how the context model would boost DBN performance when training sequence number is limited. We train the DBN and context models with respectively 5, 10, 15, 20, and 22 sequences for each event, and evaluate the trained models on the same testing set. The overall event recognition rates with respect to different training sequence numbers are given in Fig. 7, where the blue squares linked with blue dashed line represent the performances of DBN and context models when training sequence is limited to 5, 10, 15, 20, 22, respectively. The DBN and context models trained with different numbers of training sequences are shown in Table 2.
lines are the baseline DBN performances, and the red circles linked
with red solid lines are the context augmented DBN performances.

From the blue curve of Fig. 7, we can see with fewer training se-
quences, the performance of baseline DBN model decreases. How-
ever, the context model can effectively boost the baseline DBN
performance even when training sequence number is limited. For
example, provided with only 5 sequences per event, the context
augmented DBN achieved an overall performance better than that
of the baseline DBN trained with 15 sequences per event. More-
over, as the training sequence number increases, the performance
of context augmented DBN generally increases faster than the per-
formance of baseline DBN model.

7.4.2. Performances with three contexts combined

We further utilize the full combined data for testing our DBN
augmented with event-object interaction context, the scene con-
text, and the event temporal context simultaneously. In Fig. 8(a),
we present the overall recognition rates with five fold cross valida-
tions. An overall 9% improvement can be realized utilizing all con-
texts. We also present the confusion matrices of baseline DBN and
DBN with three contexts in Fig. 8 (b) and (c) respectively.

From the confusion matrices shown in Fig. 8 (b) and (c), we can
see the context information incorporated by our proposed approach
greatly improves the recognition performance of baseline DBN.
The performance of baseline DBN is affected by factors like
tremendous intra-class variations and low resolution, and tends
to mis-classify similar events like LAV and UAV etc. On the other
hand, the context augmented DBN approach can recognize each
event in a more balanced manor, and alleviate the mis-classifica-
tions between similar events (e.g. LAV and UAV, OAT and CAT, etc.).

We further provide the statistics about the computation time of
our model in this experiment, where we use a computer with 2.93
GHz Intel Core i7 CPU and 16 GB memory. Since our system would
train both DBN and context models offline, and recognize the
events during testing online, we emphasize most on minimizing
the computational costs for testing. Different from the traditional
dynamic junction tree method (Ghahramani, 1997) for DBN evalu-
ation, we use forward propagation (Rabiner, 1989) which is not
only exact but also much faster. For the 897.8 s of videos belonging
to the eight events in the combined dataset, the baseline DBN mod-
el without context information takes 11.44 s for testing, and the
DBN with context information takes 11.96 s for testing. These
numbers translate to about 0.4 ms per frame, much faster than real
time.

7.5. Comparison with other models using context

As discussed in Section 2 and 6, context is defined as the sur-
rroundings, environment, background and settings which help clarify
the meaning of an event. It is generally regarded as the extra
information that is available or can be retrieved both during train-
ning and testing. Thus, many approaches use context information
directly as extra feature describing the surrounding environment,
and directly use classifiers like support vector machine (SVM) to
incorporate these contextual features. To compare with such tradi-
tional approaches also using context, we choose SVM as the base-
line both without and with context.

We compare our models with SVM not using and using the
three contexts (the event-object interaction context, the scene con-
text, and the event temporal context) on the combined dataset in
the same experiment settings as the settings discussed in Sec-
Table 3, it is clear the three contexts used as features can
improve the SVM performance by over 1%. And, even though our
generative baseline DBN model performs not as well as the dis-
criminative SVM classifier, our proposed context augmented DBN
can improve the baseline DBN by over 9%. In all, our proposed ap-
proach is around 3% better than SVM that also uses the same con-
text information.

Please cite this article in press as: Wang, X., Ji, Q. Context augmented Dynamic Bayesian Networks for event recognition. Pattern Recognition Lett. (2013),
http://dx.doi.org/10.1016/j.patrec.2013.07.015
8. 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 event DBN model with three types of contexts: the object, scene and event temporal contexts. Experiments on real scene surveillance datasets show that the contexts can effectively improve the event recognition performance even under great challenges like large intra-class variations and low image resolution. Comparison with the existing context model also demonstrate the superior performance of our model.

Our proposed context approach can be further applied to other surveillance vision tasks like object detection, tracking, track association etc. In these tasks, context information from multiple sources would also help improve performance on these tasks. Moreover, event recognition could also serve as a mutual context for these tasks, since event category itself can put strong constraints on corresponding objects.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version, at http://dx.doi.org/10.1016/j.patrec.2013.07.015.

References


Table 3

Comparison of overall recognition rates of SVM without and with contexts, the baseline DBN model, and the proposed model incorporating contexts to DBN.

<table>
<thead>
<tr>
<th>Models</th>
<th>Performance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM without context</td>
<td>25.35</td>
</tr>
<tr>
<td>SVM with context</td>
<td>26.48</td>
</tr>
<tr>
<td>Baseline DBN</td>
<td>20.28</td>
</tr>
<tr>
<td>DBN + Context</td>
<td>25.30</td>
</tr>
</tbody>
</table>

Please cite this article in press as: Wang, X., Ji, Q. Context augmented Dynamic Bayesian Networks for event recognition. Pattern Recognition Lett. (2013), http://dx.doi.org/10.1016/j.patrec.2013.07.015