A Hierarchical Context Model for Event Recognition in Surveillance Video

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

Abstract

Due to great challenges such as tremendous intra-class variations and low image resolution, context information has been playing a more and more important role for accurate and robust event recognition in surveillance videos. The context information can generally be divided into the feature level context, the semantic level context, and the prior level context. These three levels of context provide crucial bottom-up, middle level, and top down information that can benefit the recognition task itself. Unlike existing researches that generally integrate the context information at one of the three levels, we propose a hierarchical context model that simultaneously exploits contexts at all three levels and systematically incorporate them into event recognition. To tackle the learning and inference challenges brought in by the model hierarchy, we develop complete learning and inference algorithms for the proposed hierarchical context model based on variational Bayes method. Experiments on VIRAT 1.0 and 2.0 Ground Datasets demonstrate the effectiveness of the proposed hierarchical context model for improving the event recognition performance even under great challenges like large intra-class variations and low image resolution.

1. Introduction

Visual event recognition is attracting growing interest from both academia and industry [24]. Visual event recognition is defined as the recognition of spatio-temporal visual patterns from videos. Much existing related work [22, 34, 39] has been focusing on recognition of basic human action/activities (like "walking", "turning around" etc.) in clean backgrounds using datasets such as KTH [28], Weizmann [9], and HOHA [19]. By contrast, in this research, we focus on recognition of real world surveillance video events that involve interactions between humans and objects with complex backgrounds. Various algorithms have been developed for event recognition. These methods can be divided into feature (descriptor) based approach and model based approach. Feature-based approach includes methods that employ various image features and treat event recognition as a classification problem. The most widely used features include the HOG feature [5] and the spatio-temporal features such as the spatio-temporal interest point (STIP) features [18] and optical flow based features [4]. These features capture local appearance or motion patterns near the interest points or optical flows. Although having been successfully applied to many applications, these features generally focus more on local patterns.

(a) (b) (c) (d)

Figure 1. Examples of event "loading" with intra-class variation.

Model-based methods for event and activity recognition include probabilistic graphical models such as Hidden Markov Models [21], Dynamic Bayesian Networks [36], Conditional Random Fields [33], and their variants. They use model to encode sematic and temporal relationships and combine it with image features. While capable of simultaneously capturing both spatial and temporal interactions, they can only capture the local spatial and temporal interactions due to the underlying Markov assumption.

Despite these efforts, surveillance video event recognition still faces difficulties even with the well-constructed descriptors or 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 target motion variation, viewpoint change, illumination change, and occlusion. Figure1 gives examples of event "loading" with large appearance variations. Second, the poor target tracking results and the often low video resolution further aggravate the problem. These challenges force us to rethink the existing data-driven and target-centered event recognition approach and to look for extra information to help mitigate the challenges. The contextual information serves this purpose well.

Contextual information refers to additional information about the target objects and its context. While not directly

describing an event, it provides information on the circumstance and environment within which the event occurs and can therefore support event recognition. Recently, context has also been increasingly used in solving computer vision problems. Contexts in general can be grouped into feature level context [2, 37], semantic relationship level context [10, 20, 38], and prior/priming information level context [32, 7]. These three levels of contexts have also been investigated for event recognition. For example, at feature level, Wang *et al*. [34] present a multi-scale spatio-temporal context feature that captures the spatio-temporal interest points in event neighborhoods. At semantic relationship level, Yao *et al*. [38] propose a context model to make human pose estimation and activity recognition as mutual context to help each other. At prior/priming information level, the scene priming information [32] has been proved to be effective for event recognition in [25, 35].

Existing work on contexts generally incorporates one type of context or context features at one level. There is not much work that simultaneously exploits different types of contexts at different levels. Since context exists at different levels and comes in different types, we believe event recognition can benefit greatly if we can simultaneously exploit contexts at different levels and systematically incorporate them into event recognition. To this goal, we introduce an unified hierarchical model that allows systematically capturing contexts at different levels and principally integrates the captured contexts with the image measurements for robust event recognition from surveillance videos. Specifically, based on a dynamic graphical model, the proposed model can capture contexts at feature level, semantic relationship level, and prior/priming information level. Through this unified model, context in the bottom (feature) level would provide diagnostic support for the event, while context on top (prior) level provides predictive knowledge on the event. The top-down and bottom-up context meet at the middle (semantic relationship) level, where the three levels of contexts are systematically integrated to yield a comprehensive characterization of events and their context.

In summary, in this paper, we introduce a hierarchical context model that allows systematically integrating contexts from different levels for accurate and robust event recognition from surveillance videos. Compared to the existing context models for event recognition, the proposed model is comprehensive, systematic, and can better handle the challenges associated with real world videos.

2. Related Work

Incorporating context into visual recognition is an active area of research in computer vision. Given the ill-posed nature with many computer vision tasks and the poor image quality, context is being increasingly employed in various computer vision tasks. A comprehensive review on context based object recognition is given in [8]. Also, the empirical study in [6] catalogues 10 possible sources of contexts that could be beneficial. Recently, there are also increasing efforts in applying context to event recognition. In general, contextual information can exist at different levels including *feature level* [2, 37, 15, 34], *semantic relationship level* [10, 38, 17], and *prior/priming information level* [32, 30, 7, 29]. Below we briefly summarize work in each category as well as the latest efforts in integrating contexts from different levels.

At *feature level*, the context provides information about the event and its surroundings at pixel level. Many context features have been introduced for activity/event recognition. Yao *et al*. [37] propose the "grouplet" which is a descriptor that captures the structured information of an image by encoding a number of discriminative visual features and their spatial configurations. Kovashka *et al*. [15] propose to learn the shapes of space-time feature neighborhoods that are most discriminative for a given category. Wang *et al*. [34] present a representation that captures the contextual interactions between interest points in both local and neighborhood spatio-temporal domains. At *semantic relationship level*, context captures relationships among basic elements of events such as the co-occurrence semantic relationships between actions, objects, scene and poses. More specifically, Gupta *et al*. [10] present a Bayesian approach for combining action understanding with object perception; Yao *et al*. [38] propose a Markov random field model to encode the mutual context of objects and humans poses in human-object interaction activities. At *prior/priming information level*, the context captures the global spatial or temporal environment, within which events may happen. The scene priming used by Torralba *et al*. [32] and Sudderth *et al*. [30] demonstrate that scene provides a good prior information for object recognition and object detection. The scene priming information [32] has also been proved to be effective for event recognition in [25, 35].

There are several approaches that also utilize hierarchical modeling to integrate contexts. Sudderth *et al*. [30] propose to model the priming hierarchy between scene, objects and parts with a Bayesian topic model. He *et al*. and Kumar *et al*. [11, 16] utilize contexts in multi-scale image level hierarchy for image labeling tasks. Li *et al*. [20] try to capture the semantic co-occurrence relationships between event, scene and objects also with a Bayesian topic model. Sun *et al*. [31] propose to combine the point-level context feature, the intra-trajectory context feature and the inter-trajectory context feature through a multiple kernel learning model. Choi *et al*. [3] use tree hierarchy based graphical model to capture the object co-occurrence and spatial relationships. Recently, Jiang *et al*. [13, 12] utilize Dirichlet process mixture model to capture semantic co-occurrence relationships between human poses and objects. The latest

approach by Zhu *et al*. [40] also exploit contexts for event recognition. While similar to our approach in spirit, our approach differs from [40] in the following aspects: 1) we propose a probabilistic hierarchical model to model and capture contexts. By contrast, their model is a structural linear model. 2) We integrate contexts from all three levels, while their model only integrates contexts at feature and semantic relationship level.

In addition, approaches like [26, 23] also utilize hierarchical probabilistic models for event and action recognition. However, these two approaches focus on capturing the hierarchy on feature, body parts and human actions, without incorporating context information beyond the target.

In summary, the existing work in context-aided event recognition focuses mostly on context at an individual level. The existing work in integration of contexts at different levels is limited to two levels. By contrast, we propose a unified model that allows integrating contexts from all three levels simultaneously. Experiments demonstrate significant performance improvement over the existing models on challenging real world benchmark surveillance videos.

3. Hierarchical Context Model Formulation

We introduce an unified hierarchical model that allows systematically capturing contexts at different levels, and principally integrate the captured contexts with the video measurements for robust event recognition from surveillance videos. Figure 2 illustrates the overall idea of our approach. We propose to model three levels of contexts: feature context, semantic context, and priming context. The feature context in the bottom level provides diagnostic support information for the event, while the priming context at the top level supplies top-down predictive knowledge on the event. The top-down and bottom-up contexts meet at the middle level (semantic relationship context), where the three levels of contexts are systematically integrated to have a comprehensive characterization of the events and their overall contexts.

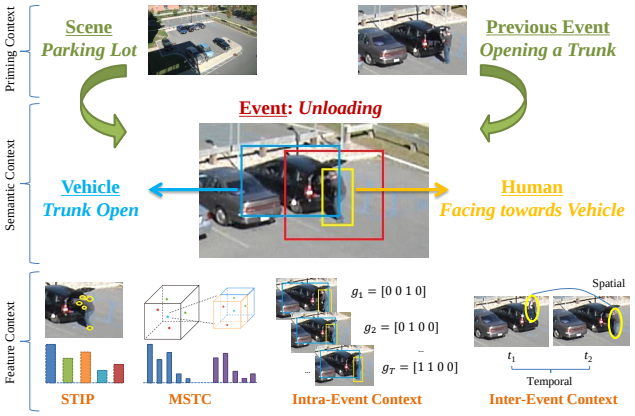


Figure 2. An example of incorporating contexts at different levels.

3.1. The Hierarchical Context Model

Specifically, we incorporate the three levels of contexts shown in Figure 2 systematically through a probabilistic hierarchical context model as shown in Figure 3. The top portion captures the prior/priming context, the middle part captures the semantic context, and the bottom part captures the feature level context. Each part consists of nodes representing respectively events, the related contexts, and their image measurements. Below we will elaborate the context modeling at each level.

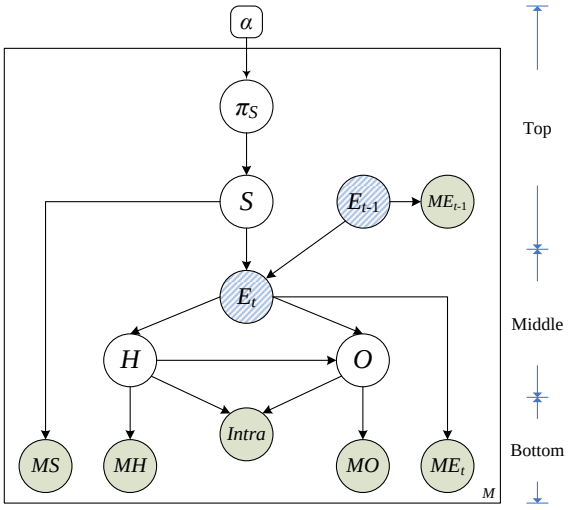


Figure 3. The hierarchical context model, where the shaded nodes represent the image measurements, the white nodes represent the latent variables, and the stripped nodes represent events to infer.

3.1.1 Semantic Context Modeling in the Middle Level

Since the middle level semantic context modeling connects the top level prior modeling with the bottom-level feature context modeling, we describe the semantic context modeling first, and will discuss the feature and prior context next.

The semantic context captures components of an event and their interactions. For this research, we are interested in events that involve interactions between humans and objects (e.g. vehicles), the semantic context modeling should therefore capture human, object, and their interactions. As shown in Figure 3, we denote an event by a discrete node E_t, and its components by discrete nodes H and O to represent human and object respectively. An event E_t can be characterized by the state of the human H, the state of the object O, and by their interactions. For example, event "person getting into vehicle" is highly correlated with human state "facing towards vehicle" and object state "door open"; also, event "person opening a trunk" has strong relations with human state "at tail of vehicle" and object state "trunk open". The interactions between H and O are captured by

the link between them, which specifies the probabilistic dependencies among their states. Furthermore, to capture the semantic contexts for different events, we connect node E_t to H and O respectively, with E_t as their joint parent node. This allows conditioning H and O on E_t such that their semantic context may vary, depending on the specific event type. Finally, as we normally do not know the exact human and object states that can best characterize different events, we treat both H and O as latent variables such that their optimal states can be learned automatically during training.

3.1.2 Feature Context Modeling at Bottom Level

Feature context modeling in the bottom part of Figure 3 consists of image features that describe the event and its context. First, to directly measure event, we use the raw event feature (i.e. spatio-temporal interest point (STIP) [18]) that provides direct observation of target event. In addition, we introduce image features to directly measure human and objects as denoted by the nodes MH and MO respectively. To measure the event context at the feature level, we utilize several types of context features. The first type of context feature is the multi-scale spatio-temporal context (MSTC) [34] that captures the spatio-temporal interest points in event neighborhoods. The second type of context feature is the inter-event and intra-event context features [40], which capture the inter-event spatio-temporal relations and intra-event human and object interactions respectively.

We combine the STIP feature, the MSTC context feature, and the inter-event context feature into the event measurement and denote it by the node ME. The intra-event context feature is denoted by the node $Intra$. It provides necessary bottom-up support for nodes H and V. The link between nodes E_t and ME_t models the distribution of ME_t given certain E_t states. Similarly, the links between H , O, and their measurements MH, MO, and $Intra$ capture dependencies of these measurements on the states of H and O. Through this captured distribution, the information from image and context features in the bottom level propagate up to provide a diagnostic support for the inference of the current event state E_t.

3.1.3 Priming Context Modeling on Top Level

The final part of the hierarchical model is the prior context at the top. It captures the related high level context that determines the likelihood of the occurrence of certain events. For this research, we utilize two types of contextual priming [32] information: the scene context and the dynamic context, though the model is generic enough to apply to other high level priming contexts.

The scene context provides an environmental context within which events occur. The scene context can dictate the likelihood for certain events to occur as well as event happening location (e.g. parking lot, shop entrance) and time (e.g. noon, dark). Scene context can therefore serve as event prior. In the model, scene is captured by a discrete scene node S, which represents different possible scene states. The link from S to E_t captures the cause and effect relation between S and E_t. To capture the prior probability of the scene node, we introduce the parameters π_S, which specify the prior probability distribution of S. To accommodate the variability with the prior probability distribution of S, i.e., the variability of π_S, we treat π_S as a random variable and introduce the hyper-parameters α to characterize its distribution. Specifically, we assume node S follows multinomial distribution with parameter π_S, and π_S follows Dirichlet distribution with the hyper-parameter α. To be able to generalize to different scenes, we assume S is a latent variable, and we will learn its hyper-parameters α during training.

The second contextual priming is the dynamic context. It provides temporal support as to what event will likely to happen given the events that have happened up to now. Event at current time is influenced by events in previous times. For example, event "loading/unloading a vehicle" typically *precedes* event "closing a trunk". The information on the previous events provides a good cue on the current event. Dynamic context can therefore serve as a temporal prior on current event. Dynamic context is captured by the E_{t-1} node. The link between E_{t-1} and E_t captures the temporal causal relation between E_t and E_{t-1}.

Both the nodes S and E_{t-1} provide top-down priming information for the inference of the current event.

3.2. Model Learning

Using the directed graphical model shown in Figure 3, we factorize the joint probability distribution of all nodes for each sample as the product of local conditional probabilities, i.e.,

$$\begin{aligned}&P(E_t, E_{t-1}, H, O, S, \pi_S, MH, MO, MS, ME_t, Intra,\\ &ME_{t-1}|\alpha) = P(\pi_S|\alpha)P(S|\pi_S)P(E_{t-1})P(E_t|S, E_{t-1})\cdot\\ &P(H|E_t)P(O|E_t, H)P(MS|S)P(ME_t|E_t)P(ME_{t-1}|E_{t-1})\cdot\\ &P(MH|H)P(MO|O)P(Intra|H, O)\end{aligned} \quad (1)$$

where each factor is a local conditional probability for each node. Specifically, the term $P(\pi_S|\alpha)$ follows Dirichlet distribution, $P(S|\pi_S)$ follows multi-nominal distribution. $P(E_{t-1})$, $P(E_t|S, E_{t-1})$, $P(H|E_t)$, $P(O|E_t, H)$, and $P(MS|S)$ can be characterized by conditional probability tables (CPTs). $P(ME_t|E_t)$, $P(ME_{t-1}|E_{t-1})$, $P(MH|H)$, $P(MO|O)$, and $P(Intra|H, O)$ follow Gaussian distributions. We propose to learn the parameters of these local distributions as follows.

We use α to denote the hyper-parameter of Dirichlet distribution π_S, use the parameters ϵ, δ, η to denote

the CPTs of $P(E_t|S, E_{t-1})$, $P(MS|S)$ and $P(E_{t-1})$ respectively. The remaining model parameters for CPTs of $P(H|E_t)$ and $P(O|E_t, H)$, as well as for the Gaussian distributions of $P(ME_t|E_t)$, $P(ME_{t-1}|E_{t-1})$, $P(MH|H)$, $P(MO|O)$, and $P(Intra|H, O)$ are denoted altogether by θ, since these parameters can be learned altogether using standard EM method in Eqn. 3. Learning of the distributions in our proposed model amounts to the estimation of the parameters α, ϵ, δ, η and θ. Given M training sequences with measurements $\{MH^m, MO^m, MS^m, ME_t^m, Intra^m, ME_{t-1}^m\}_{m=1}^M$ and event labels $\{E_t^m, E_{t-1}^m\}_{m=1}^M$, we maximize the joint log-likelihood as

$$\begin{aligned}
&\max_{\alpha,\epsilon,\delta,\eta,\theta} \sum_{m=1}^{M} \log P(E_t^m, E_{t-1}^m, MH^m, MO^m, MS^m, \\
&\qquad ME_t^m, Intra^m, ME_{t-1}^m|\alpha, \epsilon, \delta, \eta, \theta) \\
&= \max_{\alpha,\epsilon,\delta,\eta} \sum_{m=1}^{M} \log P(E_t^m, E_{t-1}^m, MS^m|\alpha, \epsilon, \delta, \eta) \qquad (2) \\
&\quad + \max_{\theta} \sum_{m=1}^{M} \log P(MH^m, MO^m, ME_t^m, Intra^m, \\
&\qquad ME_{t-1}^m|E_t^m, E_{t-1}^m; \theta) \qquad (3)
\end{aligned}$$

The objective in Eqn. 3 involves marginalization of two discrete latent nodes H and V. Its optimization can be solved using standard EM algorithm. The objective in Eqn. 2 involves marginalization of two coupled latent nodes S and π_S where S is discrete and π_S is continuous. The optimization of Eqn. 2 requires the variational Bayes based EM technique due to the complexity brought in by the continuous latent node π_S. In the following, we provide more details regarding the optimization of objective in Eqn. 2.

For training sample m, we can obtain the variational lower bound of $\log P(E_t^m, E_{t-1}^m, MS^m|\alpha, \epsilon, \delta, \eta)$ through Jensen's inequality as

$$\begin{aligned}
&\log P(E_t^m, E_{t-1}^m, MS^m|\alpha, \epsilon, \delta, \eta) \geq \\
&\int \sum_S q(\pi_S^m, S^m) \log \frac{P(\pi_S^m, S^m, E_t^m, E_{t-1}^m, MS^m)}{q(\pi_S^m, S^m)} \mathrm{d}\pi_S^m = \\
&E_q\left[\log P(\pi_S^m, S^m, E_t^m, E_{t-1}^m, MS^m)\right] - E_q\left[\log q(\pi_S^m, S^m)\right] \\
&\triangleq L(\gamma^m, \phi^m; \pi_S^m, S^m) \qquad (4)
\end{aligned}$$

where $q(\pi_S^m, S^m)$ is the variational distribution that can be factorized as $q(\pi_S^m, S^m|\gamma^m, \phi^m) = q(\pi_S^m|\gamma^m)q(S^m|\phi^m)$, with γ^m and ϕ^m as the variational parameters both in dimension K. The maximization of the variational lower bound $L(\gamma^m, \phi^m; \pi_S^m, S^m)$ with respect to γ^m and ϕ^m is equivalent to the minimization of the Kullback-Leibler divergence between $q(\pi_S^m, S^m|\gamma^m, \phi^m)$ and $P(\pi_S^m, S^m|E_t^m, E_{t-1}^m, MS^m; \alpha, \epsilon, \delta, \eta)$. Hence, by maximizing this lower bound, we can get optimized approximation of $\log P(E_t^m, E_{t-1}^m, MS^m|\alpha, \epsilon, \delta, \eta)$.

Our variational Bayes based learning algorithm would then follow the variational EM method:

E-Step: Maximize lower bound $L(\gamma^m, \phi^m; \pi_S^m, S^m)$ with respect to variation parameters γ^m and ϕ^m for each sample $m = 1, ..., M$ as

$$\max_{\gamma^m,\phi^m} L(\gamma^m, \phi^m; \pi_S^m, S^m)$$

This maximization can by solved by iteratively updating the variation parameters ϕ^m and γ^m through

$$\phi_i^m \propto \epsilon_{iu}\delta_{iv}\eta_w \exp\left\{\psi(\gamma_i^m) - \psi(\sum_{j=1}^{K}\gamma_j^m)\right\} \qquad (5)$$

$$\gamma_i^m = \alpha_i + \phi_i^m \qquad (6)$$

where $i \in \{1, ..., K\}$. ϵ_{iu}, δ_{iv} and η_w reflect the probabilities $P(E_t|S, E_{t-1})$, $P(MS|S)$ and $P(E_{t-1})$ with $S = i$ and nodes E_t, MS and E_{t-1} in states u, v and w. Also, $\psi(\cdot)$ is the digamma function.

M-Step: Maximize the joint lower bound of all samples with respect to parameters $\alpha, \epsilon, \delta, \eta$ as

$$\max_{\alpha,\epsilon,\delta,\eta} \sum_{m=1}^{M} L(\gamma^m, \phi^m; \pi_S^m, S^m)$$

which can be solved through a gradient descent method.

3.3. Model Inference

Given an unknown event sequence E_t with observed measurements (i.e. MH, MO, MS, ME_t, $Intra$, and ME_{t-1}), we recognize its event category e^* by maximizing its posterior probability given all measurements as

$$e^* = \arg\max_{E_t} P(E_t|MH, MO, MS, ME_t, Intra, ME_{t-1}; \alpha)$$

This posterior probability is proportional to the joint probability $P(E_t, MH, MO, MS, ME_t, Intra, ME_{t-1}|\alpha)$, which can be further decomposed as

$$\begin{aligned}
&P(E_t, MH, MO, MS, ME_t, Intra, ME_{t-1}|\alpha) = P(ME_t| \\
&E_t) \sum_{H,O} P(H, O|E_t)P(Intra|H, O)P(MH|H)P(MO|O)\cdot \\
&\sum_{E_{t-1}} P(ME_{t-1}|E_{t-1}) \int \sum_S P(\pi_S, S, E_t, E_{t-1}, MS|\alpha)\mathrm{d}\pi_S
\end{aligned}$$

The calculation of most terms above are straightforward. However, the integration term involves the marginalization of two coupled latent nodes π_S and S to obtain probability $P(E_t, E_{t-1}, MS|\alpha)$. Its exact calculation is intractable. On the other hand, we have defined the variational lower bound of $\log P(E_t, E_{t-1}, MS|\alpha)$ as in Eqn 4. By resorting to the variational inference through maximizing the lower bound, we can get the approximate estimation of probability $P(E_t, E_{t-1}, MS|\alpha)$. For each testing sample, the maximization of the lower bound still follows the iterative method given in Eqn. 5 and Eqn. 6.

4. Experiments

We demonstrate the effectiveness of our hierarchical context model with two surveillance datasets: the VIRAT 1.0 Ground Dataset and VIRAT 2.0 Ground Dataset [24]. Both datasets focus on real world surveillance video events that involve interactions between humans and objects with complex backgrounds.

The VIRAT 1.0 Ground Dataset contains approximately 3 hours of surveillance videos from realistic scenes of different school parking lots. There are in total six types of events: *Loading a Vehicle* (LAV), *Unloading a Vehicle* (UAV), *Opening a Trunk* (OAT), *Closing a Trunk* (CAT), *Getting into a Vehicle* (GIV), and *Getting out of a Vehicle* (GOV). All these six types of events are person-vehicle interaction events.

The VIRAT 2.0 Ground Dataset extends the VIRAT 1.0 Ground Dataset to over 8 hours of videos containing realistic surveillance scenes of school parking lots, as well as shop entrance, outdoor dining area and construction sites. Besides the six types of person-vehicle interaction events defined in VIRAT 1.0 Ground Dataset, the VIRAT 2.0 Ground Dataset also includes five more events as: *Gesturing* (GST), *Carrying a Object* (CAO), *Running* (RUN), *Entering a Facility* (EAF), and *Exiting a Facility* (XAF).

For both datasets, we use half of the data for training, and the rest for testing.

4.1. Baselines

To assess the effectiveness of the proposed model for incorporating contexts, we use the STIP based BOW feature on target with SVM classifier as our first baseline **STIP+SVM**. This baseline uses only the information from the recognition target without incorporating any contexts. Besides the STIP+SVM baseline, we compare with another baseline called **Context+SVM** that concatenate STIP based BOW feature on target with all the context features including MSTC, intra-event and inter-event context features. The Context+SVM baseline would then provide an evaluation on how incorporating context information as feature would perform for event recognition.

To further compare our model performance with state of the art performances, we list the performances of several most related approaches: the feature based approach proposed by Reddy *et al*. [27] that utilizes histogram of spatiotemporal gradients feature extracted from event bounding boxes for event recognition, the approach combining BOW feature with SVM classifier [14] which is very popular for action and activity recognition, the sum-product network approach by Amer *et al*. [1] utilizing space-time arrangements of primitive actions, and the structural model proposed by Zhu *et al*. [40] that integrates different contexts as input of the structural model.

4.2. Performance on VIRAT 1.0 Ground Dataset

We first present the model performance on VIRAT 1.0 Ground Dataset. Figure 4 shows the per-event recognition accuracy, and the average recognition accuracy over all events for the two baselines (STIP+SVM and Context+SVM) and our proposed hierarchical context model. The Context+SVM baseline performs better than the STIP+SVM baseline in terms of average recognition accuracy over six events. This result indicates the context information is beneficial for event recognition. More importantly, our proposed hierarchical context model performs better than the two baselines for five of the six events, and improves the average recognition accuracy from 39.91% (STIP+SVM) and 52.96% (Context+SVM) to 65.78%, which is a 25% absolute improvement over the STIP+SVM baseline, and a 12% absolute improvement over the Context+SVM baseline. This comparison demonstrates that our hierarchical context model is very effective on simultaneously incorporating the feature level, the semantic level and the prior level contexts. Utilizing all these levels of context information through our model can significantly improve the event recognition performance.

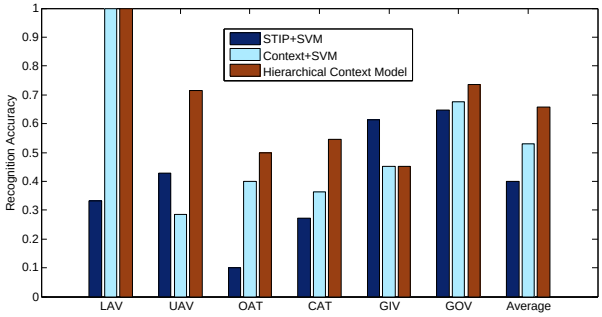
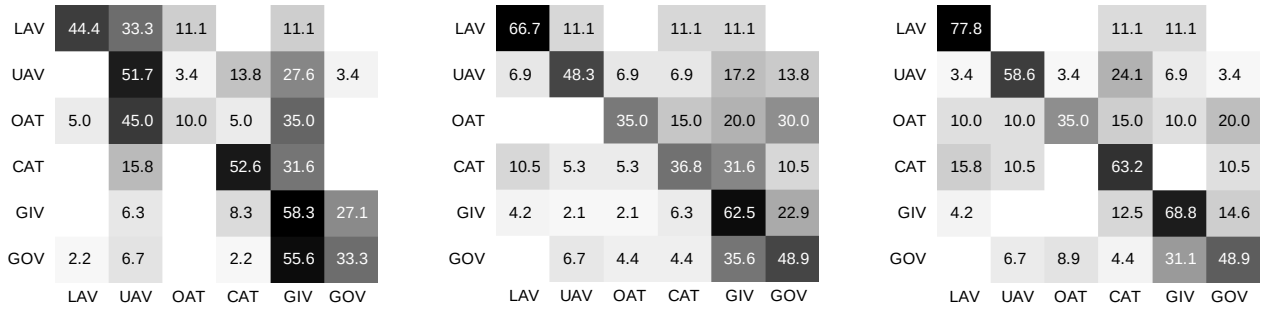


Figure 4. Recognition accuracy comparison on VIRAT 1.0 Ground Dataset. For all six events, the average recognition accuracy is 39.91% for STIP+SVM, 52.96% for Context+SVM, and 65.78% for the proposed hierarchical context model.

In Table 1, we compare with related approaches for the recognition of six events on VIRAT 1.0 Ground Dataset. For each approach, the recognition accuracy for each event and the average recognition accuracy of all six events are listed. From this comparison, we can see that the context based approaches including [40] and our approach generally outperform the approaches in [27] and [14] that do not use context information. Also, our proposed model outperforms [40] on the average recognition accuracy of six events on VIRAT 1.0 Ground Dataset.

4.3. Performance on VIRAT 2.0 Ground Dataset

The VIRAT 2.0 Ground Dataset contains 11 types of events, among which 6 events (LAV, UAV, OAT, CAT, GIV, GOV) involve the interactions between person and vehicle.



(a) STIP+SVM (b) Context+SVM (c) Our Model

Figure 5. Confusion matrices for the recognition of six person-vehicle interaction events on VIRAT 2.0 Ground Dataset with the STIP+SVM, Context+SVM and our proposed hierarchical context model.

Table 1. Comparison with state of the art approaches for event recognition on VIRAT 1.0 Ground Dataset. In this comparison, our proposed model performs the best in terms of average recognition accuracy over all six events.

Accuracy %	Reddy *et al.* [27]*	BOW+ SVM [14]	Zhu *et al.* [40]	Our Model
LAV	10.0	42.8	52.1	**100**
UAV	16.3	57.2	57.5	**71.4**
OAT	20.0	39.3	69.1	50.0
CAT	34.4	33.4	72.8	54.5
GIV	38.1	48.2	61.3	45.2
GOV	61.3	53.8	64.6	**73.5**
Average	35.6	45.8	62.9	**65.8**

*: This accuracy is read out from the bar graph in [27] with the method providing best average accuracy.

We first test our approach on recognizing these 6 events in Section 4.3.1. In Section 4.3.2, we test on recognizing 11 events, and compare the performance with state of the art.

4.3.1 Six Events Involving Person-Vehicle Interaction

Figure 6 compares the performances of STIP+SVM, Context+SVM and our proposed model on the per-event recognition accuracy and the average recognition accuracy for the recognition of six events involving person-vehicle interaction. For this comparison, our proposed hierarchical context model can consistently outperform or compete with the two baseline approaches for each event.

To evaluate the effectiveness of contexts at each level, we experiment with an additional **Bottom+Middle** model which is the hierarchical context model without top level. Our evaluation shows that by adding the feature-level context, the performance improves by 8% (from 41.74% by STIP+SVM to 49.70% by Context+SVM as in Figure 6) over using only target features. By adding the semantic level to the feature level, the Bottom+Middle model further improves the result by about 5% to 54.24%. Finally, by adding the top level, the performance improves by another 4% to a final accuracy of 58.70%. These results show the importance of contexts in each level. And, joint modeling contexts at all three levels reaches the best performance.

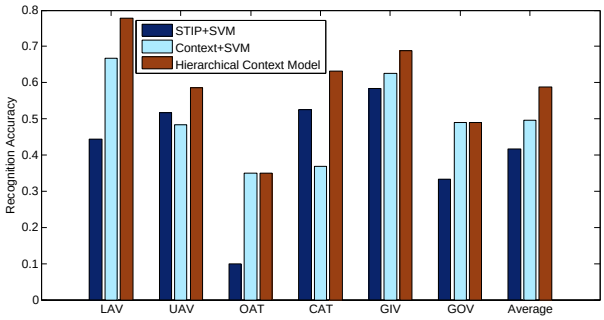


Figure 6. Recognition accuracy comparison for six events involving person vehicle interaction on VIRAT 2.0 Ground Dataset. For these six events, the average recognition accuracy is 41.74% for STIP+SVM, 49.70% for Context+SVM, and 58.70% for the proposed hierarchical context model.

For this experiment, the confusion matrices for the STIP+SVM, Context+SVM and our proposed hierarchical context model are further provided in Figure 5. In Figure 5(a), we can see the STIP+SVM approach still faces difficulties on distinguishing pairs of event that are similar in appearance. For example, in Figure 5(a), some mismatches occur between the two events "getting into a vehicle" (GIV) and "getting out of a vehicle" (GOV), or between the two events "loading a vehicle" (LAV) and "unloading a vehicle" (UAV). Comparatively, the Context+SVM approach alleviate such mismatch between similar events due to the usage of contextual features such as the intra-event and inter-event context feature that can provide additional clues besides the event appearance. Moreover, our proposed hierarchical context model can obviously reduce the mismatch between similar pairs of events with the incorporation of prior level, semantic level and feature level contexts simultaneously.

4.3.2 Performance with All Events

We further experiment on VIRAT 2.0 Ground Dataset with all 11 events, and compare with the performances of the state of the art approaches by Amer *et al*. [1] and Zhu *et al*. [40] respectively. The overall performance comparison is given in Table 2, where the recall and precision are used to evaluate the recognition performance. We can see our proposed model can outperform both approaches on the recognition of 11 events on VIRAT 2.0 Ground Dataset.

Table 2. Comparisons with state of the art methods for recognition of all 11 events on VIRAT 2.0 Ground Dataset.

	Amer *et al*. [1]	Zhu *et al*. [40]	Our Model
Precision	72%	71.8%	74.73%
Recall	70%	73.5%	77.42%

5. Conclusion

In this paper, we propose a hierarchical context model to systematically integrate feature level context, semantic level context, and prior level context for accurate and robust event recognition in surveillance videos. Compared to existing approaches that generally incorporate contexts from one level, one major contribution of this work is to build up a comprehensive model that can integrate contexts from all three levels simultaneously. We develop complete model learning and inference algorithms to tackle the challenges brought in by the model hierarchy. In experiments, we evaluate our model performance on both VIRAT 1.0 and 2.0 Ground Datasets for recognizing the real world surveillance video events that involve interaction between humans and objects with complex backgrounds. The results with the proposed hierarchical context model show significant improvements over the baseline approaches that also utilize context. Comparisons with state of the art methods also demonstrate the superior performance of our model.

Acknowledgments: This work is supported in part by Defense Advanced Research Projects Agency under grants HR0011-08-C-0135-S8 and HR0011-10-C-0112, and by the Army Research Office under grant W911NF-13-1-0395.

References

[1] M. R. Amer and S. Todorovic. Sum-product networks for modeling activities with stochastic structure. In *CVPR*, pages 1314–1321. IEEE, 2012. 6, 8

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI, IEEE Transactions on*, 24(4):509–522, 2002. 2

[3] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010. 2

[4] R. Cutler and M. Turk. View-based interpretation of real-time optical flow for gesture recognition. In *Automatic Face and Gesture Recognition*, 1998. 1

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1

[6] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, pages 1271 –1278, june 2009. 2

[7] A. C. Gallagher and T. Chen. Estimating age, gender, and identity using first name priors. In *CVPR*, pages 1–8. IEEE, 2008. 2

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

[9] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *PAMI, IEEE Transactions on*, 29(12):2247 –2253, dec. 2007. 1

[10] A. Gupta and L. Davis. Objects in action: An approach for combining action understanding and object perception. In *CVPR*, pages 1 –8, june 2007. 2

[11] X. He, R. S. Zemel, and M. A. Carreira-Perpinán. Multiscale conditional random fields for image labeling. In *CVPR*, volume 2, pages II–695, 2004. 2

[12] Y. Jiang, H. Koppula, and A. Saxena. Hallucinated humans as the hidden context for labeling 3d scenes. In *CVPR*, 2013. 2

[13] Y. Jiang, M. Lim, and A. Saxena. Learning object arrangements in 3d scenes using human context. *ICML*, 2012. 2

[14] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Inter. Conf. on Image and Video retrieval*, 2007. 6, 7

[15] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *CVPR*, 2010. 2

[16] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *ICCV*, volume 2, pages 1284–1291, 2005. 2

[17] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. *PAMI*, 2012. 2

[18] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005. 1, 4

[19] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, pages 1 –8, june 2008. 1

[20] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *ICCV*, pages 1 –8, oct. 2007. 2

[21] F. Lv and R. Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *ECCV*, pages 359–372. Springer, 2006. 1

[22] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, pages 2929 –2936, june 2009. 1

[23] J. C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *CVPR*, pages 1–8, 2007. 3

[24] S. Oh and et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR*, pages 3153 –3160, june 2011. 1, 6

[25] S. Oh and A. Hoogs. Unsupervised learning of activities in video using scene context. In *ICPR*, pages 3579 –3582, 2010. 2

[26] S. Park and J. K. Aggarwal. A hierarchical bayesian network for event recognition of human actions and interactions. *Multimedia systems*, 2004. 3

[27] K. K. Reddy, N. Cuntoor, A. Perera, and A. Hoogs. Human action recognition in large-scale datasets using histogram of spatiotemporal gradients. In *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pages 106–111. IEEE, 2012. 6, 7

[28] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR*, volume 3, pages 32 – 36 Vol.3, aug. 2004. 1

[29] X. Shen and Y. Wu. A unified approach to salient object detection via low rank matrix recovery. In *CVPR*, pages 853–860, 2012. 2

[30] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, 2005. 2

[31] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li. Hierarchical spatio-temporal context modeling for action recognition. In *CVPR*, 2009. 2

[32] A. Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 53(2):169–191, 2003. 2, 4

[33] D. L. Vail, M. M. Veloso, and J. D. Lafferty. Conditional random fields for activity recognition. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, page 235. ACM, 2007. 1

[34] J. Wang, Z. Chen, and Y. Wu. Action recognition with multiscale spatio-temporal contexts. In *CVPR*, pages 3185 –3192, june 2011. 1, 2, 4

[35] X. Wang and Q. Ji. Incorporating contextual knowledge to dynamic bayesian networks for event recognition. In *ICPR*, 2012. 2

[36] G. Yang, Y. Lin, and P. Bhattacharya. A driver fatigue recognition model based on information fusion and dynamic bayesian network. *Information Sciences*, 180(10):1942–1954, 2010. 1

[37] B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010. 2

[38] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, pages 17 –24, june 2010. 2

[39] F. Zhu, L. Shao, and M. Lin. Multi-view action recognition using local similarity random forest and sensor fusion. *Pattern Recognition Letters*, 2013. 1

[40] Y. Zhu, N. Nayak, and K. Roy-Chowdhury. Context-aware modeling and recognition of activities in video. In *CVPR*, 2013. 3, 4, 6, 7, 8