Exploiting Spatio-Temporal Scene Structure for Wide-Area Activity Analysis in Unconstrained Environments

Nandita M. Nayak, Yingying Zhu, and Amit K. Roy-Chowdhury

***Abstract*—Surveillance videos typically consist of long duration sequences of activities which occur at different spatio-temporal locations and can involve multiple people acting simultaneously. Often, the activities have contextual relationships with one another. Although context has been studied in the past for the purpose of activity recognition to a certain extent, the use of context in recognition of activities in such challenging environments is relatively unexplored. In this paper, we propose a novel method for capturing the spatio-temporal context between activities in a Markov random field. The structure of the MRF is improvised upon during test time and not pre-defined, unlike many approaches that model the contextual relationships between activities. Given a collection of videos and a set of weak classifiers for individual activities, the spatio-temporal relationships between activities are represented as probabilistic edge weights in the MRF. This model provides a generic representation for an activity sequence that can extend to any number of objects and interactions in a video. We show that the recognition of activities in a video can be posed as an inference problem on the graph. We conduct experiments on the publicly available UCLA office dataset VIRAT dataset to demonstrate the improvement in recognition accuracy using our proposed model as opposed to recognition using state-of-the-art features on individual activity regions.**

***Index Terms*—Context-aware activity recognition, Markov random field, wide-area activity analysis.**

I. Introduction

Activity recognition is a challenging task in realistic environments. A long-term wide-area surveillance video usually consists of multiple people entering and exiting the scene over a period of time. Therefore, it is hard to predict the number of activities occurring in the scene and the number of people involved in those activities. This variability in the number of actors and the number of action executions within the sequence is what we term as an "unconstrained" environment in this paper. Most existing activity recognition algorithms focus on the region where an activity occurs while ignoring the contextual information in the surroundings. Such methods place assumptions on the number of objects, scale and viewpoint of the scene and may not be equally effective in more challenging environments. Often, it has been found that examining the surroundings of an activity under consideration in a scene can provide useful clues about the activity. This information obtained from the surroundings is termed as "context" of the activity.

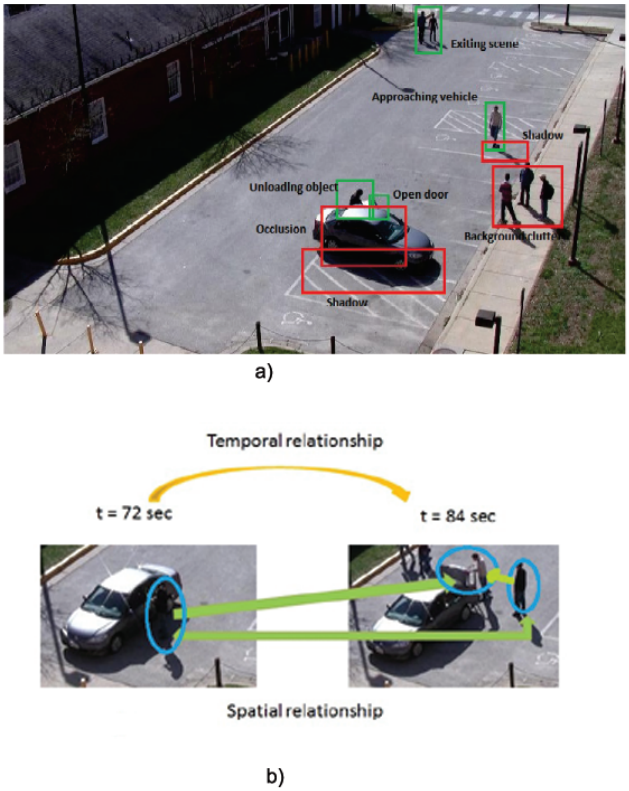

Fig. 1. a) An example scene from a surveillance video in a parking lot demonstrating that different different activities happen together and can influence each other. "Open door", "unloading vehicle" and "approaching vehicle" are related since they pertain to the same vehicle. Other objects in the scene can cause background clutter making the recognition task challenging. b) the spatio-temporal relationship between two activities in a video.

In this paper, we propose to use activities themselves as providing contextual information to other activities in their vicinity. The modeling of this contextual information along with a traditional activity recognition system can provide improved recognition rates in challenging environments.

The use of context is actively being explored in computer vision today. The use of any data in the video which does not directly correspond to the object or activity being analyzed can be termed as context. Consider a wide area surveillance

N. Nayak (email: nandita.nayak@email.ucr.edu) is with the Computer Science department at University of California, Riverside, CA, 92521 USA.

Y. Zhu (email: yzhu010@ucr.edu) and A. K. Roy-Chowdhury (email: amitrc@ee.ucr.edu) are with the Electrical Engineering department at University of California, Riverside, CA, 92521, USA.

This work has been partially supported by NSF grant IIS-0712253, and DARPA STTR award W31P4Q-11-C0042 through Mayachitra, Inc.

scene consisting of multiple actors performing a series of activities. Unlike sports videos which are governed by a fixed set of rules, these videos are unconstrained and contain a variable number of objects and activities. By unconstrained, we mean that the activities might be related but do not unfold according to set rules. In such long duration sequences, we can expect that several activities would influence each other causally while some others might occur independently. However, inferring these causalities is not trivial due to the presence of multiple actors. Also, tracking in such sequences can be challenging due to the presence of clutter and occlusion. In this work, we propose to model the spatio-temporal context between individual activities in a long duration sequence using a Markov random field. Since the number of actors can vary from one sequence to another, we propose to construct the graphical model which is specific to a test sequence.

Most existing activity recognition approaches aim at recognizing atomic activities or a single interaction in a short video clip. Real world videos tend to have a large amount of intra-class variation as well as clutter and noise which makes the recognition task difficult. Therefore, although the standard recognition methods can be applied here, it is difficult to obtain a high accuracy results with existing classifiers. A typical example of an outdoor wide area scene is shown in Figure 1 a). The different challenges in recognition are marked on the figure. Figure 1 b) shows the spatial and temporal relationships between two activities in a video. The presence of multiple activities, however, also imply that we now have more information available to us about the scene as a whole, as compared to a small clip containing a single atomic activity. Activities in a video are often related to each other. For example, the fact that a person opens the trunk of a car makes it very likely that he might place or retrieve an object from the trunk. In addition, if we knew that the person had just exited a facility, it is more likely that he will place an object rather than retrieve it. Therefore, the occurrence of one activity can provide us a context which can be used to recognize another related activity. In this work, we wish to demonstrate a method to model this context and utilize the information to recognize activities in a complex video.

The key idea behind our approach is that, if two activities are related, they can be expected to occur within a small spatio-temporal vicinity. The spatial separation, temporal separation and the association frequency of these activities can therefore be modeled as context for recognition of these individual activities. Given a collection of videos and a set of baseline classifiers for atomic activities, we wish to learn the spatio-temporal relationships between atomic activities and model them. The relationships are learnt from the training data. We propose to have a Markov Random Field model over the test sequence, with the edge potentials modeling the spatio-temporal relationships between them. The baseline classifiers (which are assumed to provide a weak classification) give us the node potentials. An inference on this MRF will help us estimate the activities in the sequence.

A. Contributions

The main contributions of this work are the following.

1) We propose a generalized formulation for modeling the contextual relationships between activities in the presence of multiple actors and when they are acting simultaneously in the scene. They could be interacting with each other or acting independently.
2) We take a probabilistic approach to modeling relationships between activities. We model the spatial relationships, temporal relationships as well as the association frequencies into the potential functions of a random field model. Inference on this graph gives us the estimate of the categories to which each activity corresponds. We perform experiments on realistic videos containing multiple activities spread over space and time with high amount of clutter and noise.
3) We define the structure of the graphical model on the test sequence rather than having a pre-defined structure. This gives us the flexibility to model different number of individuals and different number of activities based on the test sequence. We demonstrate that studying the pair-wise relationships between activities during training is sufficient for this purpose. We also propose a way of iteratively modifying the graph structure to arrive at one which is more likely to capture important relationships, thereby increasing the activity recognition confidence.

B. Overview

An overview of our proposed model for activity recognition in activity sequences is shown in Figure 2. Given a long-term video, the goal of our approach is to estimate the category to which individual activity belongs. We assume that we have some training videos available, each of which have one or more sequences of activities occurring in different spatio-temporal regions. Each spatio-temporal region where a potential activity takes place is termed as an activity region. We also have available a set of baseline classifiers $C = \{c_1, c_2..c_N\}$, which can output a probability of an activity region y belonging to a particular class c_i, i.e. $P(c_i|y)$.

A typical surveillance video, such as a parking lot video (shown in Figure 1) contains several activities occurring simultaneously or in succession in different portions of the scene. The number of objects, people and activities change from one video sequence to another. Having identified the activity regions in a video using the baseline classifiers, and having clustered them into sequences which are potentially related to each other, we explore three key aspects to improve the accuracy of recognition: 1) The relationship in the spatial locations of activities, 2) the relationship in the temporal locations of activities and 3) the probability of association of two given activities, i.e., the probability that one activity might occur in the vicinity of another.

These concepts are modeled by a Markov random field (MRF). Since the MRF is used to model the context information across activities, we choose the nodes of the MRF as atomic activities rather than pixels or image regions, as is commonly done in image segmentation. Each edge represents the spatio-temporal context between the activity nodes that it connects. The node potentials are obtained using the likelihood

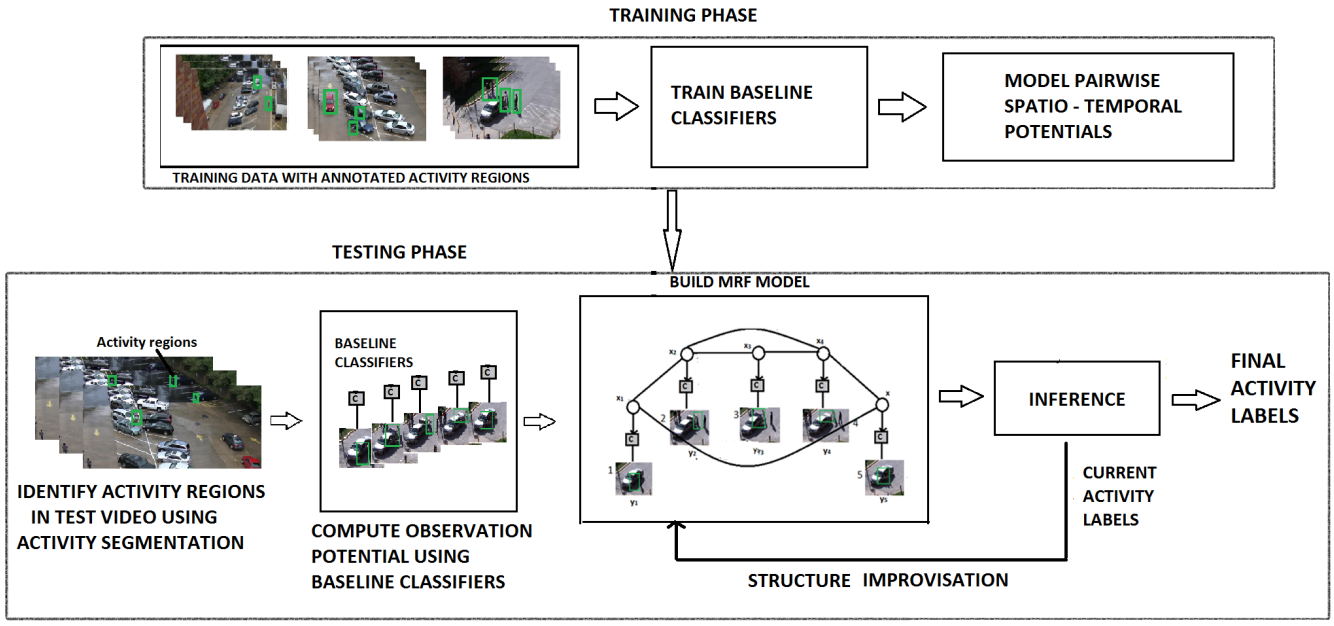

Fig. 2. Figure shows the illustration of our proposed method. Training involves modeling the pairwise spatio-temporal relationships between different activity regions which are provided in annotations as mentioned in Section II-C2. For a test video, activity regions are identified using the method presented in Section III-B. Using the potentials from training data and observation potentials as described in Section II-C1, the node labels are inferred (Section II-D). We also propose a method to improve the structure of the graph to capture the best dependencies (Section II-E. The MRF model is enlarged in Figure 3.

of the activities given by the baseline classifiers. Since the number of activities differs from one sequence to another, the structure of the MRF is tuned to the test sequence in an iterative manner based on the spatio-temporal locations of activities. The edge potentials are learnt from the training data. We perform inference on the resulting MRF to estimate the activities in the test sequence. We conduct experiments on the UCLA office dataset containing indoor office sequences and the publicly available VIRAT dataset containing parking lot videos.

C. Related Work

A major thrust of research in complex activity recognition has been in the selection of features and their representations, most of which have dealt with single activity clips [1]. Different representations have been used in activity recognition, such as space time interest points (STIP) [2] and histogram of optical flow [3]. In this work, we deal with long duration videos and explore the contextual information between different activities in the video to arrive at a representation of the scene.

Graphical models are commonly used to encode relationships in video analysis. A grid based belief propagation method was used for pose estimation in [4]. Stochastic and context free grammars have been used to model complex activities in [5]. Co-occurring activities and their dependencies have been studied using Dependent Dirichlet Process - Hidden Markov Models (DDP-HMMs) in [6]. In our work, we propose a Markov random field framework which can handle varying number of actors and activities.

Spatio-temporal relationships have played an important role in the recognition of complex activities. Methods such as [7]and [8] explore spatio-temporal relationships at a feature level. The spatial and temporal relationships between space-time interest points have been encoded as "feature graphs" in [9]. Although such methods have been applied to multiple activities occurring simultaneously, it may not be practical to construct such graphs over long term video sequences and do not explore the relationships across activities. Complex activities were represented as spatio-temporal graphs representing multi-scale video segments and their hierarchical relationships in [10]. Most of these papers focus on the modeling of low level features for recognition. Variable length Hidden Markov models are used to identify activities with high amount of intra class variabilities in [11]. In this paper, we have modeled the spatio-temporal relationships between different activities which form a higher level representation.

Context has been widely used in the past in the task of object recognition, and more recently in activity recognition. Spatial relationships between objects have been modeled using graphs for new object category discovery in [12]. The authors in [13] use objects as context for activities and vice versa. Similarly, association of tracks with activities and the generation of a high level storyline model using AND-OR graphs has been performed in an EM framework in [14]. Contextual information between different actors in group activities has been studied in [15]. Spatial context between objects and activities has been modeled in [16]. Spatio-temporal context in structured videos with manually defined rules has been modeled using Markov Logic Networks in [17]. We propose a generalized formulation for context modeling that is suited for unconstrained video sequences such as outdoor surveillance videos. The key aspects of such sequences is that there is no constraint on the number of actors in the scene or the number of activities in the sequence. Also, different actors in the scene may act independently or interact with each other if

they choose to do so.

Some of the previous approaches such as [18] assume a known structure of the graph for context representation. Models such as the AND-OR graphs or other tree structures have been suggested in the past [19] [14] for modeling sports sequences and office environments. These models however, are more suited for structured environments where there are a set of rules governing the behavior of people such as in sports, or where the number of objects/activities or the combinations of sub-activities are limited as in an office environment. Applying such models to unconstrained sequences can be laborious due to the exponential number of combinations of activities which have to be learnt here to construct such models. Similarly, papers such as [20] use context to infer a collective activity using single person activities. In such sequences however, it is assumed that all participating persons/objects contribute to the collective activity. Whereas, in a typical surveillance scenarios, different actors may or may not be interacting with each other, therefore such models cannot be directly applied here. The authors in [21] deal with recognizing a single activity over multiple cameras by topology inference, person re-identification and global activity interpretation. Here, we are dealing with a set of different activities which may or may not be correlated, therefore a Markov random field is a more suited model to capture these complex spatio-temporal relationships. Social roles for hierarchical representation of activities in sports videos is explored in [22]. Most of this work deals with short duration videos or with videos with a pre-defined structure such as sports videos. We propose to define the structure of the graph on the test sequence rather than use a pre-defined structure.

The next section explains the construction of the MRF in detail.

II. Graphical Representation of Activities

To begin with, we will define the commonly used terminologies in the paper for a clear understanding of our proposed method.

A. Definitions

- **Activity** - A meaningful event in the video which we wish to identify. Our objective is to assign every activity a class label in the range $c_1..c_N$.
- **Activity region** - A spatio-temporal volume in which the activity takes place. An activity region A_i is represented by its spatial and temporal centroids s_i and t_i.
- **Activity Sequence** - A set of activities which occur in close proximity with each other and can have causal influences on each other. Each activity sequence is modeled as a Markov random field and evaluated.

B. Proposed Model

The goal of our algorithm is to model the space-time relationships between the activities in a scene using a Markov random field. The MRF is an undirected graph $G = (V, E)$, with a set of nodes V and a set of edges E. Given a video sequence to be recognized, we first construct an MRF over all probable related activities in the sequence. Each node denotes an activity and an edge represents the spatio-temporal relationship between two activities. There are a set of observations $Y = \{y_1..y_n\}$ and a set of hidden variables $X = \{x_1..x_n\}$ for a sequence of n activities. An observation node y_i denotes the image observation of an activity, which are the features computed over an activity region. The output of the baseline classifiers for each activity is used to compute an observation potential. A hidden node denotes an atomic activity to be estimated. A node x_i can be defined as

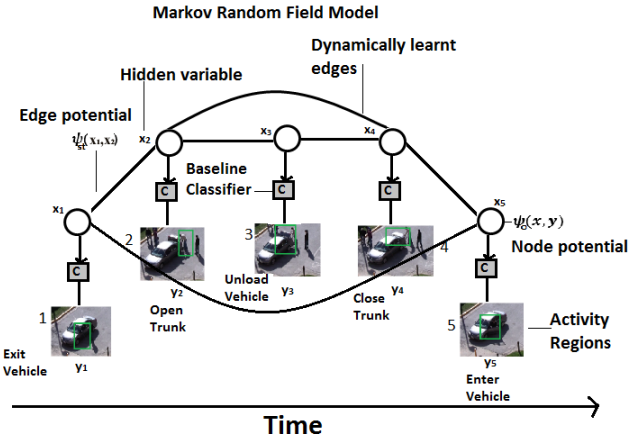

Fig. 3. Figure shows the Markov random field constructed over a spatio-temporal volume for an activity sequence. Shown in the figure are the activity regions which form the observation variables y. The baseline classifier output forms the observation potential. The labels of the activities which have to be predicted constitute the hidden nodes x. The edges of the graph are learnt iteratively.

$$x_i = (c_i, s_i, t_i), \tag{1}$$

where x_i denotes a node, c_i denotes the activity class to which it belongs, s_i is its spatial location and t_i denotes its temporal location. The MRF is given by

$$\Psi = \frac{1}{Z} \prod_{i,j \in E} \psi_{st}(x_i, x_j) \prod_{i \in V} \psi_o(x_i, y_i), \tag{2}$$

where Ψ is the overall potential. Here, we assume that the MRF factors over the edges. There are two kinds of potentials associated with the graph. $\psi_{st}(x_i, x_j)$ is the edge potential which is the spatio-temporal relation between two hidden nodes connected by an edge and $\psi_o(x_i, y_i)$ is the observation potential of a node. Z is the normalization constant. The illustration of our proposed graphical model is shown in Figure 3.

C. Potential Functions

The node observation potentials and the spatio-temporal edge potentials are defined as given below.

1) Observation Potential: The observation potential or the node potential is the evidence of the activity obtained from the video data. These are obtained from the image observations of the activities which are the baseline classifiers. We have

one baseline classifier per activity class, the output of which is the probability of the given activity belonging to a particular category. We use a Bag-Of-Features approach over space-time interest points [23] as our baseline classifiers due to its popularity for recognition of atomic activities. Specifically, space-time interest points based on Harris and Forstner operaters are computed over the training set. A feature vector is generated for each point. During training, a codebook is build by clustering and quantizing these features. Each category of activity is modeled as a distribution over this vocabulary. The interest points are computed over the test video and regions with significant number of points from the vocabulary are said to be the activity regions, denoted as the observation variables y_i. A discriminative classifier such as a multiclass SVM classifier is used to compute the probability of an activity region belonging to a particular category $P(c_j|y_i)$. These probabilities are learnt jointly over the training data. The observation potential is therefore defined as

$$\psi_o(x_i, y_i) = p(x_i|y_i, C), \tag{3}$$

where ψ_o is the observation potential, y_i is the observation variable and C is the set of baseline classifiers. It is to be noted that any other set of features or algorithm can also be used for the baseline classifiers.

2) Spatio-temporal Potential: The spatio-temporal potential is defined on edges connecting the activity variables in the graph. Actions which are within a spatio-temporal distance of each other are assumed to be related to each other. There are three components to this potential: the spatial component, the temporal component and the association component. The spatial component models the probability of an activity belonging to a particular category given its spatial configuration with its neighbor. Similarly, the temporal component models the probability of an activity belonging to a particular category given its temporal distance with its neighbor. The association component is the probability of two activities being within a pre-defined spatio-temporal vicinity of each other. The spatial and temporal components are modeled as normal distributions whose parameters μ_s, σ_s, μ_t and σ_t are computed using the training data. The spatial component is given by

$$\psi_s(x_i, x_j) = \mathcal{N}_{sd}(\|s_i - s_j\|^2; \mu_s(c_i, c_j), \sigma_s(c_i, c_j)), \tag{4}$$

$$\psi_t(x_i, x_j) = \mathcal{N}_{td}(\|t_i - t_j\|^2; \mu_t(c_i, c_j), \sigma_t(c_i, c_j)). \tag{5}$$

where $\mu_s(c_i, c_j)$,$\sigma_s(c_i, c_j)$, $\mu_t(c_i, c_j)$ and $\sigma_t(c_i, c_j)$ are the parameters of the distribution of relative spatial and temporal positions of the activities, given their categories. The association probability f_{ij} is computed as a ratio of the number of times an activity category c_j has occurred in the vicinity of activity category c_i to the total number of times the category c_i has occurred. Therefore, the spatio-temporal potential is given by

$$\psi_{st}(x_i, x_j) = f_{ij}\psi_s(x_i, x_j)\psi_t(x_i, x_j) \tag{6}$$

In a general case, these potentials would be learnt jointly over a pre-defined graph. However, in our case, although we deal with long sequences of activities, we have found that the activity pairs which are the closest spatio-temporal neighbors provide sufficient contextual information to be used as a context prior. For example, a person opening the trunk makes it very likely that he will also close the trunk. Therefore, the pairwise potentials are learnt independently for each pair of activities. We examine all activities within a close spatio-temporal range of each other and model their pairwise spatio-temporal relationships. This will account for any occlusions/misses and false positives of activities. It also gives us the flexibility to determine the structure of the graph and the number of connections based on the test sequence. The number of different functions to be evaluated is therefore, $N(N-1)$ for a set of N activity categories. The parameters for any two categories c_i and c_j can be learnt by maximizing

$$D = \sum_k \log \psi_{st}(x_i^k, x_j^k), \tag{7}$$

where x_i^k, x_j^k are the k^{th} training example of categories i and j.

D. Inference

Inference in a graphical model involves computing the marginal probabilities of the hidden or unknown variables given an evidence or an observed set of variables. We choose the belief propagation method for estimation of parameters. Since there are loops in our model, the loopy belief propagation is used. Although this algorithm is not guaranteed to converge, it has shown excellent empirical performance [24].

At each iteration, a node sends messages to its neighbor. All nodes are updated based on the messages from their neighbors. Consider a node $x_i \in V$ with a neighborhood $N(x_i)$. The message sent by a node $x_i \in V$ to its neighbor $x_j \in V, (x_i, x_j) \in E$ can be given as

$$m_{x_i,x_j}(x_j) = \alpha \int_{x_i} \psi_{st}(x_i, x_j)\psi_o(x_i, y_i) \prod_{x_k \in N(x_i)} m_{x_k,x_i}(x_i) dx_i \tag{8}$$

The marginal distribution of each activity region is given by

$$p(x_i) = \alpha\psi_o(x_i, y_i) \prod_{x_j \in N(x_i)} m_{x_j,x_i}(x_i) \tag{9}$$

The activity label which has the highest marginal distribution is assigned to the region.

E. Structure Improvisation

As mentioned before, since it is hard to predict the number of activities in the scene in advance, rather than having a pre-defined graphical structure or a pre-defined set of relationships, we propose to construct the graphical model on the test sequence. Starting with an initial model, we suggest an iterative approach to improve this structure in a way that might capture the important relationships between nodes, thereby improving the recognition scores. Entropy of a node x_i is defined as $E(x_i) = -\sum_{j=1}^{N} P(c_j|x_i)\log_2(P(c_j|x_i))$, where $P(c_j|x_i)$ denotes the posterior probability of x_i belonging to category c_j and N is the number of activity categories. We propose to use the entropy of a node as a measure of the confidence

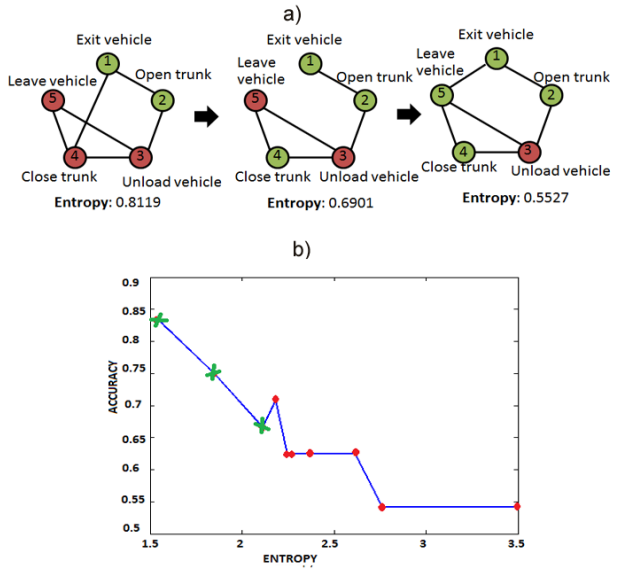

Fig. 4. a) Figure shows three iterations of structure learning for a set of five activities. The ground truth label of each node is marked in the figure. The color of the nodes indicates whether the predicted label matches the true label (green) or not (red). The change in structure is visible in the edges. We see that, by choosing the edges which tend to decrease the entropy, we are more likely to increase the recognition accuracy.b) A plot of accuracy vs entropy for a video containing 24 activities. It can be seen that as entropy decreases, recognition accuracy tends to increase. The graphs from a) are obtained in the points marked in green.

of the system in labeling the node. This is based on the intuition derived from [12] that when the system is confident of classifying a node x_i as belonging to a class c_i, we can expect $P(c_i|x_j)$ to have a high value while $P(c_k|x_i)\forall k \neq i$ to be low, thereby lowering the entropy of the node. Therefore, we wish to improve the structure of the graph in such a way that the change lowers the overall entropy of the system. Although there is no guarantee that this will always increase the accuracy of the system, in most cases the confidence of the classifier is reflective of the performance of the system.

The first step in the analysis of activities using an MRF is to construct the graph. Ideally, we would want an edge to connect an activity with one that triggers it and the one it triggers. In practice, it is difficult to infer the causalities of activities by just observing their spatial or temporal locations, i.e., given that activity y_i occurs before activity y_j does not imply that y_i triggers y_j. We make an assumption that two activities can be related to each other only of they are within a certain spatio-temporal distance of each other. However, since there could be multiple people in the scene, there could be several activities occurring close to each other. We propose a greedy hill climbing method to construct the MRF as follows.

To begin with, we construct a graphical model where every node is connected to two other nodes which have the least spatio-temporal separation. In each iteration, we run loopy belief propagation to estimate the marginal probabilities. We randomly select an edge to add or delete, one at a time, till a maximum entropy for the sequence is reached or the maximum number of allowed iterations is reached. To add an edge, we start with the nodes which have the least spatio-temporal distance, and remove edges between nodes which are farther apart. We limit the edges of the graph using a Gaussian prior [13]. We also fix the maximum connectivity of a node so as to limit the number of loops in the graph. Three iterations of structure learning for a sample sequence of 5 activities is shown in Figure 4 a). Figure 4 b) shows the accuracy of recognition of activities for a single video with varying structures along with the corresponding entropy. The sequence contains 24 activities, 5 of which form the activity sequence which is illustrated in Figure 4 a). It can be seen that in most cases, the decrease in entropy of the system has resulted in an increase in recognition accuracy.

The overall algorithm of our approach is presented in Algorithm 1.

Algorithm 1 Algorithm for labeling activities in a test sequence using our context model

Input: $\mathcal{S}_{\mathcal{R}} = \{V_1 \dots V_{N_{\mathcal{R}}}\}$ Set of training videos containing activity annotations
An activity sequence $\mathcal{Y}_{\int}$ containing n activities occurring in close spatio-temporal vicinity $\{y_1 \dots y_n\}$.
Output: Labels of activities $\{x_1 \dots x_n\}$
Training: Train baseline classifiers $c_1 \dots c_N$ for N activities and model the spatio-temporal potential $\psi_{st}(x_i, x_j)$ between all pairs of activities using annotated training videos using Eqn (6).
Testing:
1) Identify activity regions using the activity segmentation algorithm.
2) Compute observation potential $\psi_o(x_i, y_i)$ for each activity segment given by the baseline classifiers using Eqn (3).
3) Initialize graph $\mathcal{G}$ containing n observation variables representing activity regions and n hidden variables representing the activity labels.
4) Run inference to generate posteriors;
5) Compute sum of posterior entropy of all hidden nodes E_{old}.
6) **while** $E_{old} > E_{thresh} || n_{edges} < n_{max}$ **do**
   Choose an edge randomly, add/delete edge;
   Run inference;
   Compute sum of posterior entropy of all hidden nodes E_{new};
   if $E_{new} < E_{old}$ **then**
   Incorporate change into the graph; $E_{new} = E_{old}$;
   end if
7) **end while**
8) Compute labels from posteriors and output labels.

F. Analysis of the Model

1) **Symmetry:** It should be noted that the order of activities cannot be ignored here. For example, the probability of the activity "approach the vehicle" followed by "enter the vehicle" is not the same as the probability of the activity "enter the vehicle" followed by "approach the vehicle" (which implies that it is a different person who is now approaching the vehicle). Therefore, the spatial potential is not symmetric ($\psi_s(x_i, x_j) \neq \psi_s(x_j, x_i)$). It is to be observed that the spatio-temporal potential for transition from c_i to c_j is learnt using examples from the training data where c_i occurs before c_j, whereas the spatio-temporal potential for transition from c_j to c_i is learnt using the training examples where c_j occurs before c_i. This takes care of the asymmetry in the potentials.
2) **Temporal proximity:** The temporal proximity is a measure of how far apart the two activities are in the

chain of activities in the scene. This is modeled in the temporal potential. The temporal potential takes care of the fact that two activities taking place far apart in the sequence is different from one activity taking place soon after another. The association potential models the frequency with which two activities occur with a particular temporal proximity.

3) **Spatial proximity:** Similar to the temporal potential, the spatial potential is an attempt to distinguish between different persons performing two activities and the same person performing the two activities.

III. Experiments and Results

A. Dataset

The goal of our approach is to model activity context in continuous videos, therefore, we perform experimentation on long duration realistic videos. Traditional datasets like Weizmann [23] and KTH [25] cannot be used to validate our system. Some other datasets like [26] contain long unsegmented video, but these activities are not related to each other and the sequence is not a realistic one. Therefore, we evaluate our system on two challenging datasets containing long duration activities: 1)The UCLA office dataset and 2)The publicly available VIRAT ground dataset [27].

The UCLA office dataset [19] consists of indoor and outdoor videos of single and two-person activities. Here, we perform experiments on the lab scene containing close to 35 minutes of video captured with a single fixed camera in a room. We work on 10 single person activities: Enter lab, exit lab, sit down, stand up, work on laptop, work on paper, throw trash, pour drink, pick phone receiver and place receiver down. There is very little variation in viewpoint, occlusion and scale here. The first half of the data is used for training and the second half for testing. Each activity occurs 6 to 15 times in the dataset.

The VIRAT dataset is a state-of-the-art activity dataset with many challenging characteristics, such as wide variation in the activities and a high amount of occlusion and clutter. It consists of surveillance videos of 11 scenes with different scales of resolution. These are parking lot videos involving single vehicle activities, person and vehicle interactions, and people interactions. There are also some group activities. This dataset consists of scenes captured on a single camera although the viewpoint can differ from one scene to the next. In any scene, the activities can occur at different orientations depending on the location. However, since these are wide-area videos, persons of interest are usually far away from the camera, the change in spatio-temporal distance with camera view is considered negligible. We have used parking lot scenes $VIRAT\_S\_0000$, $VIRAT\_S\_0401$ and $VIRAT\_S\_0502$ for the first set of experiments and all data for the second set. The length of the videos vary between $2-15$ minutes and containing up to 30 activities in a video. For every scene, the first half is used for training and the second half for testing.

We perform two sets of experiments on the VIRAT dataset, one on Release 1 and the other on Release 2 of the data. For Release 1, there are 6 activities which are annotated:

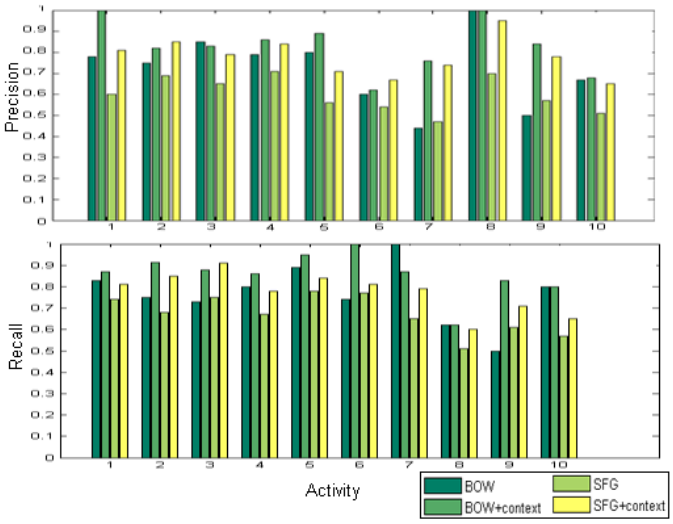

Fig. 5. The figure shows the precision and recall obtained on the UCLA office dataset and its comparison with the Bag-Of-Features baseline classifier and SFG [9]. The activities are: 1 - enter room, 2 - exit room, 3 - sit down, 4 - stand up, 5 - work on laptop, 6 - work on paper, 7 - throw trash, 8 - pour drink, 9 - pick phone, 10 - place phone down.

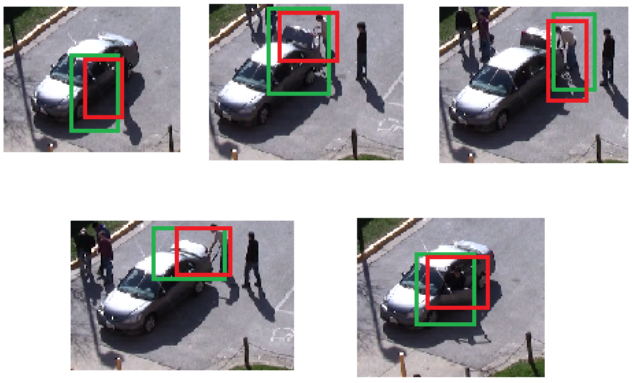

Fig. 6. The figure shows some examples of segmentation of activity regions. The obtained segmentation is marked in green while the true segmentation is marked in red.

Person entering vehicle, person exiting vehicle, person opening trunk, person closing trunk, person loading vehicle and person unloading vehicle. In release 2, additional 5 activities have been added: person carrying an object, person gesturing, person running, entering and exiting a facility. For release 1, we have provided comparison with the baseline Bag-of-Words classifier as well as the state-of-the-art String of Feature Graphs [9] method. For release 2, we show comparison with the baseline Bag-of-Words classifier.

B. Pre-processing

To label meaningful activities in a long-duration wide area video, the first step is to identify the spatio-temporal location of activities. We call this step as "activity segmentation". The video is first divided into overlapping time windows of fixed duration. Activity region computation is performed on

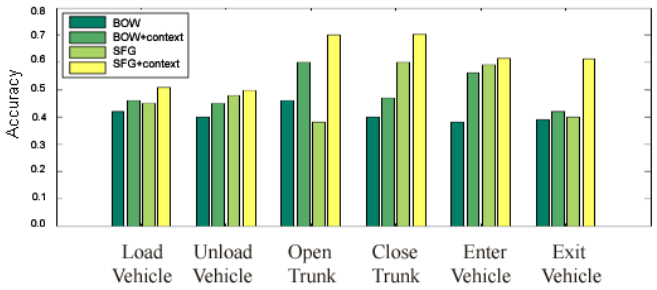

Fig. 7. Figure shows the accuracy of our method with the VIRAT release 1 dataset for six activities and its comparison with the baseline approaches Bag-of-Words and SFG [9] approach. The activities are: 1 - loading, 2 - unloading, 3 - open trunk, 4 - close trunk, 5 - enter vehicle, 6 - exit vehicle.

windows of three scales. Here each window consists of 30, 60 and 120 frames with an overlap of half the number of frames. Feature points are computed for each time window which contains a track and the time window is spatially clustered into as many regions as the number of tracks in the window. Here, we use the Space-Time Interest Points (STIP) [23] as our features.

For each time window, the baseline classifiers are used to assign a probability of the window belonging to a particular activity. All activities which do not correspond to the set of "interesting" activities are considered as "background activities". We also train a baseline classifier for background activities. For each set of overlapping windows, the window which has achieved the highest probability is chosen as the activity region. All regions which correspond to background activities are eliminated. Recognition is carried out on the remaining activity regions. An example of activity segments identified in a sequence is shown in Figure 6.

A limitation of this approach is that, when the segmentation algorithm fails to detect an activity segment, it is eliminated from further processing, thereby missed detections are not corrected.

C. Methodology

We used a randomly selected set of half the data for training and the other half for testing. During the training, we assume that the activity segmentation and the activity labels are available to us. We normalize all distances with respect to the scale of the video to make the approach invariant to scale. A threshold was set on the spatio-temporal distance between activities to determine if the relationship between them has to be modeled. We used the distance threshold as a bounding box of 4 times the average dimensions of the person in the scene and a time threshold of 20 seconds. These values have been fixed experimentally. The graphical model is constructed on individual activity sequences. Classification over an entire activity sequence is carried out using the proposed method. For each activity region in the sequence, the baseline classifier is applied to generate the observations. A graph is constructed based on the spatio-temporal distances between activities. Inference is carried out on the graph using the MRF parameters

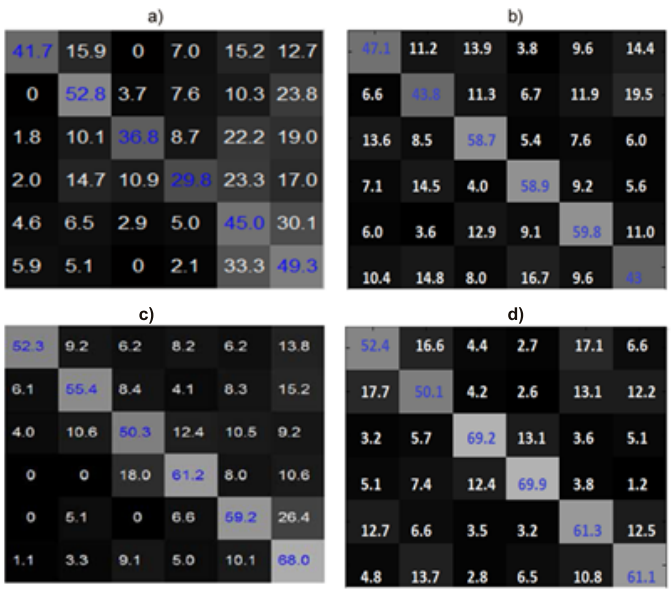

Fig. 8. The Figure shows the confusion matrix on VIRAT release 1 data. a)Result of applying the baseline classifier BOW to the data. b) Result of applying BOW+context on the data. c) Result of SFG baseline classifier. d) Result of SFG + context. The activities are: 1 - loading, 2 - unloading, 3 - open trunk, 4 - close trunk, 5 - enter vehicle, 6 - exit vehicle. The corresponding increase in recognition accuracy is evident from the graph.

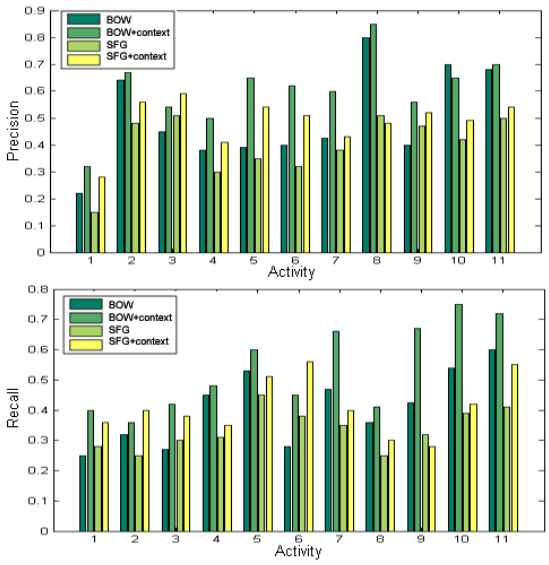

Fig. 10. The figure shows the precision and recall obtained on the VIRAT release 2 dataset and its comparison with the Bag-Of-Features and SFG approaches. The activities are: 1 - person loading an object to a vehicle, 2 - person unloading an object from a vehicle, 3 - person opening a vehicle trunk, 4 - person closing a vehicle trunk, 5 - person getting into a vehicle, 6 - person getting out of a vehicle; 7 - person gesturing, 8 - person running, 9 - carrying load, 10 - entering facility, 11 - exiting facility.

computed during training. Labels are assigned to each activity region based on the posterior probabilities.

D. Analysis of the Results

1) UCLA office dataset: For the UCLA dataset, we consider only single person activities in a high resolution video

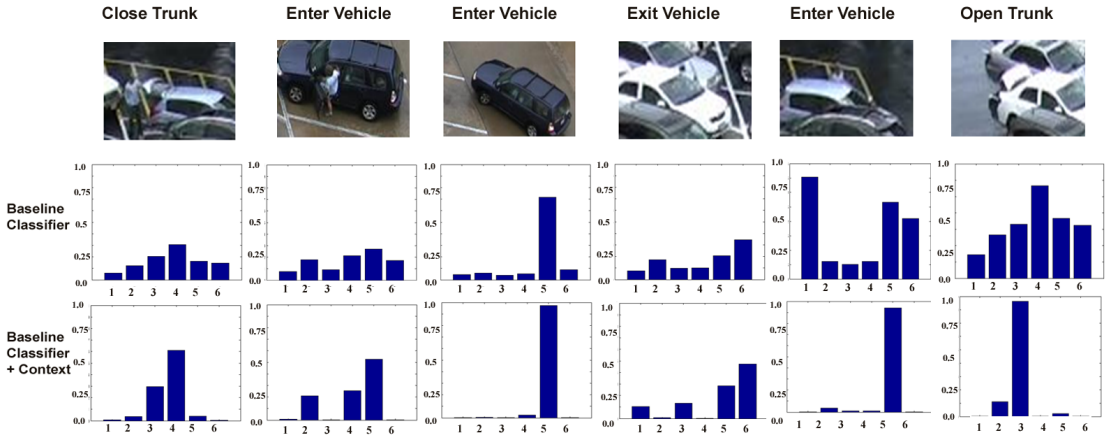

Fig. 9. The comparison of the prior probabilities which are the output of the baseline classifiers with the posterior probabilities which is the output of our algorithm for a set of six activities. The output of our algorithm is seen to have a more well defined peak (less uncertainty) as compared to the baseline classifier. For the last two, it is seen that the addition of context corrects an incorrect classification. The activities in order are: 1 - person loading an object to a vehicle, 2 - person unloading an object from a vehicle, 3 - person opening a vehicle trunk, 4 - person closing a vehicle trunk, 5 - person getting into a vehicle, 6 - person getting out of a vehicle

with little variation in viewpoint and occlusion. Although this dataset has been used for experimentation in [19], the events which have been classified for the lab data and the accuracy of recognition for each event have not been provided by the authors. Therefore, we provide comparison to the baseline methods used here, which is the Bag-of-Words and the SFG [9]. In both cases, it can be seen that the addition of context improves performance. The Bag-of-Words classifier gives an overall accuracy of 75.4%.For some activities, the BOW classifier was able to identify all instances, therefore no further improvement was possible. An overall accuracy of 86.7% was achieved with the addition of context. For the SFG method, an overall accuracy of 62.3% was achieved while the addition of context gave an accuracy of 77.9%. The values of precision and recall for BOW and BOW+context, SFG and SFG+context are shown in Figure 5.

2) VIRAT dataset: We compare our approach with two well known approaches: the Bag-of-Words approach and the String of Feature Graphs (SFG) approach which is a recent method that provides state-of-the-art performance on multi-object data in realistic videos. This method also models spatio-temporal relations at the feature level.

For the VIRAT release 1 data, we demonstrate our method using the BOW as well as SFG as baseline classifiers in Figure 7. We have also shown the results of the baseline classifiers for comparison. We can see that our method performs better than the SFG method in most cases. An overall accuracy of 40% was obtained using BOW and 51.3% was obtained using the SFG method. The usage of our method on BOW resulted in an overall accuracy of 52.4% while the usage of our method on SFG resulted in an accuracy of 61.5%. The accuracy of recognition for activities "loading" and "unloading" was found to be slightly lower than the rest since they involve similar gestures. The confusion matrix for the 6 activities using BOW. BOW+context, SFG and SFG + context is shown in Figure 8.

In Figure 9, we illustrate the difference between the output of the baseline classifier and our algorithm for different activities like enter vehicle, exit vehicle, open and close trunk. It can be seen that the output of our algorithm has a more well defined peak in probability, which in turn means less uncertainty in prediction as compared to the baseline classifier. This shows that the confidence of classification can be increased with the use of context. In the last two cases, the addition of context corrects an incorrect classification (represented by the highest probability).

The second set of experiments was conducted on the release 2 of VIRAT dataset. This dataset contains five additional activities - person carrying load, gesturing, running, entering and exiting facility. These activities add some additional context information to the data. We provide the precision-recall values for each activity as well as the comparison with Bag-Of-Features and SFG approaches in Figure 10. Here also, we find that the addition of context helps in better recognition in both cases. The overall accuracy of BOW+context was 52.6% while BOW had an accuracy of 41.3%. The overall accuracy of SFG was 37.8% while the overall accuracy of SFG+context was 46.4%. It was seen that the activities "enter vehicle" and "load vehicle" were often confused with each other in the absence of context. But the use of context tells us that if a person opens the trunk, he is likely to load it, whereas if the person opens a door, he is likely to enter it. This contextual information was captured by our model and brought about a an improvement in the performance. It was seen that the method shows an improvement over the baseline classifiers in the case of partial occlusion as well as noise due to shadows and clutter.

IV. Conclusion

In this paper, we have shown that the spatio-temporal relationships between different activities in a scene can be used as context in the recognition of activities. We illustrated

a scheme based on graphical models used to learn the spatio-temporal relationships from training video. We inferred the most probable set of labels for the activities in the test video given their spatio-temporal configurations and observation potentials generated from weak classifiers.

In future, there are different directions in which this work can evolve. We plan to create an integrated framework which can identify interesting activity regions as well as recognize them using contextual information. We also plan to extend the method to be able to correct missed detections and false positives using the contextual information available. It might also be possible to model this contextual information using a discriminative approach such as structural SVMs.

REFERENCES

[1] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys*, 2011.

[2] I. Laptev and T. Lindeberg, "Local descriptors for spatio-temporal recognition," in *First International Workshop on Spatial Coherence for Visual Motion Analysis*, 2004.

[3] R. Chaudhary, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Computer Vision and Pattern Recognition*, 2009.

[4] M. Lee and R. Nevatia, "Human pose tracking in monocular sequence using multilevel structured models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[5] M. Ryoo and J. Aggarwal, "Recognition of composite human activities through context-free grammar based representation," in *Computer Vision and Pattern Recognition*, 2006.

[6] D. Kuettel, M. Breitenstein, L. V. Gool, and V. Ferrari, "What's going on? discovering spatio-temporal dependencies in dynamic scenes," in *Computer Vision and Pattern Recognition*, 2010.

[7] M. Ryoo and J. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *International Conference on Computer Vision*, 2009.

[8] Y. Zhang, X. Liu, M.-C. Chang, W. Ge, and T. Chen, "Spation-temporal phrases for activity recognition," in *European Conference on Computer Vision*, 2012.

[9] U. Gaur, Y. Zhu, B. Song, and A. Roy-Chowdhury, "String of feature graphs analysis of complex activities," in *International Conference on Computer Vision*, 2011.

[10] W. Brendel and S. Todorovic, "Learning spatiotemporal graphs of human activities," in *International Conference on Computer Vision*, 2011.

[11] K. Tang, L. Fei-Fei, and D. Koller, "Learning latent temporal structure for complex event detection," in *Computer Vision and Pattern Recognition*, 2012.

[12] Y. J. Lee and K. Grauman, "Object graphs for context aware category discovery," in *Computer Vision and Pattern Recognition*, 2010.

[13] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *Computer Vision and Pattern Recognition*, 2010.

[14] A. Gupta, P. Srinivasan, J. Shi, and L. Davis, "Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos," in *Computer Vision and Pattern Recognition*, 2009.

[15] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *Pattern Analysis and Machine Intelligence*, 2011.

[16] A. Prest, C. Schmid, and V. Ferrari, "Weakly supervised learning of interactions between humans and objects," INRIA, Technical Report, 2010.

[17] V. I. Morariu and L. S. Davis, "Multi-agent event recognition in structured scenarios," in *Computer Vision and Pattern Recognition*, 2011.

[18] J. Varadarajan, R. Emonet, and J. Odobez, "Bridging the past, present and future: Modeling scene activities from event relationships and global rules," in *Computer Vision and Pattern Recognition*, 2012.

[19] M. Pei, Y. Jia, and S.-C. Zhu, "Parsing video events with goal inference and intent prediction," in *International Conference on Computer Vision*, 2011.

[20] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *Computer Vision and Pattern Recognition*, 2011.

[21] C. C. Loy, T. Xiang, and S. Gong, "Time-delayed correlation analysis for multi-camera activity understanding," *International Journal of Computer Vision*, 2010.

[22] T. Lan, L. Sigal, and G. Mori, "Social roles in hierarchical models for human activity recognition," in *Computer Vision and Pattern Recognition*, 2012.

[23] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *International Conference on Computer Vision*, 2005.

[24] Y. Li and R. Nevatia, "Key object driven multi-category object recognition, localization and tracking using spatio-temporal context," in *European Conference on Computer Vision*, 2008.

[25] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *Internation Conference on Pattern Recognition*, 2004.

[26] M. Ryoo, C. Chen, J. Aggarwal, and A. Roy-Chowdhury, "An overview of contest on semantic description of human activities (sdha) 2010," in *International Conference on Pattern Recognition*, 2010.

[27] S. Oh and et al, "A large-scale benchmark dataset for event recognition in surveillance video," in *Computer Vision and Pattern Recognition*, 2011.

Nandita M. Nayak Nandita M. Nayak received her Bachelor's degree in Electronic and Communications Engineering from M. S. Ramaiah Institute of Technology, Bangalore, India in 2006 and her Master's degree in Computational Science from Indian Institute of Science, Bangalore, India. She is currently a Ph.D. candidate in the Department of Computer Science in University of California, Riverside. Her main research interests include image processing and analysis, computer vision and artificial intelligence.

Yingying Zhu Yingying Zhu received her Bachelor's degree in Engineering in 2004 from Shanghai Maritime University. She received her Master's degree in Engineering in 2007 and 2010 from Shanghai Jiao Tong University and Washington State University, respectively. She is currently pursuing the Ph.D. degree within Intelligent Systems in the Department of Electrical Engineering at the University of California, Riverside. Her main research interests include computer vision, pattern recognition and machine learning, imgage/video processing and communication.

Amit K. Roy-Chowdhury Amit K. Roy-Chowdhury received his Masters degree in systems science and automation from the Indian Institute of Science, Bangalore, India, and the Ph.D. degree in electrical engineering from the University of Maryland, College Park. He is a Professor of electrical engineering and a Cooperating Faculty in the Department of Computer Science, University of California, Riverside. His broad research interests include the areas of image processing and analysis, computer vision, and statistical signal processing and pattern recognition, where he has over 100 technical publications. His current research projects include intelligent camera networks, wide-area scene analysis, motion analysis in video, activity recognition and search, video-based biometrics (face and gait), and biological video analysis. He is a coauthor of two books - Camera Networks: The Acquisition and Analysis of Videos over Wide Areas and Recognition of Humans and Their Activities Using Video. He has been on the organizing and program committees of multiple computer vision and image processing conferences and is serving on the editorial boards of multiple journals.