

Context-Aware Activity Forecasting

Anirban Chakraborty and Amit K. Roy-Chowdhury*

Electrical and Computer Engineering, University of California, Riverside, USA

Abstract. In this paper, we investigate the problem of forecasting future activities in continuous videos. Ability to successfully forecast activities that are yet to be observed is a very important video understanding problem, and is starting to receive attention in the computer vision literature. We propose an activity forecasting strategy that models the simultaneous and/or sequential nature of human activities on a graph and combines that with the interrelationship between static scene cues and dynamic target trajectories, termed together as the ‘activity and scene context’. The forecasting problem is then posed as an inference problem on a MRF model defined on the graph. We perform experiments on the publicly available challenging VIRAT ground dataset and obtain high forecasting accuracy for most of the activities, as evidenced by the results.

1 Introduction

In computer vision literature, one major topic of interest is to automatically detect and recognize human activities in a video. The methods developed in the literature on activity recognition range from analyzing simple individual actions such as those discussed in [1, 2] to more natural and complex human activities involving one or more actors in the scene [3–5]. However, these methods provide ‘after-the-fact’ recognition once the activity of interest is complete. Forecasting activities into the future much before they are observed is an important problem for many application scenarios and can be useful in designing anomalous event detection schemes. However, it hasn’t yet received much attention in the computer vision community.

We have seen some recent developments in the field of activity prediction or forecasting and two classes of such problems have been introduced in the literature. The first class of problems looks into early recognition of ongoing activities [6–8] and is defined in the literature as an inference of the ongoing activity given temporally incomplete observations. In this problem, the first few frames of the video sequence containing an activity are observed and an early classification of the ongoing activity needs to be achieved. The second class of the problems seeks to forecast future activities in continuous videos [9] well before they are observed. This problem can be generally stated as an anticipation about future activity classes in a continuous video, where no observation of any

* Corresponding author. E-mail: amitrc@ece.ucr.edu

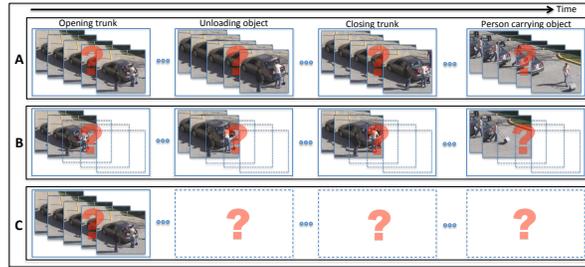


Fig. 1. Different types of problems in human activity analysis. The figure shows four consecutive activity sequences for an actor - opening the trunk of a vehicle, unloading an object from the vehicle, closing the trunk, and the actor carrying the unloaded object, performed in that order. Three categories of activity analysis problems are presented on these sequences. (A) The classic activity recognition problem: each of the activity sequences is fully observed before the activity labels are predicted. (B) Early prediction of ongoing activity: only a few initial frames per activity sequence is observed and the goal is an early prediction of the activity classes from these incomplete observation sets. (C) Forecasting of future activities in absence of observation: at any point of time in a continuous video all activities occurring upto that time point are observed and the goal is to forecast the labels for future activities without the availability of observation for any of them.

future activity is available and all past activities are observed. The differences between these two problems and how each of them are principally different from a standard activity recognition problem are described through Fig. 1.

In this paper, we propose a method that not only attempts to solve the problem of forecasting unseen future activities (the second class of problem) but also jointly recognizes the activities that have already taken place and were observed. In most cases, it can be observed that activities performed by an actor occur following fixed temporal sequences. For example, if a person carries a bag and walks towards the trunk of a parked car, s/he is most likely to open the trunk, load the bag into it and then close the trunk. Also, for collective activities it can often be seen that actions of the actors involved are strongly synchronized with each other within a spatio-temporal window. All of these are collectively termed in this paper as ‘activity and scene context’ and we leverage upon these contextual information for successful recognition of observed activities and forecasting of unobserved future activities.

We formulate the joint recognition and forecasting problem probabilistically. The past, present and future activities in a video are modeled as the nodes of a graph and the activity and scene context are modeled as a Markov Random Field on the proposed graph structure. Then a suitable inference strategy is adapted for recognition and forecasting of the activity classes. We show experiments on a challenging and realistic activity dataset - the VIRAT ground dataset release 2 [10]. This dataset comprises of long duration video clips, each containing multiple activities that take place either simultaneously or sequentially, thereby making

these datasets both challenging and suitable for testing the proposed spatio-temporal context based activity forecasting method.

2 Relation to Existing Work

In computer vision research, majority of the works related to human activity in video has focused on the task of recognition of simple to more complex activities [11]. Many existing works exploring context focus on spatio-temporal relationship of features [3, 12], interactions of objects and actions/activities [13, 5, 14], AND-OR graph based scene representation [15, 16]. Methods such as [17–22] studied spatio-temporal relationship between activities in a scene.

There have been some recent works on the emerging topic of early recognition of ongoing activities. The method in [6] approached this problem by representing an activity as an integral histogram of spatio-temporal features and subsequently used a novel dynamic bag-of-words approach to model how these feature distributions change over time. Authors in [7] developed a ‘spatial-temporal implicit shape model’ which characterizes the space time structure of the sparse activity features extracted from a video and the early recognition is done using a random forest structure. The authors in [8] proposed a max-margin framework based on structured SVM to recognize partially observed events. However, these methods rely on the availability of a partial set of information for the ongoing activity where a typical activity forecasting problem should be able to forecast probable future activities well before the start of the activity segments.

Very recently, the activity forecasting problem was introduced in [9]. The authors combined semantic scene labeling with inverse optimal control to forecast probable actor trajectories, which in turn help predict destinations and future actions. However, there are a number of differences between our method and [9]. [9] investigates the effect of the static scene environment on future activities, whereas we use both static cues from the scene and dynamic cues from target trajectories and model their interrelationships for forecasting future activities. Unlike a pure trajectory based approach in [9], we combine the target trajectory information with the motion based activity recognition methods in a dynamical model. Finally, we show results on the recent release of VIRAT dataset containing 11 diverse activities where [9] tested their forecasting method on a dataset of three activities.

3 Overview of the Proposed Method

In this work, we propose a strategy to jointly recognize and forecast activities in long duration continuous videos. The method attempts to recognize activities that have already been observed in a video while forecasting the most probable categories of future activities, yet to be observed in that video sequence.

A typical surveillance video contains multiple activities occurring simultaneously or in succession at different portions of the scene. In such videos, it can be observed that a specific activity by an actor is often followed by another activity

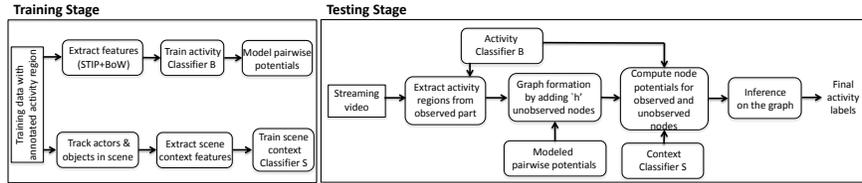


Fig. 2. The overall pipeline: Training and Testing.

by the same actor and this pattern repeats itself through and across the videos given the similarity in scenes. Therefore, an actor’s future activities can often be inferred from one or more of its previous activities. Moreover, in group activities where multiple actors are involved, one actor’s observed activity pattern can help us forecast another’s future actions. We call this ‘activity context’ and it can be modeled on the edges of an *activity graph* to aid recognition and prediction. As the number of observed activities can increase with time, the graph formation strategy is dynamic with an aim to keeping the size of the graph constant. This is discussed in Sec. 4.1.

After graph formation, a ‘Markov Random Field’ (MRF) (see Sec. 4.2) is defined on the graph. The edge potentials defined on each of the edges of the graph are modeled using the frequencies of occurrence of pairs of activities in a tight spatio-temporal proximity (Sec. 4.3) and are directly learned from a set of annotated training videos. The node potentials for the observed nodes (observed activities) are obtained using the likelihood of the activities, given by a set of activity classifiers when applied on the features (STIP+BoW) extracted from the observed activity regions (Fig. 2). The node potentials for all the unobserved nodes (unobserved future activities) are initially set as uniform distributions in absence of any other specific information. However, as in most cases, the activity can be characterized by the proximity and motion of the actor relative to a number of key points and detected secondary objects in the scene. These scene specific information, termed as the *scene context*, help modifying the observation/node potential of the first unobserved activity node in immediate future for every actor (Fig. 3). Please note that these scene contexts and the previously introduced spatio-temporal activity context are collectively termed as ‘activity and scene context’.

An object detector and a person tracker are employed to extract and estimate various scene context features (see Sec. 4.4) for each actor in the scene at each time point. A trained classifier, when applied on the scene context features extracted from the observed video, provides us with the observation potentials of the aforementioned nodes. The edge potentials remain fixed for the graph across all time points. Finally, the joint activity recognition and forecasting problem can be posed as an inference problem on the MRF just described, which is solved using an iterative ‘message passing’ algorithm.

4 Activity Forecasting Framework

Let a complete continuous video clip be V , v_t being the portion of V that is observed upto time t and let v'_t be the portion that is yet to be observed. Therefore, $v_t \cup v'_t = V$. v_t contains a number of activity regions and the set of K most recent observations from these activity regions is given as $Y = \{y_1, y_2, \dots, y_K\}$. More clearly, the observation y_k denotes the image observation of an activity, i.e., the features computed from the k^{th} activity region amongst the most recent K activity observations. A subset of these observations is the set of observed activities by one individual actor. If there are n^o actors $O = \{o_1, o_2, \dots, o_{n^o}\}$ in the scene at time t , the set of activities by actor $o_i \in O$, observed so far would be $Y^i = \{y_1^i, y_2^i, \dots, y_{N^i}^i\}$. Further, we define a forecasting horizon h over which we intend to do activity forecasting. Note that h is not a time window, rather it denotes the number of future activities per actor we would be predicting ahead of the current time instant. Therefore, we can define a total of $(K + n^o \cdot h)$ variables representing the hidden activity labels, which we estimate. Let the set of these labels be $X_t = \{x_1, x_2, \dots, x_K, x_{K+1}, \dots, x_{K+n^o \cdot h}\}$. The two subsets of this label set are the one containing the labels with associated observations, $X_t^{obs} = \{x_1, x_2, \dots, x_K\}$ and another containing the labels for which no observation is available, $X_t^{unobs} = \{x_{K+1}, x_{K+2}, \dots, x_{K+n^o \cdot h}\}$. Let the hidden variable/label for k^{th} activity by actor o_i be represented as $x_k^i \in X_t$.

In the next subsections we introduce the structure of an ‘activity graph’, the potential functions associated with an MRF defined on it and how to do recognition and prediction as inference on this MRF to obtain the labels of the hidden states X_t .

4.1 Activity Graph Formation

A graph is built with the atomic activities (both observed and unobserved) as nodes and the activity contexts are modeled on the edges of the graph. The characteristics and definitions of various components of the graph are, as follows,

Each node in the graph is an atomic activity. Let the set of all the nodes in the graph at any given time instant t be \mathcal{N}_t . Let a node corresponding to an activity by actor o_i be n_k^i . Then, $n_k^i \in \mathcal{N}_t$. The hidden variable corresponding to the node n_k^i is x_k^i (the activity label), the value of which is to be estimated. An edge between two activity nodes represents the spatio-temporal context between them. Let the set of all the edges in the graph be \mathcal{E}_t . The nodes corresponding to the already observed activities in the video are called observed nodes (\mathcal{N}_t^{obs} , blue nodes in Fig. 3). The unobserved activities are represented by the unobserved nodes (\mathcal{N}_t^{unobs} , white nodes in Fig. 3). Observed edges are those which connect two observed nodes (\mathcal{E}_t^{obs} , blue lines in Fig. 3). If both the terminal nodes of an edge are unobserved, it is called an unobserved edge (\mathcal{E}_t^{unobs} , red lines in Fig. 3). If an edge connects two nodes one of which is observed and the other unobserved, it is called a semi-observed edge ($\mathcal{E}_t^{semi-obs}$, black lines in Fig. 3).

Two observed nodes are connected by an edge if the corresponding activities occur in a predefined spatio-temporal proximity. But for the unobserved

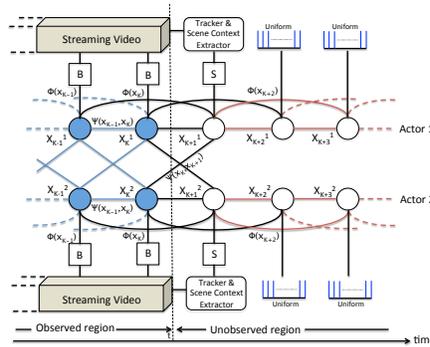


Fig. 3. A snapshot of the graph structure for activity forecasting for two actors in the scene at any time instant ‘ t ’. ‘B’ denotes a trained activity classifier for observed activity recognition and ‘S’ denotes a scene-context classifier.

nodes and edges, this strategy cannot be adapted as we are unaware of both the spatial location and time of any future activity. Even the exact number of future activities in a video clip at any observational time point is also unknown. Therefore, whenever an activity is observed, we add one more unobserved node (corresponding to the actor of that activity) in the graph and drop the node corresponding to the oldest observed activity. Thus the total number of observed (K) and unobserved ($n^o.h$) nodes remains constant through the video. Please note that an actor might exit the scene, or the video sequence might end before all the future activity nodes are observed. The unobserved nodes are time ordered and two consecutive unobserved nodes pertaining to the activities to be performed by the same actor are connected using an *unobserved edge*. Second order connections are also made for two unobserved nodes. Finally, the *semi-observed edges* are used to connect the last two observed nodes per actor and the two unobserved nodes in the immediate future.

4.2 Markov Random Field Modeling

The set of random variables associated with nodes \mathcal{N}_t is $X_t = \{x_1, x_2, \dots, x_K, x_{K+1}, \dots, x_{K+n^o.h}\}$, which are to be estimated given all observations Y_t . These random variables correspond to the state of each node in the graph and the support for each of these variables is the candidate set of activities (C).

Then the overall MRF is expressed as

$$P(X_t; Y_t) = \frac{1}{Z} \prod_{k=1}^{K+n^o.h} \phi(x_k, y_k) \prod_{\substack{(k,l) \\ : (n_k, n_l) \in \mathcal{E}_t}} \psi(x_k, x_l) \quad (1)$$

Here $\phi(x_k, y_k)$ represents the node potential of any node $n_k \in \mathcal{N}_t$, and $\psi(x_k, x_l)$ is the edge potential from node n_k to node n_l . To estimate the optimal

state for every node, we have to maximize $P(X_t; Y_t)$. Towards that objective, we first estimate the approximate marginal distributions $P(x_k; Y_t)$ at each node using a belief propagation scheme as described later. The optimal states that maximize the posterior distribution could be then estimated by maximizing the marginals independently.

4.3 Edge/Activity Context Potential

The activity context potential is defined on the edges of the graph, in each of \mathcal{E}_t^{obs} , \mathcal{E}_t^{unobs} , $\mathcal{E}_t^{semi-obs}$. This potential function models the association between any two activities occurring immediately one after the other or in close spatio-temporal succession. For any two nodes n_k^i and n_l^j (the corresponding labels being x_k^i and x_l^j respectively) such that $(n_k^i, n_l^j) \in \mathcal{E}_t$, the inter-activity potential is given as,

$$\begin{aligned} \psi \left(x_k^i = c_m, x_l^j = c_n \right) &= f_{mn,1}^s \text{ if } i = j, |l - k| = 1 \\ &= f_{mn,2}^s \text{ if } i = j, |l - k| = 2 \\ &= f_{mn}^d \text{ if } i \neq j \end{aligned} \quad (2)$$

All these values $f_{mn,1}^s$, $f_{mn,2}^s$ and f_{mn}^d are computed from the annotated training data. $f_{mn,1}^s$ is computed as the ratio of the number of times the same actor performs the activities c_m and c_n immediately one after another to the total number of times the activity c_m is performed in the training data. $f_{mn,2}^s$ is computed as the number of times the same actor performs activities c_m and c_n with the gap of exactly one activity in between them, and it is expressed as a ratio to the total number of times the activity c_m is performed. Finally, f_{mn}^d is obtained as the ratio of the number of times activities c_m and c_n are performed in a close spatio-temporal vicinity by two different actors to the number of times c_m is observed in the video. The same spatio-temporal proximity thresholds are also used in forming the graph, as discussed in Sec. 4.1.

For computing the activity context, only close spatio-temporal neighbors (1st and 2nd order connections) are considered, as we have observed that sub-sequences of relatively smaller length show stronger trends in repeating themselves than the longer activity sequences. Thus in the training videos, we examine all such 2 and 3-tuples of activity sub-sequences and model their pairwise relationships. This also helps us in correcting for any false positives and missing activities.

4.4 Node Potentials

The node potential is the likelihood of occurrence of a particular type of activity as observed in the video data. As there are specifically two types of nodes in our graph (observed and unobserved), we devise separate strategies for computing node potentials for these two categories of nodes.

Observed Nodes: From the annotated training data, we identify the activity regions and we train one activity classifier, the output of which is the probability of a given activity belonging to a particular category. Features at these activity regions are the observation variables and if any of the observation variables is associated with the k^{th} observed activity by actor o_i , it is denoted as y_k^i . A classifier can be employed to estimate the probability of an observation y_k^i resulting from an activity belonging to a particular category $c_p \in C$. Thus, if the set of trained baseline classifiers is B , then the observation/node potentials of the node n_k^i is given as

$$\phi(x_k^i, y_k^i) = p(x_k^i | y_k^i, B), \text{ if } n_k^i \in \mathcal{N}_t^{obs} \quad (3)$$

Although we have mentioned a particular feature and type of baseline classifier in the experiments section, any other discriminative classifier and low level motion features could be used for this purpose.

Unobserved Nodes: The node potentials, thus obtained above, are potentials for the observed nodes. However, for an unobserved node n_k^i , observation y_k^i is yet to be obtained (i.e., $y_k^i = \emptyset$) and hence a future activity is equally likely to belong to any category out of the M possible activity types in the dataset, i.e.,

$$\phi(x_k^i, y_k^i = \emptyset) = (1/M)\mathbf{1}^T, \text{ if } n_k^i \in \mathcal{N}_t^{unobs} \quad (4)$$

Although no low level motion feature is available for a future activity, its likelihood of being categorized as a specific activity can sometimes be substantially improved over $1/M$ with the help of some secondary observations from the scene, termed as the ‘scene context’ through this paper.

Scene Context Classifier: Often times, an activity is characterized by its interaction with other objects in the scene. For example, in [10], ‘opening trunk’/‘closing trunk’, ‘loading a vehicle’/‘unloading a vehicle’, ‘getting into a vehicle’/‘getting out of a vehicle’ - all these activities have at least one thing in common, i.e. the actor interacts with a parked car in all of them. Similarly, ‘entering a facility’/‘exiting a facility’ are both associated with a detectable entry-exit point of a facility in the scene, probably the doorway of a building. Therefore, knowledge about the locations of these objects, key scene elements and whether an actor is going to interact with either of them in near future could help us ascertain that the future activity belongs to a much smaller subset of all possible activities. These information, as a whole, is termed as the ‘scene context’. It is represented by a set of variables comprising of the locations/bounding boxes of all the secondary objects, and key points in the scene that are related to one or more types of activities, location and motion information of the actor relative to these objects/key points. Please note that the scene context is computed individually for every actor in the scene and the values of them naturally change with time. Details on such context features in relation to experiments on VIRAT data is given in Sec. 5.

The computed scene context features are averaged over a predefined time window to generate a smoothed scene context feature vector per actor at each time point. Let, at time point t , the scene context feature computed for actor o_i be $f_t^{o_i} = \langle f_{t,1}^{o_i}, f_{t,2}^{o_i}, \dots, f_{t,Nf}^{o_i} \rangle$. As these features are computed in between two successive activities, the pair $(f_t^{o_i}, a_{k+1})$ completes the representation of the scene context, where $f_t^{o_i}$ is computed at a time t , after which the next activity o_i is going to perform is a_{k+1} . Such features for all the actors over the entire training dataset are combined and a scene-context classifier S is trained. Given a test video, at each time point, whenever we want to run the recognition and prediction, we compute the scene context features. If an actor o_i has already performed k activities and its computed scene context features at time t is $f_t^{o_i}$, then the classifier S provides us with the likelihood of the next activity (X_{k+1}^i) that o_i is going to perform, which is also the node potential for the first unobserved activity node for o_i at time t . Therefore,

$$\phi(x_{k+1}^i, y_{k+1}^i = \emptyset) = p(x_{k+1}^i | f_t^{o_i}, S), \quad (5)$$

where $n_k^i \in \mathcal{N}_t^{obs}$, $n_{k+1}^i \in \mathcal{N}_t^{unobs}$. For all other future unobserved nodes for actor o_i , the node potentials remain uniform (see Eqn. 4) until the next observation is obtained. It can be noted that as $f_t^{o_i}$ is time varying, the estimated node potential also changes from frame to frame and needs to be re-estimated.

4.5 Inference: Loopy Belief Propagation

The next step is to do the inference on the MRF, which involves the computation of the marginal probability distributions for the states x_k of each node $n_k \in \mathcal{N}_t$, given the observations Y_t . For computation of the marginals at each node, we choose to use *Loopy Belief Propagation* (LBP) based on the *Sum-Product* algorithm [23]. If LBP converges at iteration L , the estimated marginals at each node would be $P^{(L)}(x_k; Y_t)$ and the MAP estimates for the most likely states is computed as $\hat{x}_k = \arg_{x_k} \max P^{(L)}(x_k; Y_t)$. This optimum state corresponds either to the recognized or the predicted label of activity node n_k depending on the type of the node.

5 Experimental Results

To assess the effectiveness of our proposed method in activity forecasting, we perform experiments on the publicly available state-of-the-art VIRAT ground dataset[10] that contains 11 activities in different scenes (see supplementary for the list of activities). We perform two similar sets of experiments corresponding to two recognition schemes used for labeling observed activities, viz. 1. An automated classifier (BOW+SVM), 2. Ground truth activity labels. For experiment set 1, we use half of the data for each scene for training our model and the rest is used for testing. For experiment set 2, however, training is only needed for the scene context based future activity classifier and only a fifth of the data in each scene is used for training and we test our method on the rest of the data.

5.1 Preprocessing

Given a test video sequence, the first task is to obtain the observed activity regions. As activity regions overlap with the motion regions in a video, a background subtraction method [24] can be used to locate the motion regions. Moving persons and vehicles are identified by using an available software [25]. Doors, bags, boxes etc. are detected by using a detector similar to [26] on the entire scene. A tracking method [27], when applied on the detected actors' bounding boxes, provides us with the trajectories of the actors.

5.2 Extraction of Scene Context Features

For our experiments on VIRAT dataset, the set of scene context features computed are - 1. Are cars parked in the scene? (1-Y, 0-N), 2.Distance from the closest parked vehicle normalized by length of diagonal of the car bounding box, 3.Heading towards the closest parked vehicle, 4. Largest overlap of the actor bounding box with the bounding box of a parked vehicle normalized by area of the actor bounding box, 5. Is there one or more entry/exit points to facilities in the scene? (1-Y, 0-N), 6. Distance from the closest entry/exit point normalized by the length of the diagonal of the actor bounding box, 7. Heading towards the closest entry/exit point, 8. Is an object seen on the actor? (1-Y, 0-N), 9. Average velocity of the actor, 10. Time elapsed since last observed activity. For other datasets, the objects of interest will be recognized from the segmented training videos and the generalized scene context features can be estimated by keeping the same relationship between actor and objects. The features are estimated at every frame for every actor using the actor track and locations of detected objects in the scene. The features extracted from the training videos are further used to train a *bag of decision trees* containing 200 fully grown trees. At every frame, the next activity class is used as label. In training videos, given a scene context feature vector extracted at any frame, the trees individually vote and the normalized votes are used as the likelihood for probable future activity class labels.

5.3 Motion Feature Extraction for Observed Activities

In experimental setup 1 (automated classifier based labels for observed activities), we have used a 'Bag-of-Features' approach over 'Space Time Interest Points' (STIP) [28] due to its popularity in the literature for recognition of atomic activities. The STIPs based on Harris and Förstner operators are computed for every activity region in the training data. Feature vectors computed at each point are clustered and quantized to generate a codebook during the training phase and each activity category is modeled as a distribution over this codebook. A multiclass SVM classifier is trained with these features and the corresponding activity labels obtained from the annotated training data. Similarly, for test video inputs, the STIPs are computed and probable activity regions are identified where a significant number of points from the trained vocabulary is observed.

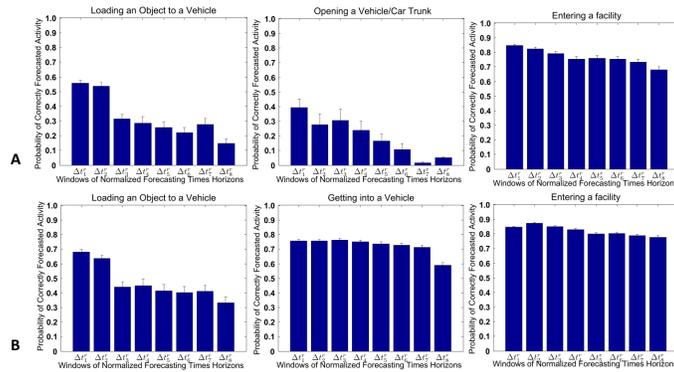


Fig. 4. Increasing trend of forecasting probabilities for different classes of activity (observed in the test set) with time. The positive direction of the time axis indicates increasing time gap from the instant at which the activity to be forecast happens. (A) Probability with which the ground truth activity is forecast as the next activity in exp. setup 1, (B) Similar increasing trend as observed in exp. setup 2.

5.4 Experiment Set 1

In this section, we present the experimental results when a classifier (BOW + SVM) is used to generate node potentials corresponding to already observed activities. At every fifth frame between two activities in a continuous video, we forecast the next activities that an actor is going to perform using the previously observed activities in the video as well as estimated scene context at that frame. At any given time before an activity is performed, the proposed method estimates the probabilities of various candidate future activity labels and these forecasting probabilities can vary with time as the actor moves and the scene context changes. In Fig. 4(A) and Fig. 5(A), we examine this variation in forecasting probabilities with time for the next activity that an actor may perform.

Let us assume that an actor has already performed an activity A_{last} upto time t_{last} (or just entered the scene) and is going to perform A_{next} at time t_{next} . At every time point t between t_{last} and t_{next} , we estimate the probability with which A_{next} is forecast as the next unobserved activity label, and thus $t^r = (t - t_{next})$ is the forecasting horizon. The average probability of forecasting the ground truth activity (A_{next}) over all instances of A_{last} in the dataset is computed and its time evolution is observed. Please note that for the same future activity performed, the time gap $[t_{last}, t_{next}]$ varies for different instances and hence is normalized between $[-1, 0]$ (-1 denotes the end time of last observed activity or the time of actor’s first appearance and 0 is the time when the next activity is going to occur in future). This time gap is split into 8 equal ranges ($\Delta t_i^r, i = 1, \dots, 8$) and the average probabilities (with standard errors) of the next ground truth activities are plotted.

Fig. 4(A) shows the time evolution of probabilities of three activity types (loading obj. to vehicle, opening trunk, entering facility) as the next activity in

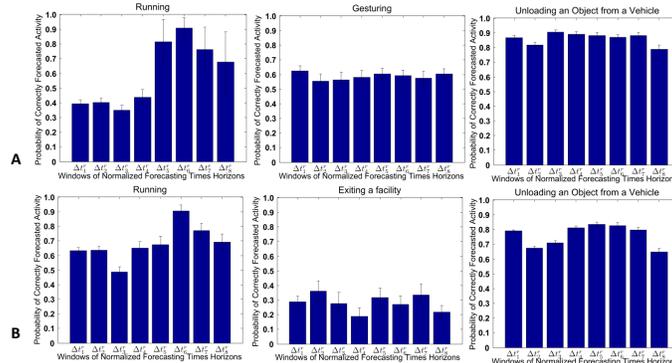


Fig. 5. Time evolution of forecasting probabilities for different classes of activity (in the test set), where no apparent trend is observed. The positive direction of the time axis indicates increasing time gap from the instant at which the activity to be forecast happens. (A) Probability with which the ground truth activity is forecast as the next activity in exp. setup 1, (B) Similar absence of trend, as observed in exp. setup 2.

exp. setup 1. It can be observed that for these activities, the probabilities rapidly increase as the forecasting horizon closes to zero (t closes to t_{next}), especially for the first two car related activities. This is because, during this time range an actor typically walks upto the vehicle and as the actor gets closer to the vehicle (t closes to t_{next}), the model gets more confident that the person is going to interact with the car and hence one of these activities is going to be performed. The last observed activity label and the spatio-temporal context further refines the forecasting probabilities to put preference to a particular activity label.

However, this increasing trend in forecasting probability is largely activity specific as for some of the activities in the dataset, there may not be any tightly associated scene context variable. Thus, even large changes in computed scene context variables minimally affect the forecasting probabilities when these activities would occur in immediate future. For VIRAT, some examples of such activities are Running, Gesturing etc. As seen in Fig. 5(A), there is no visible trend in the time evolution of forecasting probabilities for these activities. Again, for activities such as ‘unloading object from a vehicle’, the relevant scene context variables (e.g. distance from a car, overlap with a car bounding box etc.) remain largely constant, thereby resulting in uniform average probabilities through the forecasting time range (Fig. 5(A)).

5.5 Experiment Set 2

The probabilities and accuracies for forecasting future activities are affected by the accuracy of the recognition module used for already observed activities. Therefore, to factor out the effect of the errors in the observed activity recognition module on the forecasting results, we repeat the same experiments as in exp. set 1 with only the classifier replaced by a perfect recognition scheme. As

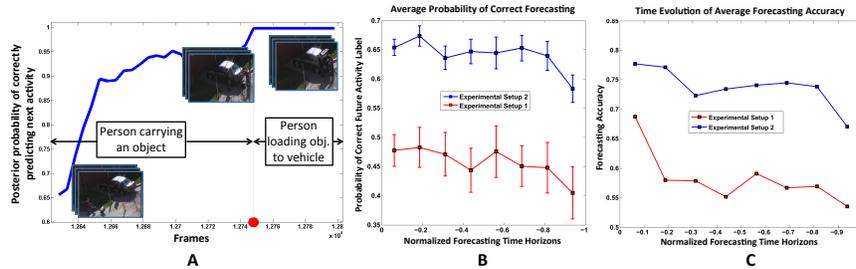


Fig. 6. (A) An example showing how the posterior probability of forecasting increases with time and stabilizes once the next observation is obtained. (B) Time evolution of forecasting probability averaged over all activity classes in exp. setups 1 and 2. (C) Average accuracy of forecasting correct labels for the immediately next unobserved activities in both the experimental setups.

we observe an activity, we retrieve its ground truth label and set the activity recognition probability for that particular activity at a very high value, and close to zero for the rest. The evolution of forecasting probabilities with normalized time horizon is shown in Fig. 4(B) for the activities that show an increasing trend in forecasting probabilities and Fig. 5(B) for the activities without any apparent trend in the time evolution of probabilities. The figures are visually similar to those for the same activities in exp. setup 1. However, the average forecasting probabilities for most of the activities are typically higher than that in the classifier based recognition case (set 1).

An example showing the increasing forecasting probability for the next unobserved activity is presented in Fig. 6(A) (in exp. setup 2). In a video segment, an actor is observed to ‘carry an object’ and the *unobserved* future activity would be ‘person loading object to a vehicle’. A parked car is detected in the scene and the posterior probability of the next activity being labeled as ‘person loading obj.’ rapidly increases as the person walks straight towards the car and gets closer to it. Fluctuation in the probability is seen due to occlusion of the detected object on person. The posterior probability gets close to 1 just before the start of the next activity. Once the next observation is obtained, the posterior represents recognition probability and remains constant for the rest of the video. The time evolutions of forecasting probability averaged over all activity classes for both exp. setup 1 and 2 are shown in Fig. 6(B) and in both the cases they show an overall increasing trend. As expected, the average probabilities in exp. setup 2 is higher than that in exp. setup 1. Similar trends are also observed in Fig. 6(C), which shows the time evolution of average forecasting accuracies. An increasing trend very similar to that of forecasting probabilities is observed.

As the proposed method is capable of forecasting activities deeper into the future beyond the immediately next activity, we also compute the time evolution of forecasting probabilities and accuracies for activities one step ahead (the ‘next-to-next’ activities). Time evolution of forecasting probabilities for the next-to-next activities are given in the supplementary. The overall forecasting accuracies

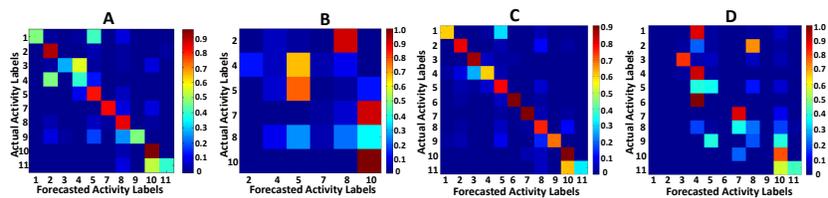


Fig. 7. Confusion matrices showing the overall forecasting accuracies obtained for each class of activity. (A-B) Accuracies for forecasting activities in immediate future and one step ahead (next-to-next) in experimental setup 1, (C-D) Accuracies for next and next-to-next activities respectively in experimental setup 2.

of all next and next-to-next activities in exp. setup 1 are shown in Fig. 7(A-B). The accuracies for most activities happening in immediate future is high. As we predict activities deeper into future, the accuracies tend to go down, as evidenced by Fig. 7(B). Similar trends are seen for activities in exp. setup 2 (Fig. 7(C-D)). Please note that, as there is no baseline activity classifier to train in exp. setup 2 and we need only 20% of the data for training the scene context classifier, we have the entire remaining dataset for testing and that is why results for all 11 activities could be investigated. With an ideal activity recognition scheme (setup 2), the expected improvements in forecasting accuracy can immediately be evidenced by the confusion matrices shown in Fig. 7(C-D).

The effect and importance of individual components of the proposed model can be understood by analyzing the results above. Fig. 4 and Fig. 6 show improvement in forecasting probability and accuracy as the scene context based node potential changes from uniform (no scene context) to a more definitive distribution. Fig. 6(B,C) and Fig. 7 also show how scene context improves forecasting over simple temporal activity context as the next unobserved activity nodes benefit from scene context, but the next-to-next activities do not.

6 Conclusion

In this paper, we have presented a novel approach towards the problem of forecasting future activities in long duration continuous videos. We have shown that the forecasting problem can be posed as a graph inference problem on a MRF where individual activities in a sequence are nodes on the graph. The method combines the spatio-temporal inter activity context and inter-relationship between actors' tracks and detected key points and objects in the scene with a standard activity recognition classifier to forecast activities that are yet to be observed. We show detailed experimental results on the challenging VIRAT[10] dataset and achieve meaningful and encouraging results. Future work would include anomalous activity detection using the method proposed in this paper.

Acknowledgement This work is partially supported by the National Science Foundation grant IIS-1316934.

References

1. Schudt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: International Conference on Pattern Recognition. Volume 3. (2004) 32–36
2. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. *Visual Surveillance and Performance Evaluation of Tracking and Surveillance* **0** (2005) 65–72
3. Ryoo, M.S., Aggarwal, J.K.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: International Conference on Computer Vision. (2009)
4. Nayak, N.M., Zhu, Y., Roy-Chowdhury, A.K.: Exploiting spatio-temporal scene structure for wide-area activity analysis in unconstrained environments. *IEEE Transactions on Information Forensics and Security* **8** (2013) 1610–1619
5. Choi, W., Shahid, K., Savarese, S.: Learning context for collective activity recognition. In: *Computer Vision and Pattern Recognition*. (2011)
6. Ryoo, M.S.: Human activity prediction: Early recognition of ongoing activities from streaming videos. In: International Conference on Computer Vision. (2011) 1036–1043
7. Yu, G., Yuan, J., Liu, Z.: Predicting human activities using spatio-temporal structure of interest points. In: *ACM Multimedia, ACM* (2012) 1049–1052
8. Hoai, M., De la Torre, F.: Max-margin early event detectors. In: *Computer Vision and Pattern Recognition*. (2012)
9. Kitani, K.M., Ziebart, B.D., Bagnell, J.A., Hebert, M.: Activity forecasting. In: *European Conference on Computer Vision*. Volume 7575 of *Lecture Notes in Computer Science*. (2012) 201–214
10. Oh, S., Hoogs, A., Perera, A.G.A., Cuntoor, N.P., Chen, C.C., Lee, J.T., Mukherjee, S., Aggarwal, J.K., Lee, H., Davis, L.S., Swears, E., Wang, X., Ji, Q., Reddy, K.K., Shah, M., Vondrick, C., Pirsiavash, H., Ramanan, D., Yuen, J., Torralba, A., Song, B., Fong, A., Chowdhury, A.K.R., Desai, M.: A large-scale benchmark dataset for event recognition in surveillance video. In: *CVPR, IEEE* (2011) 3153–3160
11. Poppe, R.: A survey on vision-based human action recognition. *Image and Vision Computing* **28** (2010) 976–990
12. Gaur, U., Zhu, Y., Song, B., Roy-Chowdhury, A.K.: A “string of feature graphs” model for recognition of complex activities in natural videos. In: International Conference on Computer Vision. (2011)
13. Yao, B., Feifei, L.: Modeling mutual context of object and human pose in human object interaction activities. In: *Computer Vision and Pattern Recognition*. (2010)
14. Lan, T., Wang, Y., Robinovitch, S.N., Mori, G.: Discriminative latent models for recognizing contextual group activities. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence*. (2012)
15. Gupta, A., Srinivasan, P., Shi, J., Davis, L.S.: Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In: *Computer Vision and Pattern Recognition*. (2009)
16. Si, Z., Pei, M., Yao, B., Zhu, S.: Unsupervised learning of event and-or grammar and semantics from video. In: International Conference on Computer Vision. (2011)
17. Zhu, Y., Nayak, N.M., Roy-Chowdhury, A.K.: Context-aware modeling and recognition of activities in video. In: *Computer Vision and Pattern Recognition*. (2013)

18. Zhu, Y., Nanyak, N.M., Roy-Chowdhury, A.K.: Vector field analysis for multi-object behavior modeling. In: *IEEE Journal of Selected Topics in Signal Processing (J-STSP)*. Volume 7. (2013) 91–101
19. N. Nayak, Y. Zhu, A.R.C.: Exploiting spatio-temporal scene structure for wide-area activity analysis in unconstrained environments. In: *IEEE Trans. on Information Forensics and Security*. Volume 8. (2013)
20. Benmokhtar, R., Laptev, I.: INRIA-WILLOW at TRECVID2010: Surveillance Event Detection. In: *TRECVID*. (2010)
21. Morariu, V.I., Davis, L.S.: Multi-agent event recognition in structured scenarios. In: *Computer Vision and Pattern Recognition*. (2011)
22. Tang, K., Fei-Fei, L., Koller, D.: Learning latent temporal structure for complex event detection. In: *Computer Vision and Pattern Recognition*. (2012)
23. Kschischang, F.R., Frey, B.J., Loeliger, H.A.: Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory* **47** (1998) 498–519
24. Zivkovic, Z.: Improved adaptive gaussian mixture model for background subtraction. In: *International Conference on Pattern Recognition*. (2004) 28–31
25. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32** (2010) 1627–1645
26. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: *Computer Vision and Pattern Recognition*. (2005) 886–893
27. Song, B., Jeng, T.Y., Staudt, E., Roy-Chowdhury, A.K.: A stochastic graph evolution framework for robust multi-target tracking. In: *European Conference on Computer Vision*. (2010) 605–619
28. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: *International Conference on Computer Vision*. (2005) 1395–1402