Context-Aware Activity Recognition and Anomaly Detection in Video

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

***Abstract*—In this paper, we propose a mathematical framework to jointly model related activities with both motion and context information for activity recognition and anomaly detection. This is motivated from observations that activities related in space and time rarely occur independently and can serve as context for each other. The spatial and temporal distribution of different activities provides useful cues for the understanding of these activities. We denote the activities occurring with high frequencies in the database as normal activities. Given training data which contains labeled normal activities, our model aims to automatically capture frequent motion and context patterns for each activity class, as well as each pair of classes, from sets of predefined patterns during the learning process. Then, the learned model is used to generate globally optimum labels for activities in the testing videos. We show how to learn the model parameters via an unconstrained convex optimization problem and how to predict the correct labels for a testing instance consisting of multiple activities. The learned model and generated labels are used to detect anomalies whose motion and context patterns deviate from the learned patterns. We show promising results on the VIRAT Ground Dataset that demonstrates the benefit of joint modeling and recognition of activities in a wide-area scene and the effectiveness of the proposed method in anomaly detection.**

***Index Terms*—Context-aware activity recognition, context-aware anomaly detection, structural model.**

I. Introduction

VIDEO surveillance systems monitor people's activities and generate alerts when anomalous activities are detected. Usually, samples of anomalous activities are rare. Given a set of normal samples, the system is trained to learn frequent patterns of normal activities using methods of activity recognition. Activities whose patterns deviate from the learned frequent patterns are detected as anomalies.

Most methods developed in the literature on activity recognition have concentrated on analyzing individual motion patterns of activities as evidenced by popular activity datasets like KTH [1] and Weizmann [2]. These methods model activities individually and aim to learn discriminative patterns for each activity class. However, activities in natural scenes rarely happen independently as shown in Fig. 1(i). In the same local spatio-temporal region, the activity of a person closing a vehicle trunk often happens before, and not after, the activity of the person getting into the vehicle. The interdependence between activity classes provides important cues for activity recognition. Jointly modeling and recognizing related activities in space and time can improve recognition accuracy. This, in turn, will help detect anomalous activities better.

It has been demonstrated in [3] that context is significant in human visual systems. As there is no formal definition of context in video analysis, we consider all the detected objects and motion regions in a local neighborhood as providing contextual information about each other. Human-object interaction has been frequently used as context in many past works [4], [5]. An

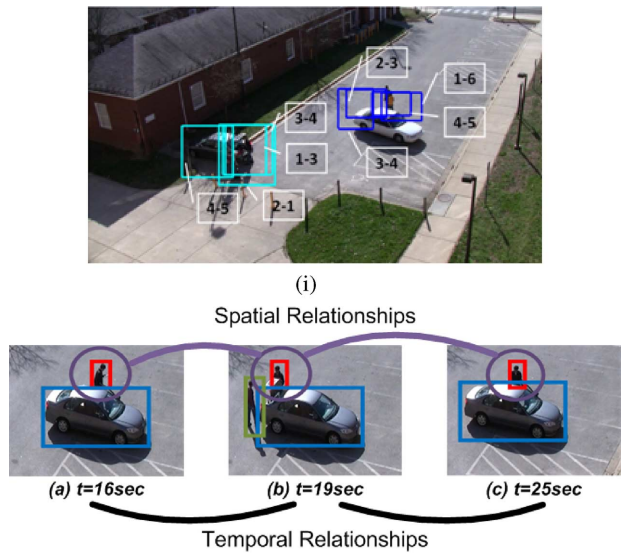


Fig. 1. (i) Example images from a wide area video (interesting activities happening within about 2 minutes are shown). Activities in the same color in each video happen in the same local spatio-temporal region. Activity classes are listed in Fig. 1 in the supplementary material. (a). For the indices, the first index denotes the temporal order of the activity in the region, while the second number denotes the activity class, e.g., 1–6 means the activity belongs to class 6 and is the first activity that happens in this video volume. (ii) Example of context in activity recognition. A person of interest is located by red bounding box, surrounding objects are located by bounding boxes of other colors, and the circles in purple indicate the motion regions of the activities. The existence of a car nearby gives information about what the person of interest is doing, and the relative position of the person of interest and the car may denote that activities in (a) and (c) are very different from activity in (b). However, it is hard to tell if the person in (a) and (c) is getting out of the vehicle or getting into the vehicle. If we knew that these activities occured around the same vehicle, we can infer with high probability that in (a) the person is getting out of the vehicle and in (c) the person is getting into the vehicle.

Manuscript received August 07, 2012; revised November 16, 2012; accepted December 07, 2012. Date of publication December 20, 2012; date of current version January 22, 2013. This work was supported in part by National Science Foundation grant IIS-0712253, and Defense Advanced Research Projects Agency STTR award W31P4Q-11-C0042 through Mayachitra, Inc. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Venkatesh Saligrama.

Y. Zhu is with the Department of Electrical Engineering, University of California, Riverside, CA 92521 USA (e-mail: yzhu010@ucr.edu).

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

A. Roy-Chowdhury is with the Department of Electrical Engineering, University of California, Riverside, CA 92521 (e-mail: amitrc@ee.ucr.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2012.2234722

example of the relationships between various activities is shown in Fig. 1(ii). Harnessing such spatial and temporal relationships could be very beneficial for activity recognition. Motivated by the above, we propose to learn the motion and context patterns of normal activities. These learned motion and context patterns are used for classifying between normal activities. Activities whose motion and context patterns deviate from the normal motion and context patterns are considered as anomalies.

A. Overview and Main Contributions

The main contribution of this work is to show how context can be exploited for activity recognition and anomaly detection in video. We focus on the joint modeling and recognition of activities in videos of a wide scene, using both motion and context information. We build upon existing well-known low-level motion and image feature descriptors and the spatio-temporal context representations that, when combined together, provide a powerful framework to model activities in continuous videos.

Activities within a spatio-temporal distance threshold are considered related to one another and are grouped into the same activity set as shown in Fig. 1(i). Activities in each set are jointly modeled and recognized. Given a set of related activities, motion and context features are extracted. A label vector, whose elements are the class labels of individual activities, is to be assigned to it. The problem now is how to measure the compatibility between its features and a candidate label of the activity set. A potential function is introduced for this purpose. The parameter of this function, which is called the joint weight vector, captures the valid motion and context patterns. The candidate label with the highest potential score is assigned to the activity set as its label vector.

In the learning process, the parameter estimation is formulated as a large-margin problem, which tries to maximize the margins around the decision plane which separates the negative and positive instances. We show how this problem can be modified to be an unconstrained convex optimization problem. Next, the modified bundle method in [6] is used to solve the optimization problem. This method iteratively searches for the increasingly tight upper and lower bounds of the objective function till convergence is reached. The learning process automatically learns and weights motion and context patterns for each activity class and each pair of classes from sets of predefined patterns. Given a test video, spatio-temporal locations of activities are detected using motion segmentation and the surrounding regions identified. Activities in each region are jointly recognized using the learned potential function through a greedy search algorithm [7], which greatly decreases the computational complexity of the inference process with negligible reduction to recognition accuracy.

Thus, our method explicitly models the spatial and temporal relationships of activities and captures useful spatio-temporal patterns for each pair of interesting activity classes during the learning process. It integrates motion features and various context features into a unified model. With the learned pattern parameters, normality factors are introduced to measure the normalcy of activities based on their motion and context features. Activities with one or more normality factors lower than the predefined thresholds (which can be learned a priori) are considered as anomalies.

II. Related Work

In this section, we only review related works on activity recognition exploring context and anomalous activity detection. For a more comprehensive review in activity recognition, please refer to recent surveys like [8].

Many existing works exploring context focus on spatio-temporal relationships of features [9], [10], interactions of objects and actions/activities [5], [11], [12], environmental conditions such as spatial locations of certain activities in the scene [13], and temporal relationships of activities [4], [14]. There has been a lot of recent interest in exploiting context-sensitive information for activity recognition. One of the approaches that has been popular for context modeling is the use of AND-OR graphs which can provide a semantic representation of the scene [15], [16]. They have been employed in applications like sports videos [15] and office scenes [16]. Results on recognizing atomic actions in such structured scenarios have also been provided. However, the applicability of AND-OR graphs for more unstructured scenarios has not been demonstrated.

Methods on the detection of anomalous activities can be divided into two categories: low-level anomaly detection and high-level anomaly detection. Approaches on low-level anomaly detection identify local spatio-temporal regions that probably contain anomalous patterns of low-level features, before high-level analysis such as object tracking and activity classification is done [17]–[19]. Some other works of this category represent activities as local spatio-temporal regions. Abnormal activities are discovered by modeling the dominant motion patterns of these local regions. In [20], a probabilistic framework was developed to detect local anomalies that have infrequent patterns with respect to their neighbors. In [21], the authors proposed an online algorithm to incrementally learn a sparse dictionary of motion features of normal instances. Spatial-temporal blocks whose motion features can not be reconstructed sparsely from the learned dictionary were identified as anomalies.

Several works have looked at the problem on high-level anomaly detection. These approaches usually identify semantically meaningful activities while detecting anomalies. In [22] activities were represented as bags of event n-grams. Disjunctive sub-classes of an activity class were discovered automatically. An information-theoretic method was used to explain the detected anomalies. In [23], [24] activities were represented by suffix trees over multiple temporal scales which efficiently extract the structure of activities by analyzing their constituent sub-events. An linear-time algorithm was proposed to detect anomalous subsequences of activities which are inconsistent with the learned Suffix Trees. In [25], attribute grammars were built to describe constraints on attributes and syntactic structure of normal events. Events which do not follow the syntax of the learned grammars or whose attributes do not satisfy these constraints were detected as anomalies. Many other works [26], [27] were based on trajectories of moving objects in videos. Dominant trajectory clusters were identified and modeled as normal while trajectories which do not fit into the learned models were detected as anomalies. These approaches work well in identifying global anomalous activities whose characteristics can be determined by underlying object trajectories.

III. Model Formulation for Context-Aware Activity Representation

In this section, a structural activity model that integrates motion features with various context features within and across activities is proposed to jointly model related activities in videos.

A. Preprocessing

Given a video, background subtraction [28] is used to locate the motion regions. Moving persons and vehicles are identified using publicly available software [29]. Bounding boxes of moving persons are obtained and used as the initialization of the tracking method developed in [30] to obtain local trajectories. We use the Spatio-temporal interest point (STIP) detector developed in [31] to generate concatenated histogram of oriented gradients (HOG) and histogram of optical flow (HOF) features for motion regions surrounding the bounding boxes of the moving objects. Thus, STIPs generated by noise, such as slight tree shaking, camera jitter and motion of shadows, are avoided.

An activity region is defined as a 3D video region with a start time and an end time to be labeled. To locate the activity regions, motion regions are first divided into temporal bins. Sliding windows of different sizes are applied to the motion regions. In the experiment, bag-of-words combined with multi-class support vector machine (BOW+SVM) [32] are used to label each window as one of the normal activity classes. Then, weighted average smoothing is applied to obtain the label of each temporal bin. Objects that occur in the images that overlap with motion regions are detected. These image features will be used for the development of the context features within activities.

B. Motion and Context Feature Descriptor

The following definitions will be used for the development of feature descriptors. An agent is a moving person of interest. Motion region of an activity at frame i denotes a circular region surrounding the objects of interest at the i^{th} frame of the activity. Activity region is the smallest rectangle region that encapsulates all the motion regions over all frames of the activity.

1) Intra-Activity Motion Feature: Features of an activity that encode the motion information extracted from low-level motion features are defined as intra-activity motion features. Assume there are M classes of normal activities (a class of normal background activities may be introduced to handle other normal activities that do not belong to the M classes). Motion features such as STIP histograms are often high-dimensional. We train a multi-class classifier, which is called as the *baseline* classifier, to generate the normalized confidence scores $x_1, \ldots, x_M$, where $\sum_{i=1}^{M} x_i = 1$, of classifying an activity as belonging to activity classes $1, \ldots, M$, and thus transforming the high-dimensional motion features to a low-dimensional space. Then, a score histogram $x = [x_1, \ldots, x_M]$ is developed as the intra-activity motion feature of an activity. In the experiments, we use BOW+SVM [32] and SFG in [10] as the baseline classifier.

2) Intra-Activity Context Feature: Features that capture the context information about the agents and relationships between the agents and the the interacting objects (e.g., the object classes, interactions between agents and their surroundings) are defined as intra-activity context features. We use common sense knowledge about normal activity classes to guide the building of the context features of activities. We define a set G of context attributes related to the scene and involved objects in the normal activities. G consists of N_G subsets of attributes that are exclusively related to certain image-level features. Since we work on the VIRAT dataset with individual person activities and person-object interactions, we use subsets of attributes for the development of intra-activity context features in Fig. 2.

Subset	Associated Attributes
G_1	moving object is a person; moving object is a vehicle; moving object is of other kind.
G_2	the agent is at the body of the interacting vehicle; the agent is at the rear/head of the interacting vehicle; the agent is far away from the vehicles.
G_3	the agent disappears at the entrance of a facility; the agent appears at the exit of a facility; none of the two.
G_4	velocity of the agent (in pixels) is larger than a predefined threshold; velocity of object of interest is smaller than a predefine threshold.
G_5	the activity occurs at parking areas; the activity occurs at other areas.
G_6	an object is detected on the agent; no object is detected on the agent.

Fig. 2. Subsets of context attributes used for the development of intra-activity context features.

For a given activity, whether the above attributes are true or not are determined from image-level detection results (e.g., detected objects). Let n_{G_i} be the number of attributes in subset G_i for $i = 1, \ldots, N_G$. For frame n of an activity, we obtain $g_i(n) = I(G_i, n)$, where $I(\cdot)$ is the indicator vector of size $n_{G_i} \times 1$ with element 1 if the corresponding attribute is true for frame n and 0 otherwise. $g_i(n)$ is then normalized so that its elements sum to 1. Note that the attribute in g_2 is determined once for each activity. For instance, the agent (person) disappears at the entrance of a facility is true, if opening entrance door is detected around the agent and the detector could not find a good match of the agent after the door is closed again. Fig. 3 shows examples of $g_i(n)$ for different activities.

Let $g_i = (1/N_f) \sum_{n=1}^{N_f} g_i(n)$, where N_f is the total number of frames associated with the activity. The $\sum_{i=1}^{N_G} n_{G_i}$-bin histogram $g = (1/N_G)[g_1 \oplus \cdots \oplus g_{N_G}]$ is the intra-activity context feature vector of the activity, where $\oplus$ denotes the vector concatenation operator.

3) Inter-Activity Context Feature: Features that capture the relative spatial and temporal relationships of activities are defined as inter-activity context feature. We develop normalized histograms sc and tc that bin the spatial and temporal relationships (defined below) of two activities into pre-determined sets SC and TC, respectively.

Spatial Context: Let $O_{a_i}(n)$ and $R_{a_i}(n)$ denote the center and radius of the motion region of activity a_i at its n^{th} frame and O_{a_j} and R_{a_j} denote the center and radius of the activity region of activity a_j. Let

$$r_s(a_i(n), a_j) = \frac{d(O_{a_i}(n), O_{a_j})}{R_{a_i}(n) + R_{a_j}}, \tag{1}$$

where $d(\cdot)$ denotes the Euclidean distance. $r_s(a_i(n), a_j)$ captures the scaled spatial overlap of the activity a_j with the n^{th} frame of a_i. The attributes in set SC are obtained by quantizing and grouping $r_s(a_i(n), a_j)$ into a predefined

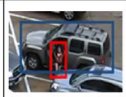
Activity	person loading	person unloading	opening trunk	closing trunk
Example Image				
$g_1(n)$	$[\frac{1}{2}\ \ 0\ \ \frac{1}{2}]$	$[1\ \ 0\ \ 0]$	$[\frac{1}{2}\ \ \frac{1}{2}\ \ 0]$	$[\frac{1}{2}\ \ \frac{1}{2}\ \ 0]$
$g_2(n)$	$[0\ \ 1\ \ 0]$	$[0\ \ 1\ \ 0]$	$[0\ \ 1\ \ 0]$	$[0\ \ 1\ \ 0]$
$g_4(n)$	$[0\ \ 1]$	$[0\ \ 1]$	$[0\ \ 1]$	$[0\ \ 1]$
$g_5(n)$	$[1\ \ 0]$	$[1\ \ 0]$	$[1\ \ 0]$	$[1\ \ 0]$
$g_6(n)$	$[1\ \ 0]$	$[0\ \ 1]$	$[0\ \ 1]$	$[1\ \ 0]$
Activity	getting into vehicle	getting out of vehicle	gesturing	carrying object
Example Image				
$g_1(n)$	$[1\ \ 0\ \ 0]$	$[1\ \ 0\ \ 0]$	$[1\ \ 0\ \ 0]$	$[\frac{1}{2}\ \ 0\ \ \frac{1}{2}]$
$g_2(n)$	$[1\ \ 0\ \ 0]$	$[1\ \ 0\ \ 0]$	$[0\ \ 0\ \ 1]$	$[0\ \ 0\ \ 1]$
$g_4(n)$	$[0\ \ 1]$	$[0\ \ 1]$	$[0\ \ 1]$	$[0\ \ 1]$
$g_5(n)$	$[1\ \ 0]$	$[1\ \ 0]$	$[1\ \ 0]$	$[1\ \ 0]$
$g_6(n)$	$[0\ \ 1]$	$[0\ \ 1]$	$[0\ \ 1]$	$[1\ \ 0]$

Fig. 3. Examples of detected intra-activity context features. The example images are shown with detected high-level image features. Object in red bounding box is a moving person; object in blue bounding box is a static vehicle; object in orange bounding box is a moving object of other kind; object in black bounding box is a bag/box on the agent.

(a) (b)

Fig. 4. (a) The image shows one example of inter-activity spatial relationship. The red circle indicates the motion region of a_i at this frame while the purple rectangle indicates the activity region of a_j. Assume SC is defined by quantizing and grouping $r_s(n)$ into three bins: $r_s(n) \leq 0.5$ (a_i and a_j is at the same spatial position at the n^{th} frame of a_i), $0.5 < r_s(n) < 1.5$ (a_i is near a_j at the n^{th} frame of a_i) and $r_s(n) \geq 1.5$ (a_i is far away from a_j at the n^{th} frame of a_i). In the image, $r_s(n) > 1.5$, so, $sc_{ij}(n) = [0\ \ 0\ \ 1]$. (b) The image shows one example of inter-activity temporal relationship. The n^{th} frame of a_i, denoted by $a_i(n)$, occurs before a_j. So, $t_{ij}(n) = [1\ \ 0\ \ 0]$. Note that t indicates the time of the corresponding frames(s) in the video.

number n_{SC} of bins as shown in Fig. 4(a). Then, the spatial relationship $sc_{ij}(n)$ of a_i and a_j at the n^{th} frame can be calculated as $\mathbf{I}(SC, n)$, where $\mathbf{I}(SC, n)$ is a $n_{SC} \times 1$ vector with 1 for an element if $r_s(a_i(n), a_j)$ belongs to the corresponding bin in SC, and 0 otherwise. The n_{SC}-bin histogram $sc_{a_i,a_j} = (1/N_f)\sum_{n=1}^{N_f} sc_{ij}(n)$ is the inter-activity spatial feature of activity a_i and a_j.

Temporal Context: The temporal context feature is defined by the following temporal relationships: n^{th} frame of a_i is before a_j, n^{th} frame of a_i is during a_j and n^{th} frame of a_i is after a_j. $tc_{ij}(n)$ is the temporal relationship of a_i and a_j at the n^{th} frame of a_i as shown in Fig. 4(b). The 3-bin histogram $tc = (1/N_f)\sum_{n=1}^{N_f} tc_{ij}(n)$ is the inter-activity temporal context feature of activity a_i with respect to activity a_j.

C. Structural Activity Model

Suppose we are interested in M activity classes. Activity set $a = \{a_i : i = 1, \ldots, N\}$ is associated with a label vector $y = \{y_i : i = 1, \ldots, N\}$, where $y_i \in \{1, \ldots, M\}$ is the label of a_i. We model the activity set by the combination of motion features of individual activities and various context features discussed above. A potential function that measures the compatibility between features of a and label y is defined as $F(a, y)$:

$$\mathbf{F}(a, y) = \sum_{i=1}^{N} \omega_{x,y_i}^T x_i + \sum_{i=1}^{N} \omega_{g,y_i}^T g_i + \sum_{i,j=1,i\neq j}^{N} \omega_{sc,(y_i,y_j)}^T sc_{ij} + \sum_{i,j=1,i\neq j}^{N} \omega_{tc,(y_i,y_j)}^T tc_{ij}, \quad (2)$$

where $x_i \in R^{D_x}$ and $g_i \in R^{D_g}$ are the motion feature and intra-activity context feature of instance a_i, D_x and D_g are the dimension of x_i and g_i respectively. $\omega_{x,y_i} \in R^{D_x}$ and $\omega_{g,y_i} \in R^{D_g}$ are the weights that capture the valid motion and intra-activity context patterns of activity class y_i. $sc_{ij} \in R^{D_{sc}}$ and $tc_{ij} \in R^{D_{tc}}$ are the inter-activity context features associated a_i and a_j. D_{sc} and D_{tc} are the dimension of sc_{ij} and tc_{ij} respectively. $\omega_{sc,(y_i,y_j)} \in R^{D_{sc}}$ and $\omega_{tc,(y_i,y_j)} \in R^{D_{tc}}$ are the weights that capture the valid spatial and temporal relationships of activity classes y_i and y_j. In general, dimensions of the same kind of feature can be different for each activity class/class pairs.

In order to form a linear function with a single parameter, we rewrite (2) as:

$$\mathbf{F}(a, y) = \omega_x^T \sum_{i=1}^{N} \varphi(x_i, y_i) + \omega_g^T \sum_{i=1}^{N} \vartheta(g_i, y_i) + \omega_{sc}^T \sum_{i,j=1,i\neq j}^{N} \psi(sc_{ij}, y_i, y_j) + \omega_{tc}^T \sum_{i,j=1,i\neq j}^{N} \phi(tc_{ij}, y_i, y_j), \quad (3)$$

where ω_x, ω_g, ω_{sc} and ω_{tc} are weight vectors defined as

$$\omega_x = [\omega_{x,1}^T \ \ \omega_{x,2}^T \ \ \cdots \ \ \omega_{x,M}^T]^T,$$
$$\omega_g = [\omega_{g,1}^T \ \ \omega_{g,2}^T \ \ \cdots \ \ \omega_{g,M}^T]^T,$$
$$\omega_{sc} = [\omega_{sc,(1,1)}^T \ \ \cdots \ \ \omega_{sc,(1,M)}^T \ \ \cdots \ \ \omega_{sc,(M,M)}^T]^T,$$
$$\omega_{tc} = [\omega_{tc,(1,1)}^T \ \ \cdots \ \ \omega_{tc,(1,M)}^T \ \ \cdots \ \ \omega_{tc,(M,M)}^T]^T,$$

and $\varphi(x_i, y_i)$ and $\vartheta(g_i, y_i)$ have non-zero entries at the position corresponding to class index y_i. $\psi(sc_{ij}, y_i, y_j)$ and $\phi(st_{ij}, y_i, y_j)$ have none-zero entries at the position corresponding to class pair (y_i, y_j).

Define the joint weight vector ω and joint feature vector $\Gamma(a, y)$ as

$$\omega = \begin{bmatrix} \omega_x \\ \omega_g \\ \omega_{sc} \\ \omega_{st} \end{bmatrix}, \Gamma(a, y) = \begin{bmatrix} \sum_i \varphi(x_i, y_i) \\ \sum_i \vartheta(g_i, y_i) \\ \sum_{i,j,i\neq j} \psi(sc_{ij}, y_i, y_j) \\ \sum_{i,j,i\neq j} \phi(tc_{ij}, y_i, y_j) \end{bmatrix},$$

where $i, j = 1, \ldots, N$. Then, the optimum label y^{opt} of x is obtained as

$$y^{opt} = \arg\max_y (\omega^T \Gamma(a, y)). \quad (4)$$

IV. Model Learning and Inference

A. Learning Model Parameters

We will now describe our method for learning the model parameters from training sets. Suppose there are P collections of activities in the training videos. Let the training set be $(A_T, Y_T) = (a_T(1), y_T(1)), \ldots, (a_T(P), y_T(P))$, where each $a_T(i)$ is an activity set and $y_T(i)$ is its label vector. Suppose there are $N_T(i)$ elements in $a_T(i)$. We use the following loss function to measure the correctness of labeling instance $a_T(i)$ with the candidate label $\widehat{y_T(i)}$:

$$\Delta\left(y_T(i), \widehat{y_T(i)}\right) = \sum_{j=1}^{N_T(i)} \Delta\left(y_T(i,j), \widehat{y_T(i,j)}\right),$$

$$\Delta\left(y_T(i,j), \widehat{y_T(i,j)}\right) = \begin{cases} 1 & y_T(i,j) \neq \widehat{y_T(i,j)} \\ 0 & y_T(i,j) = \widehat{y_T(i,j)} \end{cases}.$$

The model learning problem is formulated as an unconstrained convex optimization problem (derivation is shown in Section III in the supplementary material):

$$\omega^* = \arg\min_{\omega} f(\omega) = \arg\min_{\omega} \frac{1}{2}\omega^T\omega + \Lambda(\omega),$$

$$\text{where } \Lambda(\omega) = C\sum_{i=1}^{P} \max(0, \Omega_\omega(i)),$$

$$\Omega_\omega(i) = \max_{\widehat{y_T(i)}} \Bigg(\Delta\left(y_T(i), \widehat{y_T(i)}\right) + \omega^T\Big(\Gamma\left(a_T(i), \widehat{y_T(i)}\right) - \Gamma(a_T(i), y_T(i))\Big)\Bigg). \tag{5}$$

1) Optimization Algorithm: The problem in (5) can be solved by the modified bundle method in [6]. It iteratively searches for the increasingly tight quadratic upper and lower cutting planes of the objective function until the gap between the two bounds reaches a predefined threshold. A cutting plane of a convex function is defined by its first-order Taylor approximation and can be calculated as [6]

$$g_\omega = \omega^T\partial_\omega\Lambda(\omega) + b_\omega, \quad b_\omega = \Lambda(\omega) - \omega^T\partial_\omega\Lambda(\omega).$$

The algorithm is effective because of its very high convergence rate [6]. The bundle method specified for problem (5) is summarized in Algorithm 1:

Algorithm 1 Learning the model parameter in 5 through bundle method **[6]**.

Input: $S = ((a_T(1), y_T(1)), \ldots, (a_T(P), y_T(P))), C, \epsilon$
Output: Optimum model parameter ω
1) Initialize ω as ω_0 using empirical values, $\mathcal{G}$(cutting plane set) $\leftarrow \varnothing$.
2) for $t = 0$ to ∞ do
3) for $i = 1, \ldots, P$ do
find the most violated label vector for each training instance using ω_t (the value of ω at the t^{th} iteration);
4) end for
5) find the cutting plane g_{ω_t} of $\Lambda(\omega)$ at ω_t:
$g_{\omega_t} = \omega^T\partial_\omega\Lambda(\omega_t) + b_{\omega_t}$,
where $b_{\omega_t} = \Lambda(\omega_t) - {\omega_t}^T\partial_\omega\Lambda(\omega_t)$;
6) $\mathcal{G} \leftarrow \mathcal{G} \cup g_{\omega_t}(\omega)$;
7) update ω:

$$\omega_{t+1} = \arg\min_{\omega} f_{\omega_t}(\omega),$$
$$f_{t+1}(\omega) = f_{\omega_t}(\omega_{t+1}),$$
$$f_{\omega_t}(\omega) = \frac{1}{2}\omega^T\omega + \max(0, \max_{j=1,\ldots,t} g_{\omega_j}(\omega));$$

8) $gap_{t+1} = \min_{t' \leq t} f_{\omega_{t'+1}}(\omega_{t'+1}) - f_{\omega_t}(\omega_{t+1})$;
9) if $gap_{t+1} \leq \epsilon$, then return ω_{t+1};
10) end for

2) Efficient Implementation: We call the label vector $\widehat{y_T(i)}$ of the i^{th} instance that maximizes $\Omega_{\omega_t}(i)$ the most-violated label of the i^{th} instance in the $(t+1)^{th}$ iteration, if $\Omega_{\omega_t}(i) > 0$. If $\Omega_{\omega_t}(i) \leq 0$ for all $\widehat{y_T(i)}$, the constraints on i^{th} instance will not be violated in the $(t+1)^{th}$ iteration. In each iteration, we need to check and find the most-violated label for each instance. Finding the most violated label is NP hard (we need to enumerate all the possible label vectors). A greedy forward search is proposed in [33] to balance the computation efficiency and algorithm accuracy.

The computation of cutting planes requires knowledge of the most violated labels for all the training instances in each iteration. When the number of training instances is large, finding violated label for each instances in each iteration is inefficient. Like other online optimization techniques, we try to shrink the working space in order to improve efficiency. During the learning process, it is often revealed early that the constraints in (5) of certain instances are unlikely to be violated. Let us consider the history of violated instances over the last k iterations. If the constraints of an instance are not violated at each of the last k iterations, it is likely that they will not be violated before the optimum solution is reached. Considering that cutting planes do not depend on these instances in the subsequent iterations. Such instances are excluded from the working space and the solution space is stored. Since this heuristic can fail, the constraints for the eliminated instances are checked after convergence. If necessary, the optimization process is restarted from the solution stored previously. Also, to ensure the algorithm does not restart frequently, we maintain a minimum number $P_{\min}$ of training instances in the working space.

B. Inference

With the learned model parameter vector ω, we now describe how to identify the optimum label vector y_{test} for an input instance a_{test}. Suppose the testing instance has n activity-based segments $a_{test} = [a_{test}(1), \ldots, a_{test}(n)]$. The greedy forward search [33] is used to find the optimum labels of the targeted activities. We greedily instantiate the segment that, when labeled as an activity class of interest, can increase the value of compatibility function F by the largest amount. The algorithm stops when all the regions are labeled or labeling any other segments decreases the value of compatibility function F. Algorithm 2 gives the overview of the inference process. While this greedy

search algorithm cannot guarantee a globally optimum solution, in practice it works well to find good solutions as demonstrated in the experimental results. The papers [7], [33] give theoretical explanation of the effectiveness of the method in finding the optimum solution to the problems of the kind.

Algorithm 2 Greedy Search Algorithm

Input: Testing instance
Output: Interested activities A and label vector Y
1) initialize $(A, Y) \leftarrow \{\emptyset, \emptyset\}$ and $F = 0$.
2) repeat
$\Delta F(a_i, y_i)_{(a_i \notin A)} = F((A, Y) \cup (a_i, y_i)) - F((A, Y))$;
$(a_i, y_i)^{opt} = \arg\max_{(a_i \notin A)} \Delta F(a_i, y_i)$;
$(A, Y) \leftarrow (A, Y) \cup (a_i, y_i)^{opt}$;
3) end if all activities are labeled.

V. Anomaly Detection

For anomaly detection, we assume that we have instances of all the normal activities. Test instances whose patterns deviate from the learned model are anomalies. Once the activity label is assigned to a test instance, we focus on the analysis of whether this activity is anomalous.

For discriminative models such as the proposed one, the separating hyperplane between two classes can be obtained by subtracting the associated weight vectors as discussed in [34]. For normal instances belonging to a certain class, the distances to their associated separating hyperplanes are expected to follow certain distributions [34], which can be estimated through kernel density estimation from the training data. An instance with infrequent distances can be considered as anomaly. For this reason, four kinds of distances, which can be used to evaluate the normality of an activity and pair of activities, are developed based on the weight vectors learned for the proposed structural model.

With weight vectors $\omega_{x,i}$, $\omega_{x,j}$, $\omega_{g_k,i}$ and $\omega_{g_k,j}$ for $k \in \{1, \ldots, N_G\}$, $i, j \in \{1, \ldots, M\}$ and $i \neq j$, we define the unbiased motion hyperplane $HP_x(i,j)$ and intra-context hyperplane $HP_{g_k}(i,j)$ by their normal vectors as

$$\begin{aligned} HP_x(i,j) &= \omega_{x,i} - \omega_{x,j}, \\ HP_{g_k}(i,j) &= \omega_{g_k,i} - \omega_{g_k,j}, \end{aligned} \tag{6}$$

where an unbiased hyperplane means a hyperplane that passes through the origin. Thus, hyperplanes $HP_x(i,j)$ and $HP_{g_k}(i,j)$, translated along the directions of their normal vectors by a constant, can separate classes i and j based on motion and intra-context features respectively. With weight vectors $\omega_{sc,(i,j)}$, $\omega_{sc,(i',j')}$, $\omega_{tc,(i,j)}$ and $\omega_{tc,(i',j')}$ for $i, j, i', j' \in \{1, \ldots, M\}$ and $(i,j) \neq (i',j')$, define unbiased hyperplanes $HP_{sc}((i,j),(i',j'))$ and $HP_{tc}((i,j),(i',j'))$ by their normal vectors as

$$\begin{aligned} HP_{sc}((i,j),(i',j')) &= \omega_{sc,(i,j)} - \omega_{sc,(i',j')}, \\ HP_{tc}((i,j),(i',j')) &= \omega_{tc,(i,j)} - \omega_{tc,(i',j')}. \end{aligned} \tag{7}$$

Similarly, $HP_{sc}((i,j),(i',j'))$ and $HP_{tc}((i,j),(i',j'))$, translated along the directions of their normal vectors by a constant, can separate class pairs (i,j) and (i',j') based on inter-context spatial and temporal features respectively.

Consider an activity a with motion feature x_a, intra-activity context feature g_a and class label y_a^{opt} generated by the structural model. Define the distance of motion feature x_a to hyperplane $HP_x(y_a^{opt}, j)$, as $d_x^a(y_a^{opt}, j)$ and distance of intra-activity context feature $g_{a,k}$, for $k = 1, \ldots, N_G$ to hyperplane $HP_{g_k}(y_a^{opt}, j)$ as $d_{g_k}^a(y_a^{opt}, j)$, where $j \neq y_a^{opt}$, $j \in 1, \ldots, M$. These distances can be calculated as

$$\begin{aligned} d_x^a\left(y_a^{opt}, j\right) &= \frac{HP_x\left(y_a^{opt}, j\right)^T \cdot x_a}{norm\left(HP_x\left(y_a^{opt}, j\right)\right)}, \\ d_{g_k}^a\left(y_a^{opt}, j\right) &= \frac{HP_{g_k}\left(y_a^{opt}, j\right)^T \cdot g_{a,k}}{norm\left(HP_{g_k}\left(y_a^{opt}, j\right)\right)}, \end{aligned}$$

where $norm()$ is the Euclidean norm. Assume activity collection A with member activities $a_1, a_2, \ldots, a_N$ related to each other in space and time with class labels $Y^{opt} = [y_1^{opt}, y_2^{opt}, \ldots, y_N^{opt}]$. $sc_{a_1,a_2}, \ldots, sc_{a_{N-1},a_N}$ and $tc_{a_1,a_2}, \ldots, tc_{a_{N-1},a_N}$ are their inter-activity context features. The distance of inter-activity context feature sc_{a_i,a_j} to hyperplane $HP_{sc}((y_i^{opt}, y_j^{opt}),(i',j'))$ is defined as $d_{sc}^{a_i,a_j}((y_i^{opt}, y_j^{opt}),(i',j'))$ (denoted as $d_{sc}^{a_i,a_j}(i'j')$ for simplicity), and distance of tc_{a_i,a_j} to hyperplane $HP_{tc}((y_i^{opt}, y_j^{opt}),(i',j'))$ is defined as $d_{tc}^{a_i,a_j}((y_i^{opt}, y_j^{opt}),(i',j'))$ (denoted as $d_{tc}^{a_i,a_j}(i'j')$), where $(i',j') \neq (y_i^{opt}, y_j^{opt})$, $i', j' \in 1, \ldots, M$. These distances can be calculated as

$$\begin{aligned} d_{sc}^{a_i,a_j}(i'j') &= \frac{HP_{sc}\left(\left(y_i^{opt}, y_j^{opt}\right),(i',j')\right)^T \cdot sc_{a_i,a_j}}{norm\left(HP_{sc}\left(\left(y_i^{opt}, y_j^{opt}\right),(i',j')\right)\right)}, \\ d_{tc}^{a_i,a_j}(i'j') &= \frac{HP_{tc}\left(\left(y_i^{opt}, y_j^{opt}\right),(i',j')\right)^T \cdot tc_{a_i,a_j}}{norm\left(HP_{tc}\left(\left(y_i^{opt}, y_j^{opt}\right),(i',j')\right)\right)}. \end{aligned}$$

The probability density distributions of distances $d_x^a(y^{opt}, j)$, $d_g^a(y^{opt}, j)$, $d_{sc}^{a_i,a_j}((y_i^{opt}, y_j^{opt}),(i',j'))$ and $d_{tc}^{a_i,a_j}((y_i^{opt}, y_j^{opt}),(i',j'))$ can be estimated from training instances using kernel density estimation [34] for $j, i', j' \in \{1, \ldots, M\}$, $j \neq y^{opt}$ and $(i',j') \neq (y_i^{opt}, y_j^{opt})$. Abnormal activities are expected to have one or more infrequent potential distance scores.

A. Anomaly Definitions

Analogous to outlier detection in data mining [35], we introduce the concepts of point anomaly, contextual anomaly, and collective anomaly, whose definitions are given in the subsections below.

1) Point Anomaly: Point anomalies are detected without any contextual information [36]. Typically, for an atomic event in a video, the motion information captured from its local motion features follow certain patterns, which have been demonstrated by the popular activity classification method – BOW+SVM [31] upon STIP features. In our case, motion pattern of each activity class is reflected in the distributions of their distances to the hyperplanes HP_x. Denote the learned structural model as M.

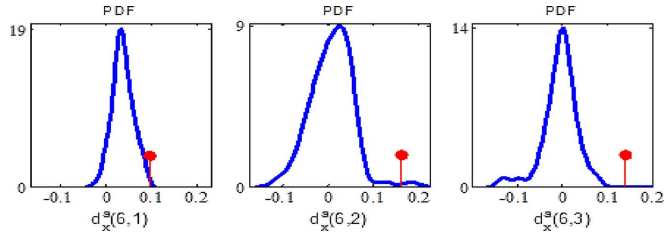


Fig. 5. Example of probability density functions of $d_x^a(y^{opt}, j)$, j =1, 2 and 3 for normal activities and the corresponding distances of a point anomaly (indicated by red circle). The first ten activities in Fig. 1 in the supplementary material are considered as normal activities, and used to train the structural model. For the point anomaly detected $y^{opt} = 6$.

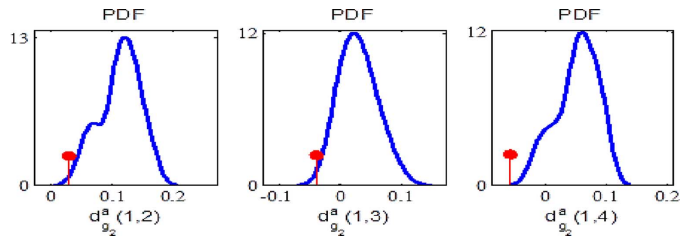


Fig. 6. Example of probability density functions of $d_{g_k}^a(y^{opt}, j), k = 2, j$ =2, 3 and 4, for normal activities and the corresponding distances of a contextual anomaly (indicated by red circle). The first ten activities in Fig. 1 in the supplementary material are considered as normal activities and used to train the structural model. For the contextual anomaly detected $y^{opt} = 1$. Intra-activity context subset G_2 is defined in Section III-B.

Given a test activity a_t with motion score histogram x_{a_t} and class label $y_{a_t}^{opt}$, define probability $p_x^{a_t}(j)$ as

$$p_x^{a_t}(j) = p\left(d_x^a\left(y_{a_t}^{opt}, j\right) < d_x^{a_t}\left(y_{a_t}^{opt}, j\right) | M\right), \quad (8)$$

for $j \in \{1, \ldots, M\}$ and $j \neq y_{a_t}^{opt}$. We use the p-test to determine whether an anomaly exists or not [37]. The (one-sided) p-value $P_x^{a_t}(j) = \min(p_x^{a_t}(j), 1 - p_x^{a_t}(j))$ measures the probability that the normal distribution of $d_x^a(y_{a_t}^{opt}, j)$ generates a value at least as extreme as $d_x^{a_t}(y_{a_t}^{opt}, j)$. The lower the p-value is, the more safe to say that the observed value does not belong to the normal distribution. So, we define the Motion-based Normality Factor MNF of a_t as the geometric mean of the associated p-values as

$$MNF(a_t) = \left(\prod_{j=1, j\neq y_{a_t}^{opt}}^{M} P_x^{a_t}(j)\right)^{1/(M-1)}. \quad (9)$$

Geometric mean is used here to measure the typical value of the set of p-values. As normal activities, which are known to us follow certain motion patterns captured by the distances, anomalous activities whose motion patterns deviate from the learned motion patterns significantly will have infrequent distances and thus a lower MNF than a threshold TH_{MNF}. Fig. 5 shows an example of distances of point anomaly.

2) Contextual Anomaly: Two kinds of attributes are generally involved with events: contextual attributes of the event, such as the location and surrounding objects; behavioral attributes include the motion of objects involved in the event. Contextual anomaly has normal behavioral attributes but abnormal contextual attributes. Given the intra-activity context feature g_{a_t} of the test activity with class label $y_{a_t}^{opt}$, define $p_{g_k}^{a_t}(j)$ as

$$p_{g_k}^{a_t}(j) = p\left(d_{g_k}^a\left(y_{a_t}^{opt}, j\right) < d_{g_k}^{a_t}\left(y_{a_t}^{opt}, j\right) | M\right), \quad (10)$$

for $k \in \{1, \ldots, N_G\}$, $j \in \{1, \ldots, M\}$ and $j \neq y_{a_t}^{opt}$. The probability that a_t belongs to class $y_{a_t}^{opt}$ based on g_k of a_t and $HP_{g_k}(y_a^{opt}, j)$ is $P_{g_k}^{a_t}(j) = \min(p_{g_k}^{a_t}(j), 1 - p_{g_k}^{a_t}(j))$. We define the Context-based Normality Factor CNF_k of a_t as the geometric mean of the associated p-values as

$$CNF_k(a_t) = \left(\prod_{j=1, j\neq y_{a_t}^{opt}}^{M} P_{g_k}^{a_t}(j)\right)^{1/(M-1)}. \quad (11)$$

If a_t has a high $MNF(a_t)$ and any of $CNF_k(a_t)$ for $k = 1, \ldots, N_G$, where N_G is the number of intra-activity context subsets, is lower than a threshold TH_{CNF}, a_t is considered as contextual anomaly. Fig. 6 shows an example of distances for contextual anomaly.

3) Collective Anomaly: A collection of activities forms a collective anomaly if the events as a whole deviate significantly from the entire training set. The collective anomaly can be further divided into sequential anomaly and co-occurrence anomaly. In our case, collective anomaly can also be considered as contextual anomaly since the detection of it utilizes the inter-activity context features – spatial and temporal relationships of activities. Assume activity collection A_t with member activities $a_1^t, a_2^t, \ldots, a_N^t$ related to each other in space and time are represented with class labels $Y_t^{opt} = [y_1^{opt}, y_2^{opt}, \ldots, y_N^{opt}]$, and $sc_{a_1^t,a_2^t}, \ldots, sc_{a_{N-1}^t,a_N^t}$ and $tc_{a_1^t,a_2^t}, \ldots, tc_{a_{N-1}^t,a_N^t}$ are their inter-activity context features. Define

$$p_{sc}^{(a_i^t,a_j^t)}(i',j') = p\left(d_{sc}^{a_i,a_j}(i'j') < d_{sc}^{(a_i^t,a_j^t)}(i',j') | M\right),$$
$$p_{tc}^{(a_i^t,a_j^t)}(i',j') = p\left(d_{tc}^{a_i,a_j}(i'j') < d_{tc}^{(a_i^t,a_j^t)}(i',j') | M\right).$$

Let

$$P_{sc}^{(a_i^t,a_j^t)}(i',j') = \min(p_{sc}^{(a_i^t,a_j^t)}(i',j'), 1 - p_{sc}^{(a_i^t,a_j^t)}(i',j')),$$
$$P_{tc}^{(a_i^t,a_j^t)}(i',j') = \min(p_{tc}^{(a_i^t,a_j^t)}(i',j'), 1 - p_{tc}^{(a_i^t,a_j^t)}(i',j')).$$

Define the Spatial Normality Factor SNF and Temporal Normality Factor TNF of (a_i^t, a_j^t) as

$$SNF\left(a_i^t, a_j^t\right) = \left(\prod_{i',j'=1}^{M} P_{sc}^{(a_i^t,a_j^t)}(i',j')\right)^{1/(M^2-1)}, \quad (12)$$

$$TNF\left(a_i^t, a_j^t\right) = \left(\prod_{i',j'=1}^{M} P_{tc}^{(a_i^t,a_j^t)}(i',j')\right)^{1/(M^2-1)}. \quad (13)$$

In (12) and (13), the condition – $(i',j') \neq (y_i^{opt}, y_j^{opt})$ – for the product is omitted for compactness of expression. If all member activities in A_{test} have high MNF and CNF values, but at least one of $SNF(a_i^t, a_j^t)$ is lower than a threshold TH_{SNF}, it is considered as a collective spatial anomaly. If at least one of $TNF(a_i^t, a_j^t)$ is lower than a threshold TH_{TNF} it is considered as a collective temporal anomaly. Fig. 7(a) shows an example

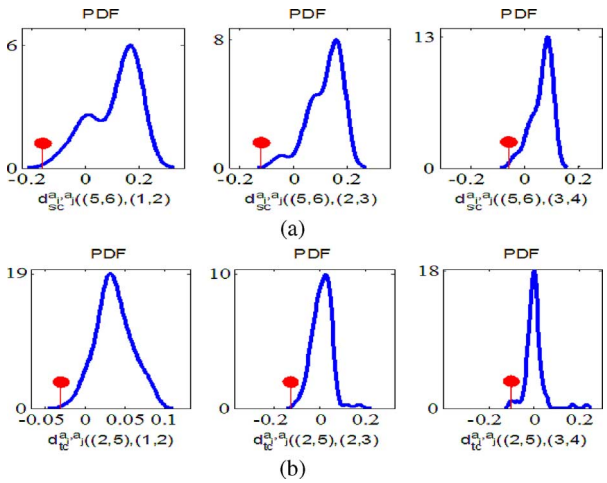


Fig. 7. (a): Example of probability density functions of $d^a_{sc}((y_i{}^{opt}, y_j{}^{opt}), (i', j'))$ for normal activities and the corresponding distances of a collective spatial anomaly (indicated by red circle). For the detected collective anomaly a_i and a_j $y_i{}^{opt} = 2$, $y_j{}^{opt} = 5$. (b): Example of probability density functions of $d^a_{tc}((y_i{}^{opt}, y_j{}^{opt}), (i', j'))$ for normal activities and the corresponding distances of a collective temporal anomaly (indicated by red circle). For the detected collective anomaly a_i and a_j $y_i{}^{opt} = 5$, $y_j{}^{opt} = 6$ (In both cases, the first ten activities in Fig. 1 in the supplementary material are considered as normal activities and used to train the structural model).

of distances of collective spatial anomaly and Fig. 7(b) shows an example of distances of collective temporal anomaly.

VI. Experimental Evaluation

To assess the effectiveness of our structural model in activity-based video modeling, we first perform experiments on the public VIRAT Ground Dataset Release 1 [38], and compare our results with two baseline methods: BOW+SVM [32] and SFG-classifier [10]. We use the SFG since it considers feature-level context and we demonstrate the effectiveness of activity-level context on top of it. Then, we work on Release 2 for anomaly detection using BOW+SVM as the baseline classifier.

A. Dataset

VIRAT Ground dataset is a state-of-the-art activity dataset with many challenging characteristics, such as wide variation in the activities and clutter in the scene. The dataset consists of surveillance videos of realistic scenes with different scales and resolution, each lasting 2 to 15 minutes and containing up to 30 events. The activities defined in the whole dataset is shown in Fig. 1 in the supplementary material. In the first set of experiments, we use Release 1 to evaluate the performance of the proposed method and assess the significance of context features in activity recognition. We examine the first six activity classes defined in release 1 as shown in Fig. 1 in the supplementary material. We randomly select half of the data for training and the rest for testing. In the second set of experiments for activity recognition and anomaly detection, we use VIRAT release 2, in which the eleven activity classes are defined as shown in Fig. 1 in the supplementary material.

B. Preprocessing

Motion regions that involve only vehicles moving are excluded from the experiments since we are only interested in person related normal and anomalous activities. For the

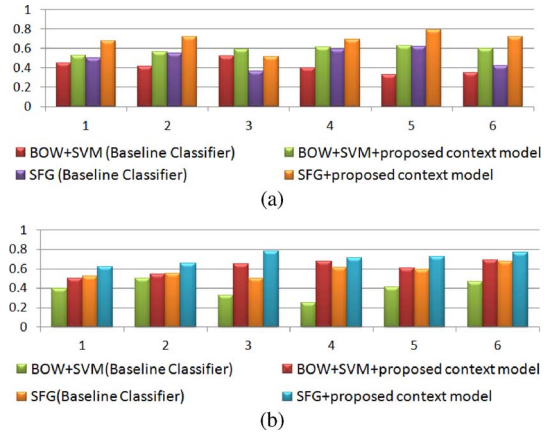


Fig. 8. Average precision (a) and recall (b) of each activity class using different methods. For each set of bars, from left to right, the associated methods are BOW+SVM (baseline classifier), our method using BOW+SVM, SFG (baseline classifier) and our method using SFG. The six activity classes considered are the first six activity classes listed in Fig. 1 in the supplementary material.

BOW+SVM classifier, $k = 1000$ visual words and a 9-nearest neighbor soft-weighting scheme are used. For the SFG-based classifier, the size of each temporal bin used is 5 frames while other settings are the same as in [10]. For the SFG method [10], activity localization is implicitly included in the recognition process. The method generates similarity scores $s \in \{0, 1\}$ between two activities. In the training process of the baseline classifier, we calculate the intra-class similarity (the average similarity score between a given activity instance and other instances of the same class) and inter-class similarity (the average similarity score between a given instance and other instances of different classes). We define the representative instances as the ones with maximum difference in their intra-class similarity and inter-class similarity. A fixed number of representative instances for each activity class are selected during the training of the baseline classifier (5 are used in the experiments).

Persons and vehicles are detected using the publicly available software [29]. Opening/closing of doors of facilities, boxes and bags are detected using method in [39] with Histogram of Gradient as the low-level feature and binary linear-SVM as the classifier. Motion score histograms described in Section III-B are generated for each activity. The score histogram of an activity contains the average similarity scores between the activity and the representative examples. For experiment 1, the intra-activity context features are built based on first two cues in Fig. 2, and all cues are used for experiment 2.

C. Recognition Results on VIRAT Release 1

We show our results on VIRAT release 1 using precision and recall scores in Fig. 8. The average recognition accuracies, measured by average precision are 40%, 59%, 51% and 68% respectively for the four methods – BOW+SVM [32], BOW+SVM+Context model, SFG [10] and SFG+Context model. The implementation using *SFG* based classifier outperforms those using *BOW + SVM* baseline classifier. However, our model increases the recognition performance uniformly by a large amount over all baseline classifiers. The results are expected since the intra-activity and inter-activity context gives the model additional information about the activities other than the motion information encoded in low-level features.

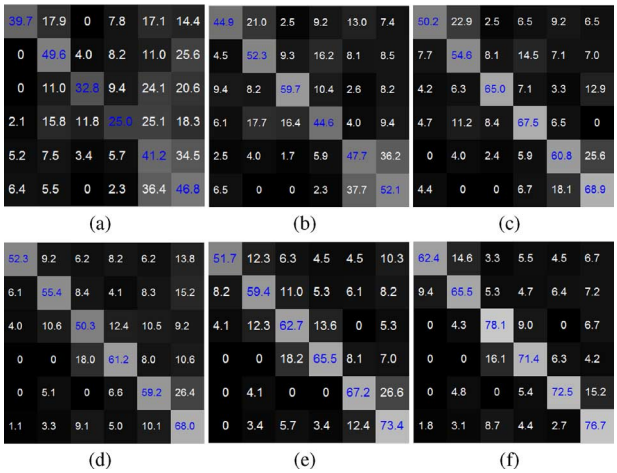


Fig. 9. Recognition Results on VIRAT Release 1. (a): Confusion matrix for BOW+SVM classifier; (b): Confusion matrix for our approach using motion and intra-activity context feature (using BOW+SVM as the baseline classifier); (c): Confusion matrix for our approach using motion and intra- and inter-activity context feature (use BOW+SVM as the baseline classifier); (d): Confusion matrix for SFG-based classifier; (e): Confusion matrix for our approach using motion and intra- and inter-activity context feature (using SFG as the baseline classifier); (f): Confusion matrix for our approach using motion and intra- and inter-activity context feature (using SFG as the baseline classifier).

Fig. 9 shows the confusion matrix for baseline classifiers and our method. The baseline classifiers often misclassify "unloading from a vehicle" as "getting out of a vehicle". However, common sense tells us that "unloading from a vehicle" often happens in the rear of the vehicle while "getting out of a vehicle" happens near the side of the vehicle. Also, the baseline classifiers often confuse "open a vehicle trunk" and "close a vehicle trunk" with each other. However, if the two activities happen closely in time in the same place, the first activity in time is probably "open a vehicle trunk". These spatio-temporal relationships within and across activity classes are captured by our model and used to improve upon the recognition performance. Although the SFG method in [10] models the spatial and temporal relationships between the features, it does not consider the relationships between various activities and thus our method outperforms the SFGs.

Finally, we show some example activities recognized using baseline classifier and examples corrected by intra-activity and inter-activity context features in Fig. 10.

D. Anomaly Detection on VIRAT Release 2

In order to access the performance of the proposed model on anomaly detection, we work on VIRAT Release 2, in which eleven activity classes are defined as shown in Fig. 1 in the supplementary material. We follow the method defined above to get the recognition results on this dataset. Fig. 11 shows the confusion matrix for VIRAT Release 2.

1) Point Anomaly: For the detection of point anomaly, we randomly selected one of the eleven activity classes as abnormal, and treated other activities as normal. Cross-validation is used to assess the performance of anomaly detection. For each run, we assume that we do not have training instances for abnormal activities, so, the activities of abnormal class are excluded from the learning process. We use BOW+SVM as the baseline classifier.

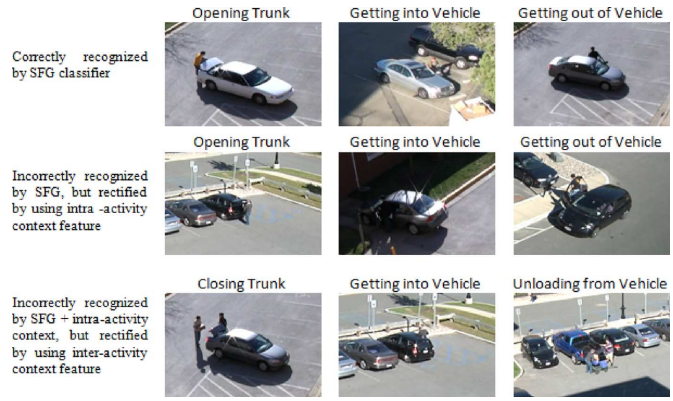


Fig. 10. Examples show the effect of context features in recognizing activities that were incorrectly labeled by the SFG classifier.

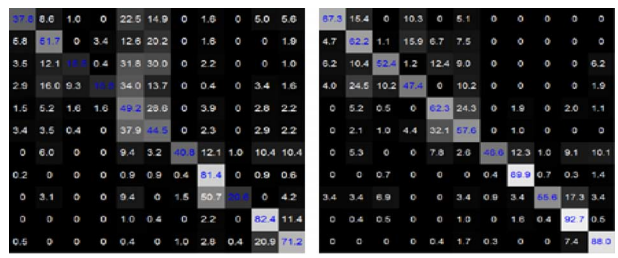


Fig. 11. Recognition Results for VIRAT Release 2. (a): Confusion matrix for BOW+SVM baseline classifier; (b): Confusion matrix for our approach using BOW+SVM as the baseline classifier. The activities considered are listed in Fig. 1 in the supplementary material.

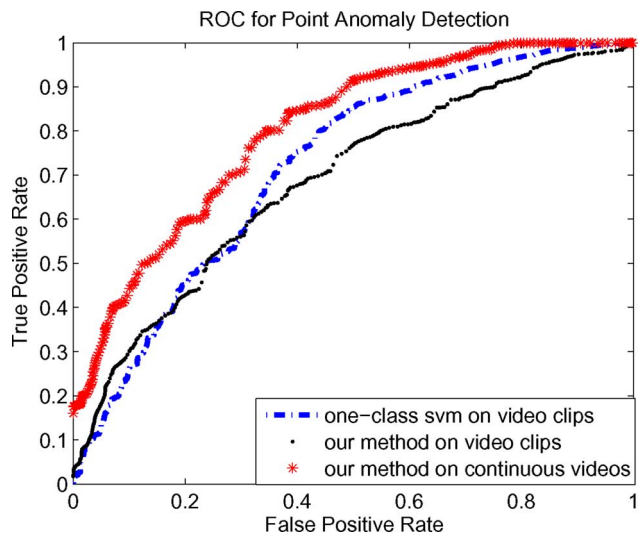


Fig. 12. ROC curves for point anomaly detection.

One-class SVM is often used for point anomaly detection [40]. To access the effectiveness of our model in detecting point anomalies, we compare our results with those using one-class SVM. For fair comparison, we also apply the proposed framework on video clips, each containing one activity of the eleven classes. Fig. 12 shows the ROC curves of BOW+SVM and our method. The areas under curve are 79.8% for our method on video clips, 72% for one-class svm on video clips and 68.5% for our method working on continuous videos.

2) Contextual Anomaly: For the detection of contextual anomalies, we consider activities that are normal in terms of motion features but with abnormal or infrequent intra-activity

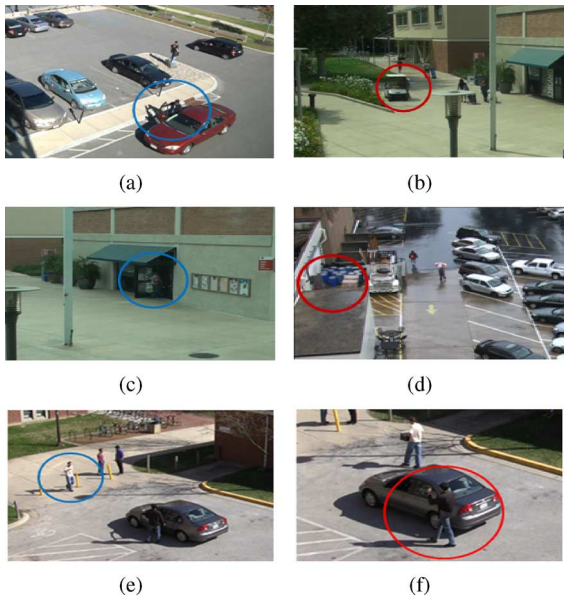


Fig. 13. (a, c, e): Examples of normal activities; (b, d, f): Examples of detected contextual anomalies. First row: Person getting into a vehicle usually occurs in the parking area (a), and the anomaly is detected when it happen in an area not for parking (b). Second row: Person exiting a facility happens at a normal exit (c), whereas an anomaly is detected when the person exits from a door that is rarely used (d). Third row: A person gesturing far from a vehicle is normal in our dataset (e), whereas in (f) the 'gesturing' occurs near the trunk of the vehicle, which is identified as a contextual anomaly.

context features as discussed in Section V-A2. The normality threshold $TH_{CNF} = 0.05$ in the experiment. Fig. 13 shows examples of detected contextual anomalies. In the first example, person getting into a vehicle usually occurs in the parking area, and the anomaly is detected when it happens in an area not for parking. In the second example, the anomaly of 'person exiting a facility' is detected when the person exits from a door of a facility that is rarely used. In the third example, the anomaly of 'gesturing' occurs near the trunk of the vehicle while others in our dataset usually occur faraway from the vehicle. None of these could have been detected without the modeling of the intra-activity context feature.

3) Collective Anomaly: Collective anomaly can be detected based on the learned inter-activity context patterns and the inter-activity contextual features of the test instances. The normality thresholds $TH_{SNF} = 0.05$ and $TH_{TNF} = 0.05$ are used. For the first example, we consider two activities – 'person getting into a vehicle' and 'person unloading an object from a vehicle'. For most of the examples in the dataset, when these two activities happen together, the unloading happens from the trunk of the car while the person enters through the driver's door. Thus, as shown in Fig. 14(a),(b), a collective spatial anomaly is detected when the unloading and entering happen near the same part of the vehicle. An example of a collective temporal anomaly is shown in Fig. 14(c),(d). The example of a 'person getting out of a vehicle' usually occurs before 'person getting into a vehicle', however, in the detected anomaly, 'person getting out of a vehicle' occurs after 'person getting into a vehicle'.

VII. Conclusion

In this paper, we present a novel approach to jointly model a variable number of activities in videos, and detect abnormal activities. We represent a video of a wide area by sets of activities that are spatially and temporally related. A structural model is proposed to learn the motion patterns and context patterns within and across activity classes from training sets of activities. The inference process tries to generate the correct labels for testing instances using the learned parameters through a greedy search method. Our experiments have shown that encapsulating object interactions and spatial and temporal relationships of activity classes can be used to significantly improve the recognition accuracy. The proposed model can detect point anomalies, contextual anomalies, and collective anomalies based on the motion and various context features.

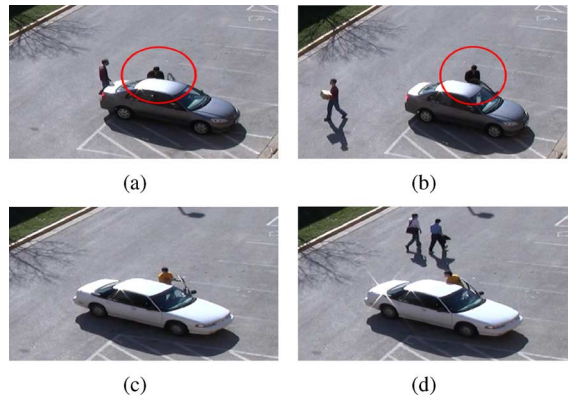


Fig. 14. Example of collective spatial anomaly and collective temporal anomalies. (a, b): we consider two activities – 'person getting into a vehicle' and 'person unloading an object from a vehicle'. For most of the examples in the dataset, when these two activities happen together, the unloading happens from the trunk of the car while the person enters the driver's door. Thus, a collective spatial anomaly is detected when the unloading and entering happen near the same part of the vehicle. (c, d): The example of a 'person getting out of a vehicle' usually occurs before 'person getting into a vehicle', however, in the detected anomaly, 'person getting out of a vehicle' occurs after 'person getting into a vehicle'.

References

[1] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. Conf. Pattern Recognition*, 2004.
[2] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, Dec. 2007.
[3] A. Oliva and A. Torralba, "The role of context in object recognition," *Trends in Cognitive Sci.*, vol. 11, no. 12, pp. 520–527, Dec. 2007.
[4] V. I. Morariu and L. S. Davis, "Multi-agent event recognition in structured scenarios," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition*, 2011, pp. 3289–3296.
[5] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human object interaction activities," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition*, 2010, pp. 17–24.
[6] C. H. Teo, Q. Le, A. Smola, and S. V. N. Vishwanathan, "A scalable modular convex solver for regularized risk minimization," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery and Data Mining*, 2007, pp. 727–736.
[7] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *Proc. Workshop Statist. Learn. in Comput. Vision, ECCV*, 2004, pp. 17–32.
[8] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Comput. Surveys*, vol. 43, no. 16, Apr. 2011, Article no. 16.
[9] M. Ryoo and J. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *Proc. IEEE Proc. Conf. Comput. Vision*, 2009, pp. 1593–1600.
[10] U. Guar, Y. Zhu, B. Song, and A. K. Roy-Chowdhury, "A string of feature graphs model for recognition of complex activities in natural videos," in *Proc. IEEE Proc. Conf. Comput. Vision*, 2011, pp. 2595–2602.

[11] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition*, 2011.
[12] T. Lan, Y. Wang, S. N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1549–1562, Aug. 2012.
[13] R. Benmokhtar and I. Laptev, Inria-Willow at Trecvid2010: Surveillance Event Detection, [Online]. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv10.papers/inria-willow.pdf
[14] K. Tang, L. Fei-Fei, and D. Koller, "Learning latent temporal structure for complex event detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition*, 2012, pp. 1250–1257.
[15] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis, "Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition*, 2009, pp. 2012–2019.
[16] Z. Si, M. Pei, B. Yao, and S. Zhu, "Unsupervised learning of event and-or grammar and semantics from video," in *Proc. IEEE Conf. Comput. Vis.*, 2011, pp. 41–48.
[17] Y. Benezeth, P.-M. Jodoin, and V. Saligrama, "Abnormality detection using low-level co-occuring events," *Pattern Recognition Lett.*, pp. 423–431, 2011.
[18] V. Saligrama, J. Konrad, and P.-M. Jodoin, "Video anomaly identification," *IEEE Signal Process. Mag.*, vol. 27, no. 5, pp. 18–33, Sep. 2010.
[19] X. Wang, X. Ma, and W. E. L. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 539–555, Mar. 2009.
[20] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition*, 2012, pp. 2112–2119.
[21] V. Saligrama and Z. Chen, "Sparse reconstruction cost for abnormal events detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition*, 2011, pp. 3449–3456.
[22] R. Hamid, A. Johnson, S. Batta, A. Bobick, C. Isbell, and G. Coleman, "Detection and explanation of anomalous activities: Representing activities as bags of event n-grams," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition*, 2005, pp. 1031–1038.
[23] R. Hamid, S. Maddi, A. Bobick, and I. Essa, "Unsupervised analysis of activity sequences using event-motifs," in *Proc. 4th ACM Int. Workshop Video Surveill. Sens. Netw.*, 2006.
[24] R. Hamid, S. Maddi, A. Bobick, and I. Essa, "Structure from statistics – Unsupervised activity analysis using suffix trees," in *IEEE Proc. Conf. Comput. Vis.*, 2007, pp. 1–8.
[25] S.-W. Joo and R. Chellappa, "Attribute grammar-based event recognition and anomaly detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition*, 2006, p. 107.
[26] I. Saleemi, K. Shafique, and M. Shah, "Probabilistic modeling of scene dynamics for applications in visual surveillance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 8, pp. 1472–1485, Aug. 2009.
[27] F. Jiang, Y. Wu, and A. K. Katsaggelos, "A dynamic hierarchical clustering method for trajectory-based unusual video event detection," *IEEE Trans. Image Process.*, vol. 18, no. 4, pp. 907–913, Apr. 2009.
[28] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *Proc. Int. Conf. Pattern Recognition*, 2004.
[29] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Discriminatively Trained Deformable Part Models, Release 4," [Online]. Available: http://people.cs.uchicago.edu/ pff/latent-release4/
[30] B. Song, T. Jeng, E. Staudt, and A. Roy-Chowdury, "A stochastic graph evolution framework for robust multi-target tracking," in *Proc. Euro. Conf. Comput. Vis.*, 2010.
[31] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. IEEE Proc. Conf. Comput. Vis.*, 2003, pp. 432–439.
[32] Y.-G. J. G.-W. Ngo and J. Yang, "Towards optimal bag of word for object categorization and semantic video retrieval," in *Proc. ACM Proc. Conf. Image and Video Retrieval*, 2007.
[33] C. Desai, D. Ramanan, and C. C. Fowlkes, "Discriminative models for multi-class object layout," *Int. J. Comput. Vis.*, vol. 95, no. 1, pp. 1–12, Oct. 2011.
[34] C. M. Bishop, *Pattern Recognition and Machine Learning*, 2nd ed. New York: Springer, 2006.
[35] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*. Amsterdam, The Netherlands: Elsevier, 2011.
[36] F. Jiang, J. Yuan, S. A. Tsaftaris, and A. K. Katsaggelos, "Anomalous video event detection using spatiotemporal context," *Comput. Vis. Image Understanding*, vol. 115, no. 3, pp. 323–333, Mar. 2011.
[37] D. Freedman, R. Pisani, and R. Purves, *Statistics*, 4th ed. New York: Norton, 2007.
[38] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai, "A large-scale benchmark dataset for event recognition in surveillance video," in *IEEE Conf. Comput. Vis., Pattern Recognition*, 2011, pp. 3153–3160.
[39] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition*, 2005, pp. 886–893.
[40] B. Scholkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, pp. 1443–1471, 2001.

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, image/video processing and communication.

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.

Amit K. Roy-Chowdhury received the Bachelor's degree in electrical engineering from Jadavpur University, Calcutta, India, the Master's 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 an Associate 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 video communications and statistical methods for signal analysis. 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), biological video analysis, and distributed video compression. He is a coauthor of the books Camera Networks: The Acquisition and Analysis of Videos over Wide Areas. He is the editor of the book Distributed Video Sensor Networks. He has been on the organizing and program committees of multiple computer vision and image processing conferences and serves on the editorial boards of a number of journals.