Detecting Group Activities With Multi-Camera Context

Zheng-Jun Zha, *Member, IEEE,* Hanwang Zhang, Meng Wang, *Member, IEEE,* Huanbo Luan, and Tat-Seng Chua

***Abstract*—Human group activities detection in multi-camera CCTV surveillance videos is a pressing demand on smart surveillance. Previous works on this topic are mainly based on camera topology inference that is hard to apply to real-world unconstrained surveillance videos. In this paper, we propose a new approach for multi-camera group activities detection. Our approach simultaneously exploits intra-camera and inter-camera contexts without topology inference. Specifically, a discriminative graphical model with hidden variables is developed. The intra-camera and inter-camera contexts are characterized by the structure of hidden variables. By automatically optimizing the structure, the contexts are effectively explored. Furthermore, we propose a new spatiotemporal feature, named vigilant area (VA), to characterize the quantity and appearance of the motion in an area. This feature is effective for group activity representation and is easy to extract from a dynamic and crowded scene. We evaluate the proposed VA feature and discriminative graphical model extensively on two real-world multi-camera surveillance video data sets, including a public corpus consisting of 2.5 h of videos and a 468-h video collection, which, to the best of our knowledge, is the largest video collection ever used in human activity detection. The experimental results demonstrate the effectiveness of our approach.**

***Index Terms*—Activity detection, context, group activity, human activity.**

I. Introduction

THE PAST few decades have witnessed a rapid proliferation of surveillance cameras and have resulted in a tremendous explosion of surveillance videos. The detection of human activities in surveillance videos attracts increasing attention from academia, industry, and security agencies, and so on. In recent years, intensive research has focused on vision-based human activity detection [1]–[3]. Most of these works are constrained in single-camera views and focus on detecting actions of a few individuals, usually one or two individuals, such as hand-waving and hugging in [4] and [5], and ObjectPut and CellToEar in TRECVid [6]–[8]. However, today, CCTV surveillance is set up with camera networks in almost every public area such as airports, campuses, and government buildings for safety and security purposes. The motivation is straightforward; the multi-sensor system can provide better views and monitor larger areas. Therefore, modeling activities under multiple cameras becomes increasingly demanding. On the other hand, group activity detection has recently attracted increasing attention. Here, group activity refers to the interactions among a group of persons; in particular, no less than three persons. The modeling of group activities has great potential for many applications, such as smart video surveillance and human computer interface.

Motivated by the above observations, this paper focuses on detecting group activities in multi-camera surveillance videos. Previous works on multi-camera activity analysis are mostly based on inter-camera topology inference [9]–[12]. Inter-camera topology is the explicit relationship of multiple cameras, including connectivity of disjoint views and time dependence of one object moving from one camera to another. These methods rely on accurate tracking results [10], [11] or clustering areas under different cameras based on visual cues [9], [12]. Unfortunately, they are hard to apply to real-world unconstrained surveillance videos. The unconstraint aspects include: 1) crowded scene and low video quality, which leads to unreliable tracking results and 2) nonconstant human crowd flow. Different from the surveillance videos of underground station, where the majority of passengers follow a regular route according to a fixed channel, human crowd monitoring over an open scene such as a building with many exits and entrances often leads to unpredictable crowd activities.

In this paper, we propose a novel approach for multi-camera group activity detection without camera topology inference. Our approach implicitly exploits intra-camera and inter-camera contextual information at the same time. While the intra-camera context encodes the relationship of motions captured by one camera, the inter-camera context captures visual cues in multi-camera views under a given group activity. Fig. 1 shows an illustration of our approach. More concretely, we first propose an effective group activity feature called VA, which stands for vigilant area. VA is a region-based feature that describes the quantity and appearance of the motion (vigilance) in an area. This feature is robust and easy to extract from a dynamic and crowded scene without reliance on any tracking. Then, we develop a discriminative graphical model with hidden variables that govern whether the low-level features, i.e., VAs are included in the determination of a certain group activity. The intra/inter-camera contexts

Manuscript received January 31, 2012; revised June 25, 2012; accepted August 17, 2012. Date of publication October 26, 2012; date of current version May 1, 2013. This work was supported by the NUS-Tsinghua Extreme Search (NExT) Project under Grant R-252-300-001-490. This paper was recommended by Associate Editor A. Signoroni.

Z.-J. Zha, H. Zhang, H. Luan, and T.-S. Chua are with the School of Computing, National University of Singapore, 117417, Singapore (e-mail: zhazj@comp.nus.edu.sg).

M. Wang is with the Hefei University of Technology, Hefei, China.

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/TCSVT.2012.2226526

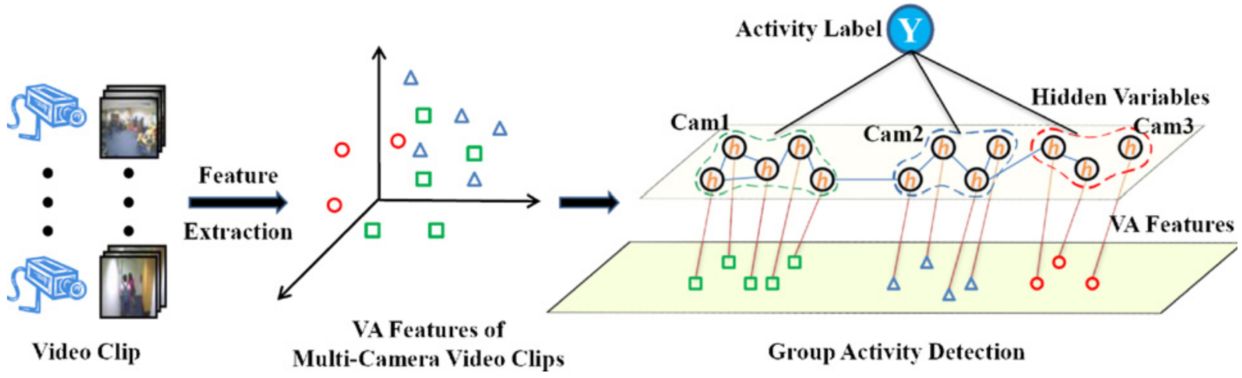

Fig. 1. Illustration of our proposed multi-camera group activity detection approach.

are interpreted by the structure of hidden variables, i.e., the connections among them. Different from the existing latent graphical models, such as hidden Markov models (HMM) and hidden conditional random field (HCRF), which assume a predefined and fixed structure for the latent variables (e.g., chain structure or tree structure), the proposed model does not fix the structure of the latent variables, but optimizes it during learning and inference. Through automatic inference of the nonfixed structure of hidden variables, the proposed model can explore the intra/inter-camera contexts intelligently for activity detection.

We apply our proposed approach to perform group activity detection on two real-world multi-camera surveillance video data sets, including a 468-h video corpus collected by nine cameras with a total footage duration of 468 h and the publicly available UCR VideoWeb dataset[1] [13], which contains 2.5 h of surveillance videos. Experimental results show that: 1) our proposed VA feature is more effective for group activity representation than local spatial-temporal descriptor such as motion scale invariant feature transformation (MoSIFT) [7] and the recently proposed dense trajectory feature [14] on real-world unconstrained surveillance videos; 2) our approach outperforms the state-of-the-art methods; and 3) multi-camera contexts are beneficial to group activity detection.

The main contributions of this paper are as follows.

1) We propose a novel approach for group activity detection in multiple-camera surveillance videos by exploiting the intra/inter-camera contextual information.
2) The intra/inter-camera contexts can be automatically inferred and well explored by our proposed discriminative graphical model without the requirement of camera topology inference.
3) We propose a novel spatiotemporal group activity feature, termed as VA, which does not rely on tracking and is thus applicable to a vast array of surveillance videos with various qualities.

The remainder of this paper is organized as follows. In Section II, we review existing works on activity analysis in surveillance videos. We introduce our proposed feature VA and its corresponding quasi-ShapeContext descriptor in Section III. In Section IV, we describe our group activity detection model in detail, including model formulation in Section IV-A, and inference and learning in Section IV-B. We report experimental results in Section V, followed by conclusions in Section VI.

II. Related Work

Vision-based human activity detection has been intensively studied in the past past decades [1]–[3]. A lot of existing works mainly focus on modeling activities in single-camera views, including activities of a few individuals (e.g., hand-waving and walking) and group activities with multiple participants (e.g., walking-in-group and stand-talk). Recently, CCTV surveillance has been set up with camera networks in almost every public area for safety and security purposes. Modeling activities under multiple cameras has attracted increasing attention. In this section, we first review previous works on single-camera activity analysis, and then introduce recent research efforts on multi-camera activity analysis.

A. Activity Analysis in Single-Camera Views

The challenge of detecting human activity in surveillance video arises from severe cluttered background and occlusion [8]. Most existing activity detection methods rely on spatiotemporal features that interpret the visual content of activities. With the spatiotemporal features, many existing machine learning methods can be employed for activity detection [2]. These features can be divided into two categories: global representation [6], [15], [16] and local representation [7], [14], [17]. In global representation, instead of locating the entire human body, partial localization of a human body part such as head [6] or multi-candidate regions for the actual body is done for preprocessing [15]. For example, Zhu *et al.* [6] detected and tracked the human head in a given video sequence. An enlarged region around the tracked head was cropped as the region of interest. A spatiotemporal descriptor, which temporally integrates the statistics of a set of response map of image gradients and optical flows, was then extracted from the region to capture the characteristics of human actions in terms of their appearance and motion patterns. Hu *et al.* [15] proposed to crop multiple regions of human body from the frames. These regions were then represented by the motion history image (MHI) feature [18]. However, the cropped regions contain noises that are not action instances. To address

[1] Available at http://www.ee.ucr.edu/ amitrc/vwdata.php.

this problem, they developed a multi-instance learning model, which is expected to learn human actions from the noisy instances. However, it is difficult to segment body parts when occlusion exists in the scene [16]. Alternative approaches to segmenting body parts based on analyzing 3-D volumes based on various features, like a histogram of gradient, a histogram of optical flow [19], or 3-D SIFT [20]. Dollar *et al.* [19] proposed a spatiotemporal feature detector, which was especially designed to extract space–time points with local periodic motions, obtaining a sparse distribution of interest points from a video. A small 3-D volume, called a cuboid, was then associated to each interest point. Each cuboid captures a visual appearance of the interest point's neighborhoods. A library of cuboid prototypes was then constructed by clustering cuboid appearances with k-means. As a result, each action was modeled as a histogram of cuboid types detected in 3-D space–time volume while ignoring their locations. Scovanner *et al.* [20] designed a 3-D variant of the SIFT descriptor, similar to the cuboid features in [19].

On the other hand, local features based on spatiotemporal interest points have shown to encourage capacity in representing activities. The local interest points are expected to be not only scale-invariant in the spatial domain but also in the temporal domain. The best characteristic of such scale-invariant local interest points is that they require no tracking [7]. A wealth of local spatiotemporal features have recently been developed. A survey of them is provided in [3], while an overall evaluation of some local features can be found in [17]. Several works extracted local features from every frame and concatenated them temporally to describe the overall motion of human activities [14], [21]–[23]. For example, Chomat and Crowley [21] proposed to use local appearance descriptors to characterize an action. They employed the motion energy receptive fields together with Gabor filters to capture motion information from a sequence of images. More specifically, local spatiotemporal appearance features describing motion orientations were detected per frame. Multi-dimensional histograms were then constructed based on the local features and used to represent actions. Zelnik-Manor and Irani [22] utilized local spatiotemporal features at multiple temporal scales. Multiple temporally scaled video volumes were analyzed. For each point in a 3-D XYT volume, they estimated a normalized local intensity gradient. A histogram of these space–time gradient features was then computed per video. Similarly, Blank *et al.* [23] calculated local features in each frame. Instead of utilizing optical flows, they calculated appearance-based local features at each pixel by constructing a space–time volume whose pixel values are the solutions to the Poisson equation. Wang *et al.* [14] recently proposed a dense trajectory feature. They sampled dense points from each frame and tracked the points based on displacement information from a dense optical flow field. Some global smooth constraints were then imposed among the points in dense optical flow fields, leading to more robust trajectories than tracking or matching points separately. Some other approaches extracted sparse spatiotemporal local interest points from 3-D volumes [19], [24]–[27]. These local spatiotemporal features possess robustness to noise, camera jitter, illumination changes, and background movements.

Beyond the detection of relatively simple activities of a few individuals, the recognition of group activity with multiple participants (e.g., fighting and gathering) has gained increasing interests in recent years [28]–[35]. One approach focuses on modeling the interactions between the participants. Lan *et al.* [30] established a discriminative graphical model with a variant structure to infer the interactions. Instead of modeling the interactions explicitly, a second approach tries to extract features that encode interaction information. Ni *et al.* [29] used different types of causality filters and the corresponding responses were considered as group activity features. In particular, three types of localized causalities, including self causality, pair causality, and group causality, were exploited to characterize the local interaction/reasoning relations within, between and among motion trajectories of different humans, respectively. Each type of causality was expressed as a specific digital filter, whose frequency responses then constituted the feature representation space. Each video segment of a certain group activity was encoded as a bag of localized causalities/fitlers. Xiang *et al.* [34] adopted the pixel change history (PCH) to measure the multi-scale temporal changes at each pixel and then formed discrete events based on detected changes in each video frame. The connected component method was adopted to group the changed pixels. The interactions were embedded in each cluster. The third approach collaboratively uses the contextual information related to participants. For example, Marszalek *et al.* [28] pointed out that human actions are frequently constrained by the purpose and the physical properties of scenes and demonstrate a high correlation with particular scene classes. They discovered relevant scene classes and their correlation with human actions. Such contextual information was then used to improve activity detection. Recently, Ryoo and Aggarwal [35] proposed a stochastic methodology for the recognition of group activities. Their system maintained probabilistic representation of a group activity, describing how individual activities of its group members must be organized temporally, spatially, and logically. In order to recognize each of the represented group activities, the system searched for a set of group members that has maximum posterior probability. A hierarchical recognition algorithm utilizing a Markov chain Monte Carlo-based probability distribution sampling was designed to detect group activities and find the acting groups simultaneously.

B. Activity Analysis in Multi-Camera Views

Existing works on group activity analysis in multi-camera views are mainly based on camera topology inference, which relies on inter-camera tracking, intra-camera tracking, or both. The topology includes overlapping fields of view (FOV), time dependency, and causality of activities captured by cameras. Generally, there are two kinds of topologies: 1) the topology of the geographical location and FOV of cameras [36] and 2) the topology of the semantic scene (or content topology) of cameras [9]. The first is beyond our research and the geological topology is known in advance in our case. The other kind of topology is significant since even if the geographical topology is fixed, the content that the camera is recording may change

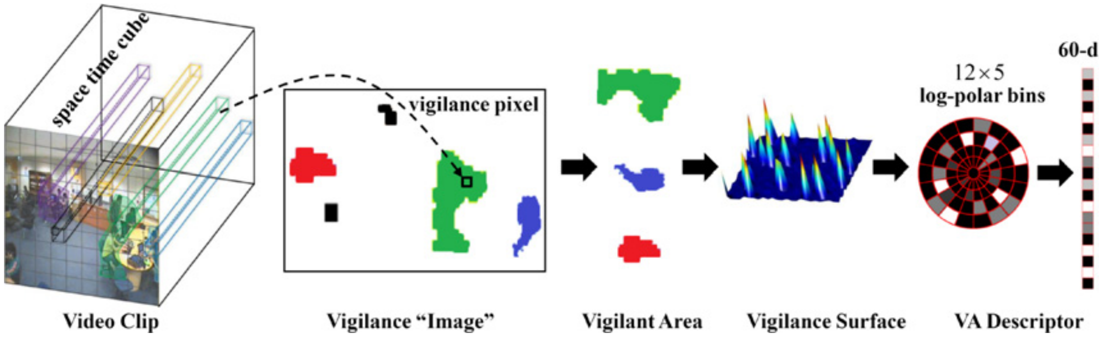

Fig. 2. Illustration of extracting VA descriptors from a video clip. A vigilance surface is constructed for each vigilance area.

over times. The crux is to discover the relations between two cameras. The solution to find the start and end zone in each camera [37], the temporal relation [10], or both [11].

Based on the topology of cameras, group activity is represented as trajectories [10], tracked blobs [38], or region clusters [9], [12]. Wang *et al.* [10] adopted a latten Dirichlet allocation model to cluster the trajectories into activities. They first tracked objects in each camera view independently. The positions and velocities of objects along trajectories were computed as features. Then, an LDA model was used to group trajectories, which belong to the same activity but may be in different camera views, into one cluster, and also model the paths that are commonly taken by objects across multiple camera views. Chang *et al.* [38] used a hierarchical agglomerative clustering algorithm to group the individual actions into group-level activities. Alternatively, Chen *et al.* [12] proposed a fixed tree-type time dependency probabilistic graph model to infer the post probability of observations in a semantic region, where low accumulative log-likelihood value indicates an anomalous activity. Their method optimized time-delayed dependencies globally. A cumulative abnormality score was introduced to replace the conventional log-likelihood score for gaining more robust anomaly detection.

The above methods implicitly assume that humans can be tracked accurately and the flow of a human crowd is constant. However, in a general surveillance setting, such assumptions are always invalid. Therefore, activity analysis based on camera topology is challenging and even infeasible in unconstrained multi-camera surveillance.

III. Activity Content Representation

Occlusion often exists in real-world surveillance videos with dynamic and crowded scenes. It is thus infeasible to obtain accurate tracking of activities. Many popular features that rely on tracking will lose their power [29], [31]. Therefore, there is a compelling need to design a new feature that does not depend on tracking to represent the visual content of group activities in dynamic and crowded scenes. We propose here a new spatiotemporal feature, termed VA. The VA feature describes the quantity and appearance of foreground motions (vigilance) in an area. The detection of VA benefits from some effective fast computer vision techniques such as background subtraction and morphological operation. As we will discuss later, VA is effective in representing group activities in dynamic and crowded scenes, and is easy to extract without the requirement of tracking results. Fig. 2 illustrates the pipeline of VA extraction. Given a video clip, we first quantize it into blocks of space–time cubes. We then calculate the accumulate number of foreground pixels in each space–time cube to characterize the quantity of motion (vigilance) in a time interval. Afterward, we represent the motion information within a video clip as a vigilance image, where each pixel corresponds to a space–time cube in the video clip, and its intensity is the vigilance of that cube. Each connected component in the vigilance image that is sufficiently large are chosen as a VA. Finally, to represent the shape and motions of each VA, we map the VA into log-polar coordinate to obtain a Quasi-ShapeContext descriptor.

Next, we will first introduce the details of VA detection and then present our quasi-ShapeContext descriptor.

A. Vigilant Area Detection

Without loss of generality, we only describe the detection of VAs in a video clip of one camera and the process can be easily extended to video clips from other cameras.

Given a video clip, the foreground is first detected using a background subtraction method [39]. Note that the length of the video clip is essential for the time scale and we should set it to a reasonable value. According to the taxonomy in [34], we set different time scales τ for different activities. The setting of τ is described in Section V-A. We denote a video clip as $v(x, y, t), t \in [0, \tau]$ where (x, y) and t are space and time coordinates, respectively. The video clip is then divided into space–time cubes in size of $10 \times 10 \times \tau$ pixels. In a space–time cube, the number of foreground pixels represents its intensity of motions. We name the intensity of motion as vigilance. The vigilance of a certain space–time cube is calculated as

$$vig(m, n) = \sum\nolimits_{\text{pixels in the cube (m,n)}} I(x, y, t)$$
$$\begin{cases} I(x, y, t) = 0, \text{if pixel}(x, y, t) \text{ is background} \\ I(x, y, t) = 1, \text{if pixel}(x, y, t) \text{ is foreground} \end{cases} \tag{1}$$

where $vig(m, n)$ is the vigilance of the (m, n)th space–time cube in the video clip.

Different from the MHI [18] and PCH [34] representations, there is no damping factor to discriminate pixels over time

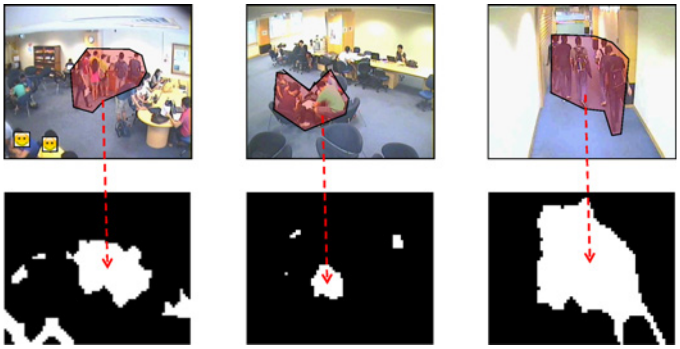

Fig. 3. Illustration of VA detected from sample video clip of *Queuing, Discussing, GroupWalking.*

because we assume that the temporal space is homogeneous in group activities. For example, *people walking to the right* and *people walking to the left* are considered the same activity as *people walking*. To eliminate the noise in the foreground such as reflection on a static object surface and minor luminance changes, we force the vigilance of a cube to be zero if it is smaller than $\eta\%$ of total $10 \times 10 \times \tau$ pixels (η is set to 1 in our work).

After obtaining the vigilance of all the cubes in a video clip, we normalize the vigilance into the range of [0, 1] and represent them in a gray-level image, called the vigilance image, where each pixel corresponds to a cube and its intensity is the vigilance of this cube. Suppose the video clip is divided in to $M \times N$ cubes; the resultant vigilance image is in size of $M \times N$. We then detect VAs from the vigilance image. Specifically, morphology operations such as dilate and erode are used to separate the connected components in the image. We remove components whose area in vigilance image is less than 8 pixels. The reason is that isolated small components are always caused by individual movement while connected large components are caused by a group activity. This threshold is empirically set to be small (i.e., 8 pixels) to alleviate the influence of the perspective distortion. As you can perceive, each component is our VA. The centroid of a VA indicates where the movement is and the contour describes what it looks like. Fig. 3 illustrates the VAs detected from sample video clips.

B. Quasi-ShapeContext Descriptor

Each VA detected above includes two types of information: vigilance and shape. Thus, the descriptor of VA should encode the shape of the contour and its internal layout while accounting for small local affine deformations. The affine deformation problem stems from different camera views and different moving pattern (e.g., walking vertically or horizontally in a view). Therefore, VAs representing the same group activity may be rotated and scaled.

Inspired by the shape context descriptor [40], we design a log-polar representation to be the VA descriptor. Given detected VAs in a video clip, we first calculate all the first principal components $\{\lambda_i\}$ of them and take the logarithm of the largest major component $\lambda^* = \max_i \lambda_i$ to be the upper bound of the log-axis. The upper bound of polar axis is set to 2π. For each VA, we transform it into the log-polar coordinates $[0, \log \lambda^*] \times [0, 2\pi]$ centered at the centroid and partition the space into 60 bins (12 angles and 5 radial intervals). Then, every pixel in the VA is allocated into its corresponding bin. The value of each bin is calculated as the maximal vigilance within it. As a result, each VA is described as a 60-dimensional descriptor. This vector is normalized by linearly stretching its values to the range of [0, 1] in order to be invariant to the differences in background.

The idea behind our quasi-ShapeContext descriptor is twofold.

1) In polar coordinates, it can be found that relative orientation and the logarithm of relative distance between points are invariant to scales. Therefore, the VA descriptor is invariant to different camera views.[2]
2) The maximal vigilance value is most informative on the motion within a bin and the uniform partition makes the descriptor more sensitive to the motions close to the centroid than those far away. Specifically, since the log function is more sensitive to the closer distance, the uniform bins record the motions in small area nearby the centroid and that in large area far away, and finally treat these motions equally. Therefore, motions nearby the centroid contribute more to its corresponding bins than motions far away.

IV. Multi-Camera Group Activity Detection

In this section, we introduce the proposed discriminative graphical model for multi-camera group activity detection. Our model aims to predict the group activity occurring in a set of video clips collected from multiple cameras in the same time interval. For the sake of simplicity, we use the term video bundle to refer to a set of video clips collected from multiple cameras in the same time interval in the remaining parts of this paper. The content of a video bundle can be represented by a set of VA descriptors extracted from the video clips within the bundle. Fig. 4 shows an illustration of video bundles and their corresponding VA descriptors. Different from the existing latent graphical models, such as HMM [41] and HCRF [42], which assume a predefined and fixed structure for the latent variables (e.g., chain structure or tree structure), the proposed discriminative graphical model optimizes the nonfixed structure of the latent variables during learning and inference. Through automatic inference of the nonfixed structure, the intra/inter-camera contexts are intelligently explored for activity detection.

Next, we will present the mathematical formulation of our discriminative graphical model, and show that this model captures the intra-camera and inter-camera contexts. We will then describe the inference and learning procedure of the proposed model.

A. Formulation

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the feature and label space, respectively. The training set is denoted by $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$. $\mathbf{x}_i \in \mathcal{X}$ is

[2] Note that spatial scaling in one single-camera view is not an important issue because the monitored area of a single camera is limited in our data. In some other applications such as far-field video surveillance, VAs can be extracted from multiple spatial scales.

a set of VAs $\{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \cdots, \mathbf{x}_{in}\}$ extracted from the ith video bundle, where n is the number of VAs in this bundle. The number of VAs can vary across bundles; for convenience of notation, we omit the dependency on the bundle index and simply refer to the number of VAs as n for every bundle. $y_i \in \{+1, -1\}$ is the corresponding label indicating whether this bundle captures a particular activity.

Given the training samples, the task is to learn a classification function $f : \mathcal{X} \rightarrow \mathcal{Y}$ from these samples. However, the label y_i associated with $\mathbf{x}_i$ does not directly indicate the latent semantics of VAs $\mathbf{x}_{ij}$. Thus, we introduce an intermediate hidden variable h_{ij} for each VA $\mathbf{x}_{ij}$, where $h_{ij} \in \mathcal{H}$ and $\mathcal{H}$ is a set of possible hidden states. This gives rise to a set of hidden variables $h_i = \{h_{i1}, \cdots, h_{in}\}$ for each sample $\mathbf{x}_i$. Intuitively, h_{ij} assigns a subactivity label to $\mathbf{x}_{ij}$, where the subactivity corresponds to certain motion patterns that are commonly shared with different activities. For example, one subactivity might correspond to *moving slightly in short-term but continuously in long-term* that is commonly observed in the activity *Studying*, *Discussing*, and *Queuing*. The subactivity corresponding to *radical moving* is shared by *GroupWalking* and *DoorOpening*. Therefore, such hidden variables implicitly capture the latent semantic meanings of the low-level features. There are connections among some hidden variables and these connections essentially encode the intra-camera and inter-camera contexts. Concretely, the connections among VAs from the same camera encode the intra-camera context, while the connections among VAs from different cameras characterize the inter-camera context. For a sample $\mathbf{x}$, we use an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent its hidden variables $h = \{h_1, h_2, \cdots, h_n\}$, where the vertices correspond to the hidden variables and the edges $(j, k) \in \mathcal{E}$ denote connections between variables h_j and h_k. As we will discuss later, the structure of graph $\mathcal{G}$, i.e., the connections between hidden variables, is automatically inferred.

Given the definitions of labels y, samples $\mathbf{x}$, hidden variables $\mathbf{h}$, and the context graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, we formulate our activity detection model as a linear discriminative function [30] as

$$F(\mathbf{x}, \mathbf{h}, y, \mathcal{G}; \mathbf{w}) = \mathbf{w}^T \Phi(y, \mathbf{h}, \mathbf{x}; \mathcal{G}) \tag{2}$$

where $\mathbf{w}$ is the model parameter and $\Phi(y, \mathbf{h}, \mathbf{x}; \mathcal{G})$ is a real valued feature function.

This model makes it feasible to incorporate the association between VAs and their hidden variables, the relation between hidden variables and labels, and the connection between hidden variables in a single unified formulation. To encode all of this information, we decompose the overall feature function $\Phi(y, \mathbf{h}, \mathbf{x}; \mathcal{G})$ into three components according to the relations between these variables

$$\begin{aligned} F(\mathbf{x}, \mathbf{h}, y, \mathcal{G}; \mathbf{w}) = \sum_{i \in \mathcal{V}} \mathbf{w}_{xh}^T \phi_{xh}(\mathbf{x}_i, h_i) \; + \sum_{i \in \mathcal{V}} \mathbf{w}_{hy}^T \phi_{hy}(h_i, y) \\ + \sum_{i, j \in \mathcal{E}} \mathbf{w}_{hh}^T \phi_{hh}(y, h_i, h_j) \end{aligned} \tag{3}$$

where ϕ_{xh} captures the association between each VA and its hidden variable, ϕ_{hy} associates the hidden variables to activity label, and ϕ_{hh} features the inter-relationship between each pair of linked hidden variables. The model parameters $\mathbf{w}$ are the combination of three parts, $\mathbf{w} = \{\mathbf{w}_{xh}, \mathbf{w}_{hy}, \mathbf{w}_{hh}\}$. In the following, we give the details of these three potentials.

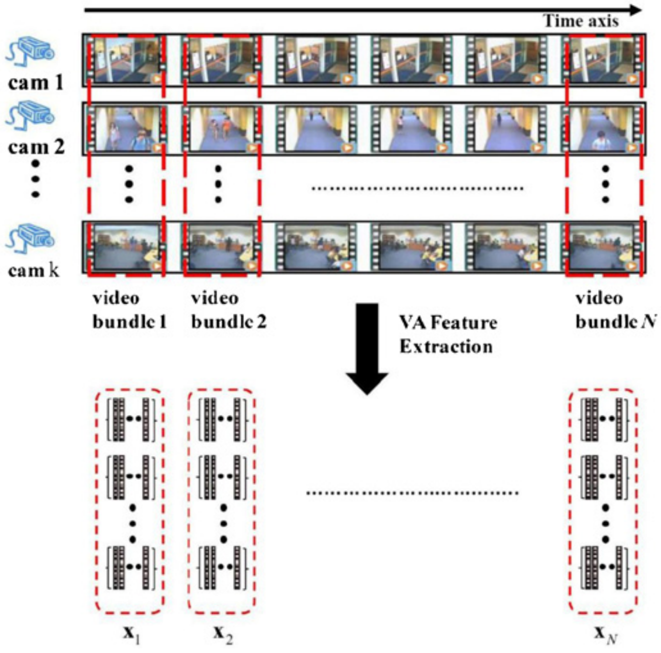

Fig. 4. Illustration of video bundles and their corresponding VA descriptors.

1) *Association Between a VA and Its Hidden Variable:* This potential models the hidden semantic of a VA. It is parameterized as

$$\mathbf{w}_{xh}^T \phi_{xh}(x_i, h_i) = \sum_{s \in \mathcal{H}} w_{xh,s}^T \cdot \delta [\![h_i = s]\!] \cdot \mathbf{x}_i \tag{4}$$

where $\delta [\![h_i = s]\!]$ is the indicator function that takes on value 1 if the argument holds true and 0 otherwise, $\mathcal{H}$ is the set of all possible hidden labels, and the parameters $\mathbf{w}_{xh}$ are the concatenation of $w_{xh,s}$ for all $s \in \mathcal{H}$. This potential function recharacterizes the VA descriptor $\mathbf{x}_i$ with a weighting factor $w_{xh,s}$ dominated by a hidden state h. Recall that each dimension of a VA descriptor $\mathbf{x}$ represents the vigilance of a part of vigilance area, the hidden label is expected to boost the vigilance values related to the corresponding subactivity while inhibiting those unrelated ones via the weight $\mathbf{w}_{xh}$.

2) *Relation Between Activity Label and Hidden Variables:* This potential is designed for modeling the compatibility between hidden variable h_i and activity label y

$$\mathbf{w}_{hy}^T \phi_{hy}(h_i, y) = \sum_{l \in \mathcal{Y}} \sum_{s \in \mathcal{H}} w_{hy,ls}^T \cdot \delta [\![h_i = s]\!] \cdot \delta \left[\!\left[y = l\right]\!\right]. \tag{5}$$

The relatedness of the hidden variables to the activity label y is recorded in the parameter $\mathbf{w}_{hy}$. For example, when we detect activity *Studying*, the hidden variables assigned to the VAs whose hidden states are the subactivity implying *slow but continuously moving* will contribute more potentials, i.e., have larger weights $\mathbf{w}_{hy}$ than those whose subactivity is *rapid moving*.

3) *Connection Between Hidden Variables:* As mentioned before, we encode the intra-camera and inter-camera contexts through the structure of the hidden variables, i.e., the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The intra-camera contextual relationship is reflected by the connections of the hidden variables corresponding to the VAs from the same camera, while the inter-camera contextual relationship is expressed by the connections of the hidden variables of VAs from different cameras. We parameterize the interaction between a pair of linked hidden variables by using the following potential function:

$$\begin{aligned} &\mathbf{w}_{hh}^T \phi_{hh}(y, h_i, h_j) \\ &= \sum_{l \in \mathcal{Y}} \sum_{p \in \mathcal{H}} \sum_{q \in \mathcal{H}} w_{hh,lpq}^T \cdot \delta \left[\left[h_i = p\right]\right] \cdot \delta \left[\left[h_i = q\right]\right] \cdot \delta \left[\left[y = l\right]\right]. \end{aligned} \tag{6}$$

Given a particular activity label y, this potential function favors the interactions among different hidden variables that reveal reasonable context.

B. Learning and Inference

We now first describe the inference of the optimal activity label y for a new sample given the model parameter $\mathbf{w}$, and then elaborate the learning of the model parameters $\mathbf{w}$ from training samples.

1) *Inference:* Suppose the model parameters $\mathbf{w}$ have been learned; the inference is to find the optimal activity label y for a new sample $\mathbf{x}$. In particular, the label y can be inferred as

$$y^* = \arg\max_{y \in \mathcal{Y}} F(\mathbf{x}, y, \mathbf{h}_y^*, \mathcal{G}_y^*; \mathbf{w}) \tag{7}$$

where $\mathbf{h}_y^*$ are the optimal states of hidden variables given a particular y, $\mathcal{G}_y^*$ is the optimal structure of the hidden variables given y. Note that the optimal structure $\mathcal{G}_y^*$ essentially interprets the intra-camera and inter-camera contexts. Specifically, $\mathbf{h}_y^*$ and $\mathcal{G}_y^*$ are inferred according to a particular y as

$$< \mathbf{h}_y^*, \mathcal{G}_y^* >= \arg\max_{\mathbf{h}_y, \mathcal{G}_y} F(\mathbf{x}, y, \mathbf{h}_y, \mathcal{G}_y; \mathbf{w}). \tag{8}$$

This crucial optimization problem is, in general, NP-hard because it involves a combinatorial search. In order to approximate the solution, we solve (8) using a coordinate descent algorithm similar to a latent support vector machine [43].

Step 1: Holding the graph structure $\mathcal{G}_y$ fixed, optimize the hidden variables for the pair $(\mathbf{x}, y)$

$$\mathbf{h}_y = \arg\max_{\mathbf{h}'} F(\mathbf{x}, y, \mathbf{h}', \mathcal{G}_y; \mathbf{w}). \tag{9}$$

Step 2: Holding $\mathbf{h}_y$ fixed, infer the graph structure $\mathcal{G}_y$

$$\mathcal{G}_y = \arg\max_{\mathcal{G}'} F(\mathbf{x}, y, \mathbf{h}_y, \mathcal{G}'; \mathbf{w}). \tag{10}$$

We use loopy belief propagation to optimize $\mathbf{h}_\mathbf{y}$ in (9). However, solving (10) is still NP-hard because the enumerations of all graph structures are exponential to the number of vertices. In order to tackle this problem, we need to make reasonable assumptions to approximate the exact solutions. We control the sparsity of the graph by a threshold d, where $d \in \{0, 2, ..., n-1\}$ refers to the degree of vertices [30]. Trivially, $d = 0$ indicates that the hidden variables are totally independent and $d = n - 1$ implies that the graph is fully connected. Generally, higher degree leads to a more complicate model, and may result in over-fitting and divergence in loopy belief propagation. Then, the optimization in (10) can be written as an integer linear programming

$$\begin{aligned} &\max_{\mathbf{z}} \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}} z_{ij} \mathbf{w}_{hh}^T \phi_{hh}(y, h_i, h_j) \\ &\text{s.t. } \forall i, j, \sum_{i \in \mathcal{V}} z_{ij} \le d, \sum_{i \in \mathcal{V}} z_{ij} \le d, z_{ij} = z_{ji}, z_{ij} \in \{0, 1\} \end{aligned} \tag{11}$$

where $z_{ij} = 1$ indicates $(i, j) \in \mathcal{E}$ and 0 otherwise. We further approximate the solution of (11) by relaxing constraint $z_{ij} \in \{0, 1\}$ to $z_{ij} \in [0, 1]$ and then rounding the solution to its closest integers.

After obtaining the optimal hidden states $\mathbf{h}_y^*$ and the optimal graph structure $\mathcal{G}_y^*$ for all possible y, we can optimize the label of a new sample $\mathbf{x}$ according to (7).

2) *Learning:* Now we describe how to learn the parameter $\mathbf{w}$ from a set of training samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$. The task of learning $\mathbf{w}$ can be formulated as a constrained optimization problem as

$$\begin{aligned} &\min_{\mathbf{w}, \xi_i} \tfrac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i \\ &\text{s.t. } \forall i, F(\mathbf{x}_i, y_i, \mathbf{h}_{y_i}^*, \mathcal{G}_{y_i}^*; \mathbf{w}) - F(\mathbf{x}_i, \bar{y}_i, \mathbf{h}_{\bar{y}_i}^*, \mathcal{G}_{\bar{y}_i}^*; \mathbf{w}) \ge 1 - \xi_i \\ &\quad\;\; \xi_i \ge 0 \end{aligned} \tag{12}$$

where $\bar{y}_i = \mathcal{Y} \backslash y_i$ and ξ_i is a penalty for margin violations of samples. C is a constant that controls the tradeoff between training error minimization and margin maximization.

The inequality constraints in (12) can be rewritten as the following equivalent equality constraints:

$$\forall i, \bar{\xi}_i = \max\{0, 1 - F(\mathbf{x}_i, y_i, \mathbf{h}_{y_i}^*, \mathcal{G}_{y_i}^*; \mathbf{w}) + F(\mathbf{x}_i, \bar{y}_i, \mathbf{h}_{\bar{y}_i}^*, \mathcal{G}_{\bar{y}_i}^*; \mathbf{w})\}. \tag{13}$$

Therefore, the constrained problem in (11) is equivalent to an unconstrained optimization problem as

$$\min_{\mathbf{w}} \tfrac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \max(0, \; 1 - F_{y_i}(\mathbf{w}) + F_{\bar{y}_i}(\mathbf{w})) \tag{14}$$

where $F_{y_i}(\mathbf{w}) = F(\mathbf{x}_i, y_i, \mathbf{h}_{y_i}^*, \mathcal{G}_{y_i}^*; \mathbf{w})$ and $F_{\bar{y}_i}(\mathbf{w}) = F(\mathbf{x}_i, \bar{y}_i, \mathbf{h}_{\bar{y}_i}^*, \mathcal{G}_{\bar{y}_i}^*; \mathbf{w})$.

Since $\bar{\xi}_i$ is nonconvex and nonsmooth, we use a nonconvex bundle method [44] to solve the optimization problem in (14). In order to apply the bundle method, we need to calculate the subgradient $g(\mathbf{w})$ of $\bar{\xi}_i$ in (13).

Calculating subgradient $g(\mathbf{w})$. The subgradient $g(\mathbf{w})$ of $\bar{\xi}_i$ is given as

$$g(\mathbf{w}) = \begin{cases} 0, \quad \text{if} \quad 1 - F_{y_i}(\mathbf{w}) + F_{\bar{y}_i}(\mathbf{w}) < 0 \\ \nabla_\mathbf{w} F_{\bar{y}_i}(\mathbf{w}) - \nabla_\mathbf{w} F_{y_i}(\mathbf{w}), \quad \text{otherwise.} \end{cases} \tag{15}$$

Next, we describe how to compute $\nabla_\mathbf{w} F_{y_i}(\mathbf{w})$, and $\nabla_\mathbf{w} F_{\bar{y}_i}(\mathbf{w}))$ can be calculated in a similar way.

Let $< \mathbf{h}_y^*(\mathbf{w}), \mathcal{G}_y^*(\mathbf{w}) >$ be the optimal solution of (8) explicitly parameterized by $\mathbf{w}$, then

$$\nabla_\mathbf{w} F_{y_i}(\mathbf{w}) = \Phi(y_i, \mathbf{h}_{y_i}^*, \mathbf{x}_i; \mathcal{G}_{y_i}^*) - \mathbf{w}^T \nabla_\mathbf{w} \Phi_{y_i}(\mathbf{w}) \tag{16}$$

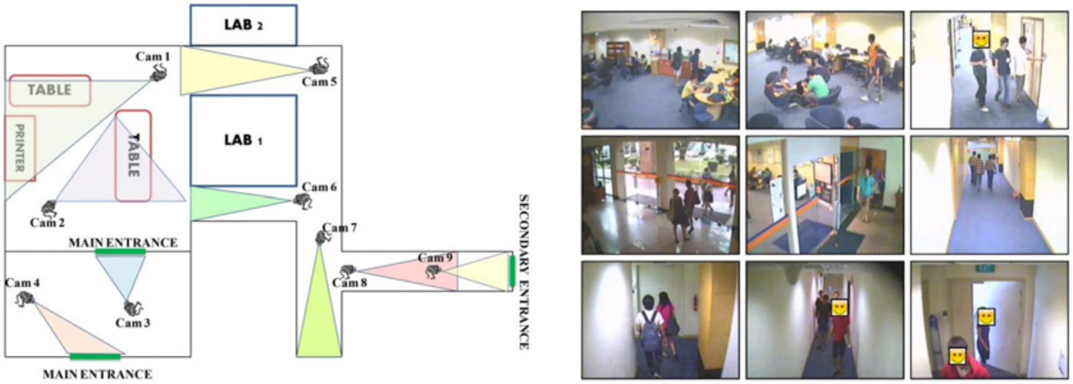

Fig. 5. Layout for the nine cameras and sample video frames from them.

where $\Phi_{y_i}(\mathbf{w}) = \Phi(y_i, \mathbf{h}^*_{y_i}(\mathbf{w}), \mathbf{x}_i; \mathcal{G}^*_{y_i}(\mathbf{w}))$. By the definition of derivative, we have

$$\nabla_{\mathbf{w}}\Phi_{y_i}(\mathbf{w}) = \lim_{||\Delta\mathbf{w}||\to\mathbf{0}} \frac{\Phi_{y_i}(\mathbf{w}+\Delta\mathbf{w}) - \Phi_{y_i}(\mathbf{w})}{\Delta\mathbf{w}} \tag{17}$$

where $\Phi_{y_i}(\mathbf{w}+\Delta\mathbf{w}) = \Phi(y_i, \mathbf{h}_{y_i}{}^*(\mathbf{w}+\Delta\mathbf{w}), \mathbf{x}_i; \mathcal{G}^*_{y_i}(\mathbf{w}+\Delta\mathbf{w}))$.

Note that $< \mathbf{h}^*_{y_i}, \mathcal{G}^*_{y_i} >$ are discrete and $\mathbf{w}$ is continuous. These properties assure that we can always find a sufficiently small $||\Delta\mathbf{w}|| \not= \mathbf{0}$ such that

$$< \mathbf{h}^*_{y_i}(\mathbf{w}+\Delta\mathbf{w}), \mathcal{G}^*_{y_i}(\mathbf{w}+\Delta\mathbf{w}) > = < \mathbf{h}^*_{y_i}(\mathbf{w}), \mathcal{G}^*_{y_i}(\mathbf{w}) > . \tag{18}$$

Therefore, the numerator on the left-hand side of (16) is zero and the subgradient $g(\mathbf{w})$ can be finally written as

$$g(\mathbf{w}) = \begin{cases} 0, & \text{if} \quad 1 - F_{y_i}(\mathbf{w}) + F_{\bar{y}_i}(\mathbf{w}) < 0 \\ \Phi_{\bar{y}_i}(\mathbf{w}) - \Phi_{y_i}(\mathbf{w}), & \text{otherwise} \end{cases} \tag{19}$$

where $\Phi_{\bar{y}_i}(\mathbf{w}) = \Phi(\bar{y}_i, \mathbf{h}^*_{\bar{y}_i}, \mathbf{x}_i; \mathcal{G}^*_{\bar{y}_i})$.

With the subgradient $g(\mathbf{w})$ of $\bar{\xi}_i$, the optimal model parameters $\mathbf{w}$ can be learned by solving the unconstrained optimization problem in (13) using the nonconvex bundle method provided by the subgradient.

V. Experiments

In this section, we evaluated our proposed VA feature and the discriminative graphical model for multi-camera group activity detection on two real-world CCTV surveillance video data sets, including a 468-h video corpus collected by nine cameras with a total footage duration of 468 h and the publicly available UCR VideoWeb data set [13], which contains 2.5 h of surveillance videos.

A. Data Set Description

468-h video data: This data set contains nine fixed and uncalibrated cameras set up on the first floor ceiling at an academic building located in a university campus. We collected videos by the cameras from 8:00 am to 9:00 pm in four days, and obtained $4 \times 9 \times 13 = 468$ h of videos in total. It is worth noting that our data set is larger than any other known surveillance video data set to our best knowledge (e.g., 50-h videos for training and 49-h videos for testing in TRECVid data set [8], 153-h videos in [9] and 177-h footage in [12]). The video stream has a size of 640 × 480 pixels at the frame rate of 4 f/s. As shown in Fig. 5, the coverage of the nine cameras includes one main entrance and one secondary entrance to the building, two corridors outside three research laboratories, and one public student studying hall. This data set is more challenging than the video corpus recorded from underground station. Besides the low frame rate, another challenge is that the crowd flow in our videos is not constant, and people may enter or exit from an unmonitored area freely. We define five group activities of interests, each involving behaviors of no less than three persons. The five activities are *Studying, Queuing, GroupWalking, DoorEntering, and Discussing*. The detailed descriptions are listed in Table I.

TABLE I
Description of Five Activities of Interest

Activity	Description
Studying	A group of students study at the table.
Queuing	A group of students queue for using a public resource.
GroupWalking	A group of students walk through an area.
DoorEntering	A group of students enter an area through a door.
Discussing	A group of students discuss with each other.

As illustrated in Fig. 4, we first divided the entire video streams into video bundles according to duration τ specified in Table II. Here, τ is the window size when detecting a given activity on a video stream. For example, for the activity *studying*, we set τ to 30 s and divided the video stream into a set of video bundles of 30 s durations. Then, we invited annotators to manually label the groundtruth of the activities on these bundles. In particular, τ is set by observing the duration of activities in the training samples. The approximately minimal and maximum durations of the five activities are 30 s and 3 h for *Studying*, 30 s and 10 min for *Queuing*, 15 s and 20 s for *GroupWalking*, 5 s and 10 s for *DoorEntering*, and 30 s and 2 h for *Disucssing*, respectively. We here set τ as the minimum duration of each activity to avoid including irrelevant content into the segmented video bundles that are positive samples with respect to a particular activity. If τ is set to be too large, for example, longer than the maximum duration of the activity, many positive video bundles will include much

TABLE II

DETAILS OF THE TRAINING DATA IN THE **468-H VIDEO** DATA SET

Activity	# of Positive	# of Negative	Duration τ
Studying	1153	11 530	30 s
Queuing	530	5772	30 s
GroupWalking	628	6420	15 s
DoorEntering	680	6800	5 s
Discussing	573	5800	30 s

irrelevant content. In other words, only a small portion of a positive bundle is truly positive. On the other hand, if τ is set as too small, for example, much shorter than the minimum duration of an activity, many positive samples will be not able to characterize the activity comprehensively. We chose the video bundles in the first three days for training and the rest as testing samples. We adopted the down-sample strategy to down-sample the negative samples in the training set. This strategy has been widely used to facilitate classifier learning in activity detection [45], [46]. Specifically, we filtered out negative samples that are completely static and then conducted temporal down-sampling on the negative samples to make the ratio between the numbers of positive and negative training samples at about 1:10. The details of the training data are listed in Table II.

UCR VideoWeb data: The Videoweb data set consists of 2.5 h of surveillance video. These videos are recorded by four to eight cameras over four days at 640 × 480 resolution with the frame rate of around 30 f/s. The average length of the video clips is 4 min. Most activities in this footage illustrate single-person actions (e.g., *Sitting down*), interactions between two persons (e.g., *Hugging*), interactions between person and object (e.g., *Tossing a ball*), and interactions with vehicles (e.g., *Entering car*). In our experiments, we evaluated three group activities, including *People milling together*, *Person joining group*, and *Person leaving group*. We conducted evaluation on the videos collected in the first three days. The data in the fourth day, which focus on the interactions with vehicles, were not used in our experiments. We separated the videos into segments with small duration and manually annotated the groundtruth over these segments. The annotated video segments were randomly split into a training set with 70% of segments and a testing set with 30% of segments.

B. Experimental Setting

To evaluate the proposed VA feature, we compared it to two existing spatiotemporal features, including the local spatiotemporal MoSIFT feature [7], which has achieved encouraging performance in TRECVid surveillance event detection, and the recently proposed dense trajectory feature [14]. MoSIFT treats a SIFT key point [47] as a MoSIFT key point if and only if this key point has sufficient motion (i.e., optical flow) in its corresponding scale space. The MoSIFT descriptor is a 256-dimensional vector that consists of 128-dimensional optical flow histogram after the original 128-dimensional SIFT descriptor. In our implementation, we compared the optical flow between every four frames and ruled out SIFT points with insufficient motion whose velocity is less than 1% of the diagonal of the frame in the corresponding scale space. The dense trajectory feature samples dense points from each video frame and tracks them based on displacement information from a dense optical flow field. In our implementation, 426-dimensional dense trajectory descriptors were extracted. We employed the k-means clustering algorithm to generate a 250-word visual vocabulary for VA and 500-word visual vocabularies for MoSIFT and dense trajectory features, respectively. The visual vocabularies were then used to generate bag-of-words (BoW) representation for each sample. Based on the BoW representation, the support vector machine (SVM) [48] was adopted as the activity detection model. For the sake of simplicity, we named these three methods VA-BoW-SVM, MoSIFT-BoW-SVM, and DT-BoW-SVM, respectively, where DT corresponds to the dense trajectory feature.

For the evaluation of the proposed discriminative graphical model (DGM), we compared our VA-DGM method to the VA-BoW-SVM approach and the existing graphical model HCRF [42]. The HCRF model was also built on the VA features for the sake of fair comparison. We termed this approach VA-HCRF. Furthermore, we compared the VA-DGM method to its single-camera counterpart VA-DGM-S and a state-of-the-art group activity modeling method based on a multi-observation hidden Markov model (MOHMM) [9]. In particular, VA-DGM-S processes the multiple camera views independently. For an activity of interest, VA-DGM-S predicts the presence of the activity on the samples (i.e., video bundles) from all the cameras independently. All the samples are then sorted together according to their confidence scores for calculating the performance metrics described in the next section.

For performance comparison, we adopted the widely used performance metric receiver operating characteristic (ROC) curve to evaluate and compare the above methods on each activity. The horizontal axis of ROC curve represents the false-positive rate (FPR) and the vertical axis represents the true-positive rate (TPR), which are calculated as

$$\begin{aligned} TPR &= TP/(TP + FN) \\ FPR &= FP/(TN + FP) \end{aligned} \tag{20}$$

where TP is the number of true-positive samples, FN is the number of false-negative samples, FP is the number of false-positive samples, and TN is the number of true-negative samples. Based on the ROC curve, we further computed the area under ROC curve (AUC) value. AUC describes the probability that a randomly chosen positive sample will be ranked higher than a randomly chosen negative sample.

The algorithmic parameters of the approaches were determined by a four-fold cross-validation on training set. For a fair comparison, the best results of all the methods were reported. Specifically, three parameters need to be estimated in our DGM model: the tradeoff parameter C in (12), the degree of vertices d in (11), and the number of hidden states. They were respectively selected from sets {10, 100, 200, 500}, {2, 4, 6, 8}, and {5, 10, 20} via a cross-validation process. Similarly, the tradeoff parameter C in SVM was selected from {10, 100, 200, 500}.

TABLE III
GROUP ACTIVITY DETECTION PERFORMANCE BY DIFFERENT APPROACHES IN TERMS OF AUC VALUES

Approach	*Studying*	*Queuing*	*GroupWalking*	*DoorEntering*	*Discussing*	Avg. AUC
VA-DGM	**0.81**	**0.74**	**0.80**	**0.92**	**0.80**	**0.82**
VA-DGM-S	0.80	0.69	0.75	0.84	0.75	0.77
VA-BoW-SVM	0.77	0.72	0.77	0.72	0.67	0.73
VA-HCRF	0.65	0.73	0.64	0.84	0.66	0.70
MOHMM	0.62	0.63	0.62	0.71	0.64	0.64
MoSIFT-BoW-SVM	0.63	0.68	0.65	0.63	0.63	0.64
DT-BoW-SVM	0.61	0.63	0.69	0.77	0.63	0.67

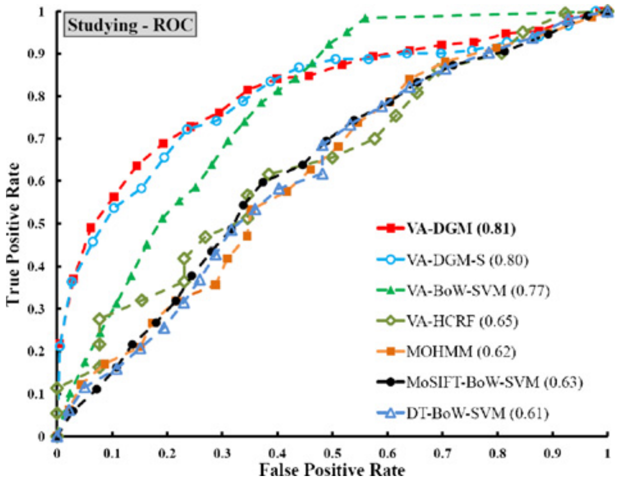

Fig. 6. ROC curves and AUC scores for the detection results of *Studying*.

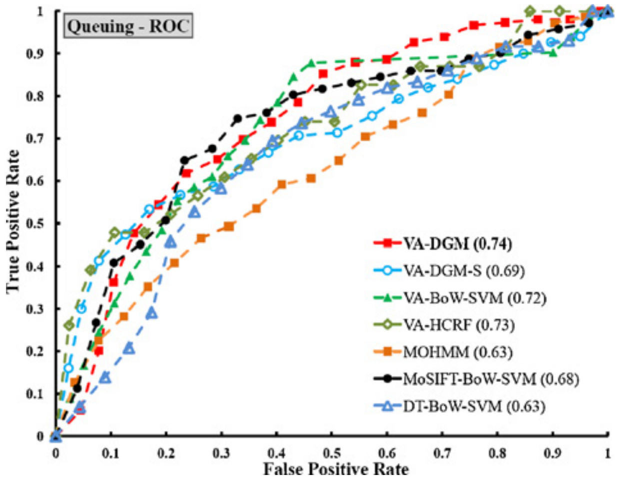

Fig. 7. ROC curves and AUC scores for the detection results of *Queuing*.

C. Experimental Results

1) *Evaluation on 468-h Video Data:* Figs. 6–10 illustrate the ROC curves of the seven methods on the five group activities in the 468-h video data, while Table III shows their AUC values on each activity and the average AUC values over all five of the activities. From these results, we can see that the proposed VA-DGM method achieves the best overall performance and outperforms others in all five of the activities.

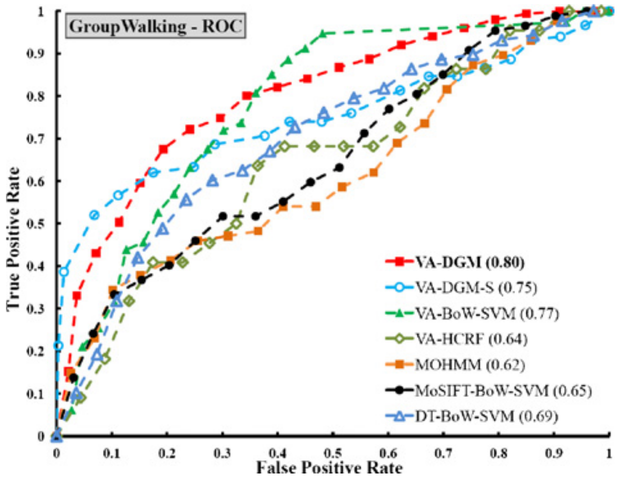

Fig. 8. ROC curves and AUC scores for detection results of *GroupWalking*.

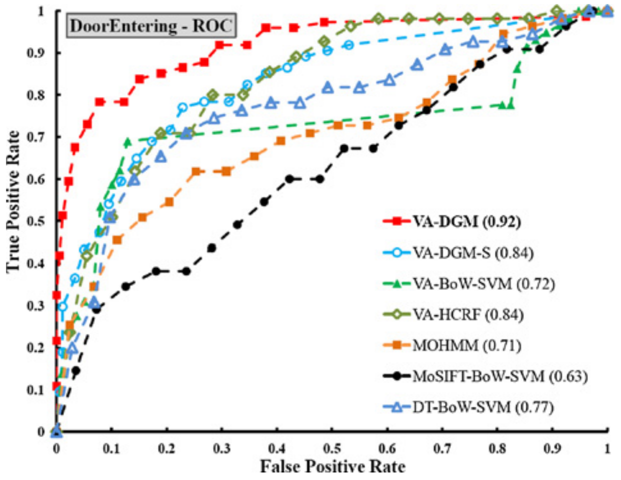

Fig. 9. ROC curves and AUC scores for detection results of *DoorEntering*.

Moreover, we can obtain the following observations.

1) *Evaluation of VA feature*: The superiority of VA-BoW-SVM to MoSIFT-BoW-SVM and DT-Bow-SVM indicates that the proposed VA feature is more effective than MoSIFT and the dense trajectory feature on group activity detection in real-world unconstrained videos. Compared to MoSIFT and the dense trajectory feature, the proposed VA feature achieves better overall performance and relative 14% and 9% improvements in terms of average AUC, respectively. Especially, for long-term group activities such as *studying*, the AUC value is relatively improved by 22% and 26% compared

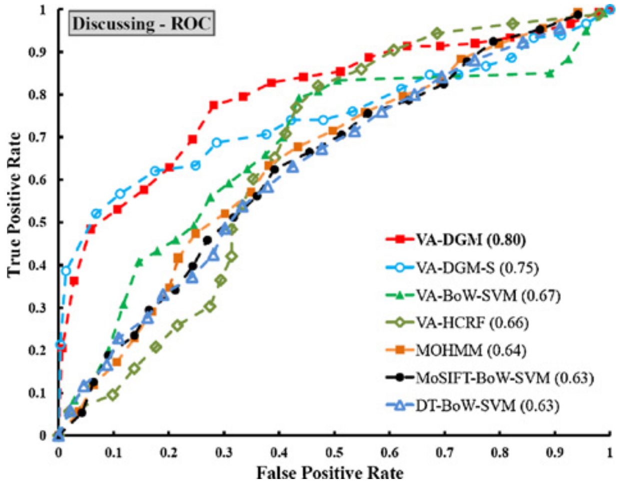

Fig. 10. ROC curves and AUC scores for the detection results of *Discussing*.

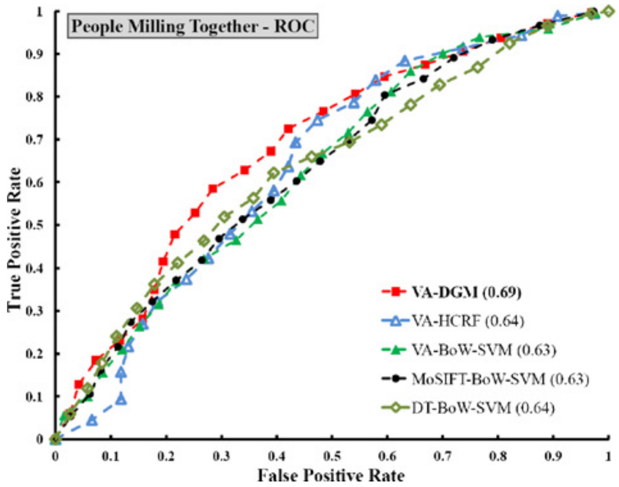

Fig. 11. ROC curves and AUC scores for the detection results of *People Milling Together*.

to MoSIFT and the dense trajectory feature, respectively. The main reason is that VA feature not only captures the accumulative motions of a group activity but also delineates the long-term appearance of the activity. In contrast, MoSIFT and the dense trajectory feature are based on local description that only encodes the instant appearance and motion of a local point. Therefore, MoSIFT and the dense trajectory feature are less descriptive and distinctive for long-term activities (e.g., *studying*) although they are effective for short-term events (e.g., CellToEar, pointing in TRECVid task [8], and *HandShake*, *Kicking* [14]). In addition, MoSIFT and the dense trajectory feature rely on the optical flow between two continuous frames. They are not effective in characterizing the motion information in videos with low frame rate, which lacks temporal consistency.

2) *Comparison to Existing Methods*: From the comparison results of the proposed VA-DGM approach and three existing methods, i.e., VA-BoW-SVM, MOHMM [9], and VA-HCRF, we can see that the VA-DGM approach obtains the best overall performance in terms of average AUC scores and outperforms the three existing method in all five of the activities. Different from HCRF, which assumes a predefined and fixed structure of latent variables, the proposed DGM does not fix the structure of latent variables, but optimizes it during learning and inference. Through automatic inference of the nonfixed structure, DGM is able to exploit the intra/inter-camera contexts more effectively. The experimental results show that DGM obtains about 17% relative improvements in terms of average AUC as compared to HCRF. Moreover, as described in Section I, most existing group activity detection methods rely on inter-camera topology inference based on accurate tracking results or clustering areas under different cameras with visual cues, such as the MOHMM method in [9]. These methods usually cannot perform well over unconstrained surveillance videos that contain a crowded scene and inconstant human crowd flow. In contrast, our proposed approach performs group activity detection without reliance on camera topology inference.

3) *Evaluation of intra-camera and inter-camera contexts*: While the VA-DGM-S method exploits the intra-camera context by using the DGM model, the VA-DGM approach leverages the intra-camera and inter-camera contexts at the same time. As shown in Table III, the improvements of VA-DGM-S over VA-BoW-SVM indicate that our DGM model performs better than the SVM with BoW representation. By modeling the intra-camera context, VA-DGM-S obtains about 5.5% improvements in terms of average AUC. From the comparison results of VA-DGM and VA-DGM-S, we can see that, by future exploiting inter-camera context, VA-DGM outperforms VA-DGM-S by around 6.5% on average AUC and performs better on all five of the activities. Specifically, the inter-camera and intra-camera contexts are interpreted by the optimized structure of hidden variables in DGM. For a testing sample, its VA descriptors are jointly modeled based on such context, where the states and structure of hidden variables are optimized on-the-fly to characterize the context. On the contrary, BoW representation of VAs is just a vector of term frequency that is assumed to be independent and hence loses important context information. These experimental results demonstrate that integrating multi-camera context within the proposed DGM method can effectively improve the activity detection performance.

2) *Evaluation on UCR VideoWeb Data:* This evaluation is targeted at studying the effectiveness of the proposed VA feature and DGM on a publicly available data set and investigating the comparison to existing spatiotemporal features and latent graphical models. Figs. 11–13 show the ROC curves of the three group activities on the UCR VideoWeb data set. The compared methods include the proposed VA-DGM method, an existing latent graphical model HCRF, two existing spatiotemporal features (i.e., MoSIFT and the DT feature), and the baseline method VA-BoW-SVM. Table IV provides the AUC values on each activity and the average AUC values.

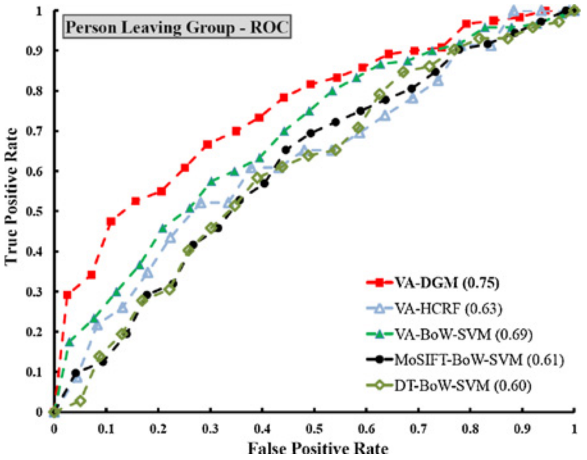

Fig. 12. ROC curves and AUC scores for the detection results of *Person Joining Group*.

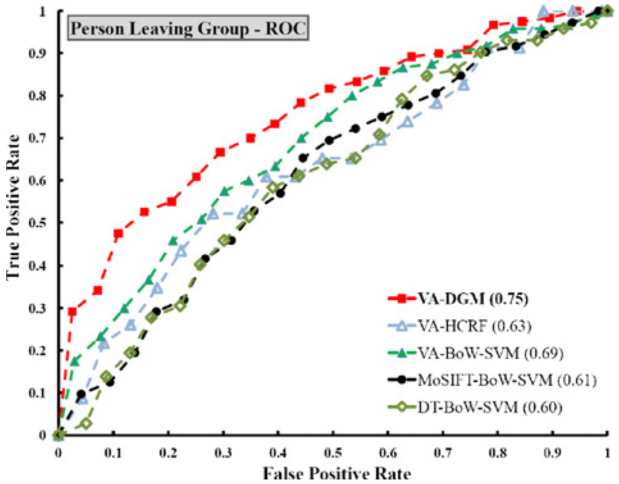

Fig. 13. ROC curves and AUC scores for the detection results of *Person Leaving Group*.

From the performance comparison between VA-DGM and VA-HCRF, we can see that the proposed DGM outperforms the existing HCRF model on all the three activities. In particular, DGM achieves 14% relative improvements over HCRF in terms of average AUC. As mentioned before, while HCRF adopts a fixed latent structure, the proposed DGM automatically optimizes the structure during learning and inference and thus can explore the intra/inter-camera contexts more effectively. On the other hand, from the comparison of VA-BoW-SVM to MoSIFT-BoW-SVM and DT-BoW-SVM, we can see that the VA feature is more effective than the two existing spatiotemporal features, i.e., MoSIFT and DT in group activity detection over real-world unconstrained surveillance videos.

VI. Conclusion and Future Work

In this paper, we proposed a new approach for detecting group activities in multi-camera surveillance videos. Different from previous works relying on camera topology inference, our proposed discriminative graphical model can simultaneously exploit intra-camera and inter-camera contexts for activity detection without topology inference. Moreover, an effective spatiotemporal feature VA was developed to represent the content of group activities. VA describes the quantity and appearance of the motion in an area. It is easy to extract from a dynamic and crowded scene without relying on any tracking. Experimental results on a 468-h real-world multi-camera surveillance video corpus and the publicly available UCR VideoWeb data set demonstrated the effectiveness of the proposed VA feature and discriminative graphical model on multi-camera group activity detection. Our proposed approach is expected to facilitate future research on group activity detection in two aspects: 1) the proposed VA feature provides an effective representation of group activity and is expected to facilitate group activity modeling and 2) the proposed DGM model exploits multi-camera context and has shown that this context is beneficial to group activity detection. This encourages future research on exploring multi-camera context. In the future, we will exploit more kinds of contextual information, e.g., scene layout, to improve the detection performance.

TABLE IV
Group Activity Detection Performance by Different Approaches in Terms of AUC Values on UCR VideoWeb Data Set

Approach	*Milling*	*Joining*	*Leaving*	Avg. AUC
VA-DGM	**0.69**	**0.73**	**0.75**	**0.72**
VA-HCRF	0.64	0.63	0.63	0.63
VA-BoW-SVM	0.63	0.66	0.69	0.66
MoSIFT-BoW-SVM	0.63	0.65	0.61	0.63
DT-BoW-SVM	0.64	0.66	0.60	0.63

The three group activities are *Milling—People Milling Together*, *Joining—Person Joining Group*, and *Leaving—Person Leaving Group*.

References

[1] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488, Nov. 2008.

[2] R. Poppe, "A survey on vision-based human action recognition," *Image Vision Comput.*, vol. 28, no. 6, pp. 976–990, 2010.

[3] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Comput. Survey*, vol. 43, no. 3, pp. 16:1–16:43, 2011.

[4] Y. Ke, R. Sukthankar, and M. Hebert, "Event detection in crowded videos," in *Proc. IEEE Int. Conf. Comput. Vision*, Jun. 2007, pp. 1–8.

[5] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 288–303, Feb. 2010.

[6] G. Zhu, M. Yang, K. Yu, W. Xu, and Y. Gong, "Detecting video events based on action recognition in complex scenes using spatiotemporal descriptor," in *Proc. ACM Int. Conf. Multimedia*, 2009, pp. 165–174.

[7] M. Chen and A. Hauptmann, "Mosift: Recognizing human actions in surveillance videos," Tech. Rep. CMU-CS-09-161, Carnegie Mellon University, Pittsburgh, PA, 2009.

[8] *TRECVid 2010 Evaluation for Surveillance Event Detection* [Online]. Available: http://www.itl.nist.gov/

[9] C. C. Loy, T. Xiang, and S. Gong, "Multi-camera activity correlation analysis," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2009, pp. 1988–1995.

[10] X. Wang, K. Tieu, and E. Grimson, "Correspondence-free activity analysis and scene modeling in multiple camera views," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 56–71, Jan. 2010.

[11] E. Zelniker, S. Gong, and T. Xiang, "Global abnormal behaviour detection using a network of CCTV cameras," in *Proc. IEEE Int. Workshop Visual Surveillance*, Oct. 2008, pp. 1473–1488.

[12] C. Loy, T. Xiang, and S. Gong, "Modelling activity global temporal dependencies using time delayed probabilistic graphical model," in *Proc. IEEE Int. Conf. Comput. Vision*, Sep. 2009, pp. 120–127.

[13] G. Denina, B. Bhanu, H. Nguyen, C. Ding, A. Kamal, C. Ravishankar, A. Roy-Chowdhury, A. Ivers, and B. Varda, "Videoweb dataset for multi-camera activities and non-verbal communication," Distributed Video Sensor Networks, 2011, pp. 325–347 2000.

[14] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2011, pp. 3169–3176.

[15] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. Huang, "Action detection in complex scenes with spatial and temporal ambiguities," in *Proc. Int. Conf. Comput. Vision*, 2009, pp. 128–135.

[16] M. Takahashi, M. Fujii, M. Shibata, and S. Satoh, "Robust recognition of specific human behaviors in crowded surveillance video sequences," *EURASIP J. Advances Signal Process.*, vol. 2010, p. 13, Feb. 2010.

[17] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatiotemporal features for action recognition," in *Proc. Brit. Mach. Vision Conf.*, 2009, pp. 1–9.

[18] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Comput. Vision Image Understanding*, vol. 104, nos. 2–3, pp. 249–257, 2006.

[19] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatiotemporal features," in *Proc. IEEE Workshop Visual Surveillance Performance Evaluation Tracking Surveillance*, Jan. 2005, pp. 65–72.

[20] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition," in *Proc. ACM Int. Conf. Multimedia*, 2007, pp. 357–360.

[21] O. Chomat and J. Crowley, "Probabilistic recognition of activity using local appearance," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 1999, pp. 36–41.

[22] L. Zelnik-Manor and M. Irani, "Event-based analysis of video," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2001, pp. 123–130.

[23] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space–time shapes," in *Proc. IEEE Int. Conf. Comput. Vision*, Oct. 2005, pp. 2247–2253.

[24] I. Laptev and T. Lindeberg, "Space–time interest points," in *Proc. IEEE Int. Conf. Comput. Vision*, Oct. 2003, pp. 432–439.

[25] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *Proc. IEEE Int. Conf. Comput. Vision*, Jun. 2005, pp. 432–439.

[26] M. S. Ryoo and J. K. Aggarwal, "Spatiotemporal relationship match: Video structure comparison for recognition of complex human activities," in *Proc. Int. Conf. Comput. Vision*, 2009, pp. 984–989.

[27] L. Ballan, M. Bertini, A. D. Bimbo, L. Seidenari, and G. Serra, "Recognizing human actions by fusing spatiotemporal appearance and motion descriptors," in *Proc. Int. Conf. Image Process.*, 2009, pp. 3533–3536.

[28] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2009, pp. 2929–2936.

[29] B. Ni, S. Yan, and A. Kassim, "Recognizing human group activities with localized causalities," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2009, pp. 1470–1477.

[30] T. Lan, Y. Wang, W. Yang, and G. Mori, "Beyond actions: Discriminative models for contextual group activities," in *Proc. Advances Neural Inform. Process. Syst.*, 2010, pp. 1216–1224.

[31] W. Choi, K. Shahid, and S. Savarese, "What are they doing? Collective activity classification using spatiotemporal relationship among people," in *Proc. Int. Conf. Comput. Vision Workshops*, 2009, pp. 1282–1289.

[32] J. Niebles, C. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *Proc. Eur. Conf. Comput. Vision*, 2010, pp. 392–405.

[33] G. Zhu, S. Yan, T. Han, and C. Xu, "Generative group activity analysis with quaternion descriptor," in *Proc. Int. Conf. Multimedia Modeling*, 2011, pp. 1–11.

[34] T. Xiang and S. Gong, "Activity based surveillance video content modelling," *Pattern Recognit.*, vol. 41, no. 7, pp. 2309–2326, 2008.

[35] M. S. Ryoo and J. K. Aggarwal, "Stochastic representation and recognition of high-level group activities," *Int. J. Comput. Vision*, vol. 93, no. 2, pp. 183–200, 2011.

[36] K. Tieu, G. Dalley, and W. Grimson, "Inference of non-overlapping camera network topology by measuring statistical dependence," in *Proc. Int. Conf. Comput. Vision*, 2005, pp. 1842–1849.

[37] D. Makris, T. Ellis, and J. Black, "Bridging the gaps between cameras," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2004, pp. 205–210.

[38] M. Chang, N. Krahnstoever, S. Lim, and T. Yu, "Group level activity recognition in crowded environments across multiple cameras," in *Proc. Int. Conf. Advanced Video Signal Based Surveillance*, 2010, pp. 56–63.

[39] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognit. Lett.*, vol. 27, no. 7, pp. 773–780, 2006.

[40] S. Salve and K. Jondhale, "Shape matching and object recognition using shape contexts," in *Proc. IEEE Int. Conf. Comput. Sci. Inform. Technol.*, 2010, pp. 471–474.

[41] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 1992, pp. 379–385.

[42] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fileds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1848–1852, Oct. 2007.

[43] C. Yu and T. Joachims, "Learning structural SVMS with latent variables," in *Proc. Int. Conf. Mach. Learning*, Jun. 2009, pp. 1169–1176.

[44] T. Do and T. Artières, "Large margin training for hidden Markov models with partially observed states," in *Proc. Int. Conf. Mach. Learning*, 2009, pp. 265–272.

[45] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce, "Automatic annotation of human actions in video," in *Proc. Int. Conf. Comput. Vision*, 2009, pp. 1491–1498.

[46] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, "Semantic model vectors for complex video event recognition," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 88–101, 2012.

[47] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[48] C.-C. Chang and C.-J. Lin. (2001). *Libsvm: A Library for Support Vector Machines* [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/libsvm

Zheng-Jun Zha (M'08) received the B.E. and Ph.D. degrees from the Department of Automation, University of Science and Technology of China, Hefei, China.

He is currently a Senior Research Fellow with the School of Computing, National University of Singapore, Singapore. His current research interests include multimedia content analysis, computer vision, social media analysis, and large-scale media search. He has authored more than 80 book chapters, journal articles, and conference publications in these areas, including TMM, TOMCCAP, TIP, TCSVT, ACM MM, CVPR, and SIGIR.

Dr. Zha received the Microsoft Research Fellowship in 2007, the President Scholarship of the Chinese Academy of Science in 2009, and the Best Paper Award in the 17th ACM International Conference on Multimedia in 2009. He is a member of the ACM.

Hanwang Zhang received the B.E. degree in computer science and technology (with CKC Hons. degree) from Zhejiang University, Hangzhou, China, in 2009. He is currently pursuing the Ph.D. degree with the School of Computing, National University of Singapore, Singapore.

His main research interests include multimedia and computer vision, developing techniques for efficient search, and recognition in image contents.

Meng Wang (M'07) received the B.E. and Ph.D. degrees from the Special Class for the Gifted Young and the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China, respectively.

He is currently a Professor with the Hefei University of Technology, Hefei. He was previously an Associate Researcher with Microsoft Research Asia, Beijing, China, and then a Core Member in a startup in Silicon Valley. After that, he was with the National University of Singapore, Singapore, as a Senior Research Fellow. His current research interests include multimedia content analysis, search, mining, recommendation, and large-scale computing. He has authored more than 100 book chapters, journal and conference papers in these areas, including TCSVT, TMM, TOMCCAP, ACM MM, WWW, and SIGIR.

Dr. Wang received the Best Paper Award in the 17th and 18th ACM International Conference on Multimedia and the Best Paper Award in the 16th International Multimedia Modeling Conference. He is a member of the ACM.

Huanbo Luan received the B.E. degree from Shandong University, Shandong, China, in 2003, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2008.

He is currently a Senior Research Fellow with the School of Computing, National University of Singapore, Singapore. His current research interests include multimedia analysis, mining, and retrieval.

Tat-Seng Chua is currently the KITHCT Chair Professor with the School of Computing, National University of Singapore (NUS), Singapore. He was the Acting and Founding Dean of the School of Computing from 1998 to 2000. He joined NUS in 1983, and spent three years as a Research Staff Member with the Institute of Systems Science (now I2R) in the 1980s. He has worked on several multi-million-dollar projects interactive media search, local contextual search, and real-time live media search. His main research interests include multimedia information retrieval, multimedia question–answering, and the analysis and structuring of user-generated contents.

Dr. Chua has organized and served as a Program Committee Member of numerous international conferences in the areas of computer graphics, multimedia, and text processing. He was the Conference Co-Chair of ACM Multimedia in 2005, the Conference on Image and Video Retrieval in 2005, and ACM SIGIR in 2008, and the Technical PC Co-Chair of SIGIR in 2010. He serves on the editorial boards of the *ACM Transactions of Information Systems* (ACM), *Foundation and Trends in Information Retrieval* (Now), *The Visual Computer* (Springer Verlag), and *Multimedia Tools and Applications* (Kluwer). He is on the Steering Committees of the International Conference on Multimedia Retrieval, Computer Graphics International, and Multimedia Modeling Conference Series. He serves as a member of international review panels of two large-scale research projects in Europe. He is the Independent Director of two listed companies in Singapore.