Graph formulation of video activities for abnormal activity recognition

Dinesh Singh[a,*], C. Krishna Mohan[a]

[a] *Visual Learning and Intelligence Group (VIGIL), Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, Kandi, Sangareddy-502285, India*

Abstract

Abnormal activity recognition is a challenging task in surveillance videos. In this paper, we propose an approach for abnormal activity recognition based on graph formulation of video activities and graph kernel support vector machine. The interaction of the entities in a video is formulated as a graph of geometric relations among space-time interest points. The vertices of the graph are spatio-temporal interest points and an edge represents the relation between appearance and dynamics around the interest points. Once the activity is represented using a graph, then for classification of the activities into normal or abnormal classes, we use binary support vector machine with graph kernel. These graph kernels provide robustness to slight topological deformations in comparing two graphs, which may occur due to the presence of noise in data. We demonstrate the efficacy of the proposed method on the publicly available standard datasets viz; UCSDped1, UCSDped2 and UMN. Our experiments demonstrate high rate of recognition and outperform the state-of-the-art algorithms.

Keywords: Abnormal Activity Recognition, Video Activity Classification, Graph Representation of Video Activity, Graph Kernel, Bag-of-Graphs (BoG)

1. Introduction

Nowadays digital video surveillance systems are ubiquitously deployed in public places for safety purpose. According to the British Security Industry Association (BSIA), approximately 4 million to 5.9 million cameras are deployed in UK [1]. This widespread use of surveillance systems in roads, stations, airports or malls has led to a huge amount of data that needs to be analyzed for safety, retrieval or even commercial reasons [2]. Anomalous event detection in crowded scenes is very important, e.g. for security applications, where it is difficult even for trained personnel to reliably monitor scenes with dense crowd or videos of long duration [2]. An anomalous event in a crowd is an event which do not confirm the

*I am corresponding author

Email addresses: cs14resch11003@iith.ac.in (Dinesh Singh), ckm@iith.ac.in (C. Krishna Mohan)

URL: cse.iith.ac.in/profile/phd/dsingh/ (Dinesh Singh)

normal appearance or dynamics of crowd. An appearance-related anomaly would be, e.g. a bicycle passing through a crowd. Moreover, sudden changes in velocity, like an abrupt increase of its magnitude and the dispersion of individuals in the crowd indicates that something unusual and potentially dangerous may have occurred [2].

In order to detect abnormal activities in surveillance videos or crowd behavior analysis, various kinds of activity modeling are proposed in the literature [3, 4, 5, 6, 7, 8, 9]. The existing models consider the object motion as the key factor for activity representation. The popular motion representation techniques are based on trajectory modeling, flow modeling, or vision based. The widely used bag-of-words (BOW) approaches [10, 11, 12] show excellent performance in action and activity recognition. A bag-of-words (BOW) approach computes a unordered histogram of visual words occurrences that encodes only the global distribution of low level descriptors, while it ignores the local structural organization (i.e. geometry) of salient points and corresponding low level descriptors. However, use of such local structure of salient points and corresponding low level descriptors should lead to discriminative video representation which further leads to better recognition of video activities.

In this work, we propose a framework for abnormal activity recognition which includes appearance and dynamics along with geometric relationships among various interactions of the entities in a video activity. First, we extract the space-time interest points and treat each interest point as a node of a graph. The edges of the graph are determined using a fuzzy membership function on the basis of closeness and the similarity of the entities associated with the interest points. If two points are close to each other, then there is a high probability that some interactions take place between the corresponding entities. In order to keep track of the objects, we also incorporate the appearance and motion of the entities using histogram of oriented gradients (HOG) and histogram of oriented optical flow (HOF). Second, a maximum margin classifier is trained on the basis of geometrical structure of the graphs formed for normal and abnormal training videos. The graph kernels are used for measuring the similarity between two graphs. The graph kernels provide robustness to the slight topological deformation in comparing two graphs because they measure on the basis of similar paths/walks. These deformations may occur due to various affecting factors like presence of noise in data. The idea of formation of video activity as a graph and use of graph kernel for their similarity measure is novel for abnormal activity recognition in surveillance videos. Finally, the combined approach provides a robust framework for the recognition of abnormal activities in surveillance videos. The experiments demonstrate the superiority of the proposed work over the existing methods which are based on dense trajectories and bag-of-words with various feature descriptors.

The rest of the paper is organized as follows: Section 2 presents related work. Section 3 describes the proposed approach for abnormal activity recognition. Section 4 discusses the experimental setup, datasets and results. The conclusions are provided in section 5.

2. Related Work

In the past decade, a considerable amount of literature is focused on the abnormal activity recognition in surveillance videos [3, 4, 5, 13, 14, 15]. The detailed surveys in [6, 7] enlighten

the progress on this topic in last decades. Wu *et al.* [14] model normal crowd patterns using chaotic invariants of Lagrangian particle trajectories based on optical flow. Saligrama *et al.* [15], presented a probabilistic framework for local anomaly detection by assuming that they are infrequent with respect to their neighbors while did not consider the relationship among local observations. Some tracking based techniques for video representation that extract trajectories of the moving objects are proposed in [11, 16]. Wang *et al.* [11] describe videos using dense trajectories that encode the shape of the trajectory, the local motion, and appearance around the trajectory. Wang *et al.* later in [16] present an improved trajectory that also takes into account camera motion. Yuan *et al.* [17] focus on different motion properties (viz; magnitude and direction) in order to detect different crowd abnormalities. They exploit the contextual evidences using structural context descriptor (SCD) to describe the relationship of the individuals, which is a concept of solid-state physics. Then the anomaly is detected by finding the large variation of SCD between newly observed frame and the previous ones. The targets in different frames are associated using a robust 3-D DCT multi-object tracker. However, it tracks only a few observers instead of analyzing the trajectories during dense crowd. The trajectory based modeling of activities is ubiquitous but unreliable in the situations where crowded scenes are present.

Some of the non-tracking based techniques are also proposed which include dense optical flow, or some other form of spatio-temporal gradients [18, 19, 20]. Reddy *et al.* proposed an algorithm to detect anomalies by inspecting motion, size and texture information [18]. It estimates object motion more precisely by computing optical flow, only for the foreground pixels. Motion and size features are modeled in small cells using computationally efficient approximated kernel density estimation technique and texture is represented using adaptively grown vocabulary. Loy *et al.* used Gaussian process regression (GPR) for multi-object activity modeling [19]. The non-linear relationship between decomposed image regions is formulated as a regression problem. It is better to characterize spatial configurations between objects, as it predicts the behavior of current region based on its past complements. However, it is unable to handle complex causalities in video scenes.

The approaches in [20, 2] consider both appearance based (spatial) and motion based (temporal) anomalies. Mahadevan *et al.* [20] proposed mixtures of dynamic textures (MDT) model to detect temporal and spatial abnormalities from unconstrained scenes. These approaches flag abnormal events based on independent location-specific statistical models but the relationship between local observations is not taken into consideration. Kaltsa *et al.* [2] incorporate swarm theory with histogram of oriented gradients (HOG) for detecting and localizing anomalous events in videos of crowded scenes. Where both motion and appearance information are considered. While histograms of oriented swarms (HOS) capture the dynamics of crowded environments, the histogram of oriented gradients (HoG) capture the appearance information. The descriptor build by combining HOS and HOG effectively characterizes each scene. The appearance and motion features extracted only within spatio-temporal volumes of moving pixels ensure robustness to local noise, increase accuracy in the detection of local, non-dominant anomalies, and achieve a lower computational cost. Kim *et al.* give space-time Markov random field (MRF) model for abnormal activity recognition in videos [21]. The nodes in the MRF graph are the grid of local regions of the video frames

where the neighbors in space and time are associated with links. At each local node, distribution of optical flow is captured to generate the model of normalcy using mixture of probabilistic principal component analyzers (MPPCA). The degree of normality of an incoming video clip is decided using learned model and MRF graph. An incremental approach is used to deal with the concept drift.

The most recent methods focus on both appearance and motion anomalies at local and global scale. Space-time interest points have been explored recently for abnormal activity recognition in surveillance videos [10]. In [10], Cheng *et al.* detect local and global anomalies via hierarchical feature representation using bag-of-visual-words (BoVW) and Gaussian process regression. The extraction of normal interactions from training videos is formulated as the problem of efficiently finding the frequent geometric relations of the nearby sparse space-time interest points (STIPs). In [11, 16], Wang *et al.* use a standard bag-of-features approach to construct separate vocabularies of 4000 visual-words for each type of low level descriptors. The low level descriptors encode information of trajectory shape, appearance using HoG, local motion using HoF, and gradient of horizontal, and vertical components of optical flow using motion boundary histograms (MBH). However, above methods use bag-of-words based approach that do not consider the geometric relationships among salient points.

Sekma *et al.* [22] used bag-of-graphs (BoG) for human action recognition that exploits the geometric relationships among trajectories. The assumption made is that entities that are related spatially are usually dependent on each other. The neighbour trajectory points are linked using Delaunay triangulation method which is invariant to affine transformations like scaling, rotation and translation. The Hungarian distance method is used for graph matching. A separate bag-of-graphs (BoG) is applied for each low level descriptor (HoG,HoF,MBH and Trajectory Shape) with 3 graph scales (3-nearest neighbour,6-nearest neighbour,9-nearest neighbour) resulting into 12 histograms that are concatenated after applying sum pooling and L_1 normalization. For classification, support vector machine is used. However, the proposed method is significantly different from this method. The first difference is in the process of generating graphs, where an edge between two points is decided through a fuzzy membership function instead of considering fixed nearest neighbours. Secondly, instead of using Hungarian distance which finds the similar set of low level descriptors, we use graph kernels for matching two graphs. The graph kernels find the similarity between two graphs which are more robust against the affine transformations as well as slight geometric deformations.

3. Proposed Work

The proposed framework for abnormal activity recognition in surveillance videos is presented in this section. The block diagram of the proposed framework is shown in Fig. 1. The proposed framework consists of three steps. In the first step, the incoming video feed is split into video clips of size T and space-time interest points in each video clip are extracted. In the second step, a set of undirected graphs of local activities is generated. The vertices of the graphs are space-time interest points and an edge represents a possible interaction. In the

third step, each activity is classified into normal or abnormal categories. Which is further classified into local abnormal activity recognition and global abnormal activity recognition. For local activity classification a max-margin classifier is trained using graph kernel SVM from training videos. For global activity recognition, bag-of-graphs (BoG) feature vectors are generated for a set of local activity graphs and a support vector machine model trained from BoG feature vectors from training videos is used to declare the global behavior. Each of these steps are discussed in detail below:

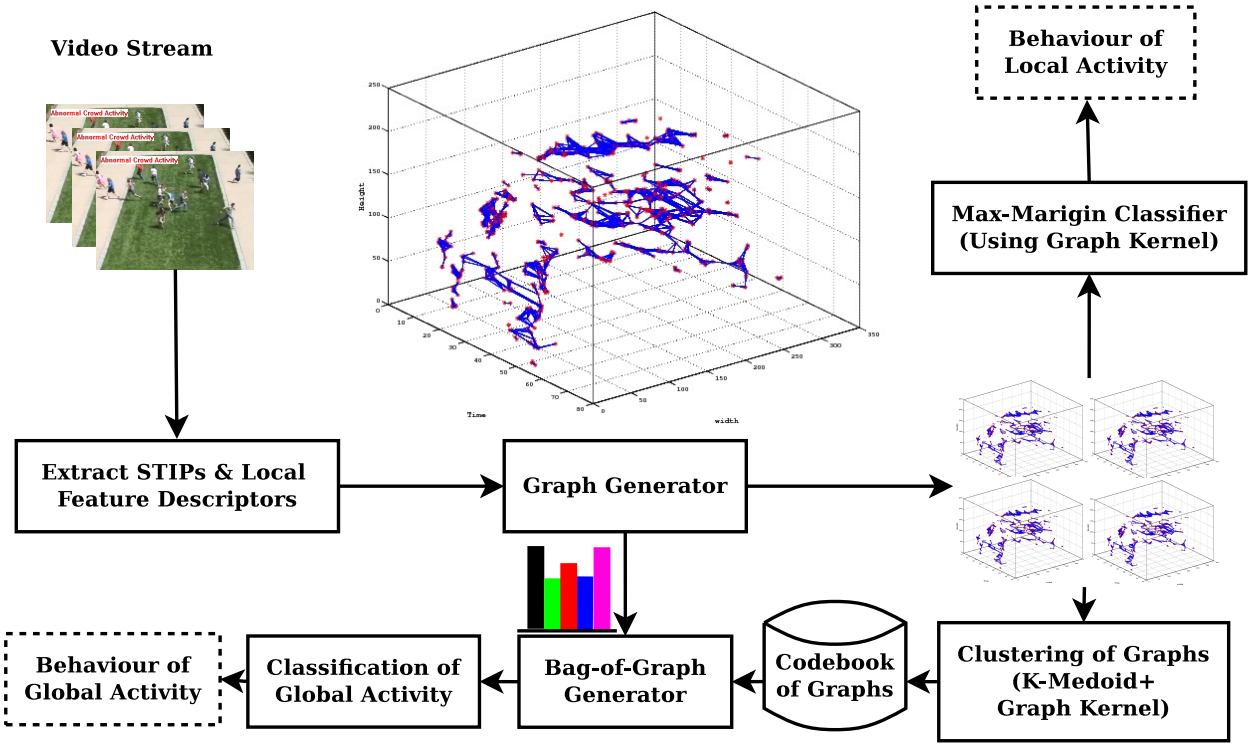


Figure 1: Block diagram of the proposed framework for abnormal activity recognition in surveillance videos.

3.1. Detection of space-time Interest Points

The space-time interest points [23] are salient points, which are the regions in $f : \mathbb{R}^2 \times \mathbb{R} \rightarrow \mathbb{R}$ having significant eigenvalues λ_1, λ_2, and λ_3 of a spatio-temporal second-moment matrix μ, which is a 3-by-3 matrix composed of first order spatial and temporal derivatives averaged using a Gaussian weighting function $g(.; \sigma_i^2, \tau_i^2)$ with integration scales σ_i^2 (spatial variance) and τ_i^2 (temporal variance). The value of μ is computed as

$$\mu = g(.; \sigma_i^2, \tau_i^2) * \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix}, \tag{1}$$

where L_x, L_y, and L_t are first-order derivatives with respect to x, y, and t of the linear scale-space representation $L : \mathbb{R}^2 \times \mathbb{R} \times \mathbb{R}_+^2 \rightarrow \mathbb{R}$ of f constructed by convolution of f

with an anisotropic Gaussian kernel $g(.;\sigma_l^2,\tau_l^2)$ with local scales σ_l^2 (spatial variance) and τ_l^2 (temporal variance). The value of the L is computed as

$$L(.;\sigma_l^2,\tau_l^2) = g(.;\sigma_l^2,\tau_l^2) * f(.). \tag{2}$$

These interest points are detected using Harris3D corner function (H) for the spatio-temporal domain by combining the determinant (det) and the trace of μ ($trace$) as follows:

$$H = det(\mu) - k\ trace^3(\mu)$$

$$H = \lambda_1 \lambda_2 \lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3, \tag{3}$$

where k is a constant. Then around each salient point $p(w,h,t)$, 72-dimensional HOG [24] and 90-dimensional HOF [25] descriptors are extracted, which together represent an interest point in the 3D space by a 162-dimensional feature vector $\mathbf{f} = \mathbb{R}^{162}$ called STIP descriptor. In this way, the STIP feature descriptors include the appearance information using HoG and motion information using HoF around the salient points. Section 3.2 presents the process of graph generation.

3.2. Graph Formulation of a Video

In previous step, we obtain a set of space-time interest points $\mathbf{P} = \{\mathbf{p}_i | \mathbf{p}_i \in (x,y,t)_{i=1}^n$ and their respective feature vectors $\mathbf{F} = \{\mathbf{f}_i\}_{i=1}^n$ for a given video. In this step, we represent the video as a graph $G(\mathbf{P},\mathbf{E})$, where $\mathbf{P}$ is the set of space-time interest points detected from previous step and $\mathbf{E}$ is the set of edge. An edge between two points $\mathbf{p}_i$ and $\mathbf{p}_j$ is decided based on μ_{ij}, which is a fuzzy membership score of the edge existence and is computed as

$$\mu_{ij} = \frac{K(\mathbf{f}_i,\mathbf{f}_j)}{||\mathbf{p}_i - \mathbf{p}_j||_2}, \tag{4}$$

where, $K(\mathbf{f}_i,\mathbf{f}_j)$ is the similarity measure between the feature vectors $\mathbf{f}_i$ and $\mathbf{f}_j$ extracted at points $\mathbf{p}_i$ and $\mathbf{p}_j$, respectively. Any geometric kernel function can be used as a similarity measure like linear kernel, polynomial kernel, RBF kernel, or sigmoid kernel. This similarity is high if feature vectors $\mathbf{f}_i$ and $\mathbf{f}_j$ are belonging to similar events and/or similar object. This shows that these points are either too close to each other so that they share lot of information during feature extraction or over the time, object at point $\mathbf{p}_i$ moved to point $\mathbf{p}_j$. The later case is significant for modelling an activity, so geometric distance $||\mathbf{p}_i - \mathbf{p}_j||_2$ between points $\mathbf{p}_i$ and $\mathbf{p}_j$ is in the denominator. Due to this, the value of μ_{ij} becomes high for points which are very close and too low for points at far distance and in both the cases we do not get significant information. However, for the points at some distance, a high value of μ_{ij} shows significance towards the existence of an event between these points. Thus, if the value of μ_{ij} is explicitly high, then we consider them as similar point and represent them using a single point which is the mid point of these points. And if the value of μ_{ij} is explicitly low, then there will be no edge. The adjacency matrix $\mathbf{A}$ of the grapg Gcan be written as

$$A_{ij} = \begin{cases} 0, & if \mu_{ij} < \mu_T\ Threshold \\ 1, & otherwise \end{cases} \tag{5}$$

Fig. 2(a) shows an adjacency matrix $\mathbf{A}$ of the graph generated from an abnormal video where people are running abruptly in all directions. The space-time interest points are arranged according to their location in 3D cube while traversing along the direction x followed by y followed by t. The black dot at location A_{ij} indicates an edge between salient points p_i and p_j. The more number of black dots around diagonal in adjacency matrix confirms that as the distance between two salient points increases, the possibility of an edge between them decreases (see Fig. 2(d)). Fig. 2(b) shows the cube of graphs corresponding to the adjacency matrix $\mathbf{A}$. Each isolated graph shown in the cube corresponds to a local action/activity in the video. The individual local activity may belong to some kind of abnormality, or a group of these local activities together may correspond to an abnormal activities. Fig. 2(c) illustrates the behaviour of an edge existence membership function, where it can be observed that the frequency of points with low membership value is high while the frequency of the high membership value is very low.

3.3. Activity Recognition

Once the video activities are represented using graphs then the next task is to classify them as normal activity or abnormal activity. This section presents framework for detecting both local and global activities.

3.3.1. Recognition of Local Abnormal Activities.

A surveillance video may contain multiple local activities occurring simultaneously. Each local activity can be represented using a graph. A max-margin classifier is trained from the collection of all the local activity graphs from the training videos. Then this classifier is used to predict the behaviour of the local activities in the test videos.

Let $\{G_i, y_i\}_{i=1}^{n}$ be the corresponding labeled graphs for n activities $\{A_i\}_{i=1}^{n}$ from N training videos $\{V_i\}_{i=1}^{N}$, where the label y_i is -1 for graphs of normal activity and $+1$ for graphs of abnormal activity. The problem for training the standard SVM [26] from the dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$ can be formulated as:

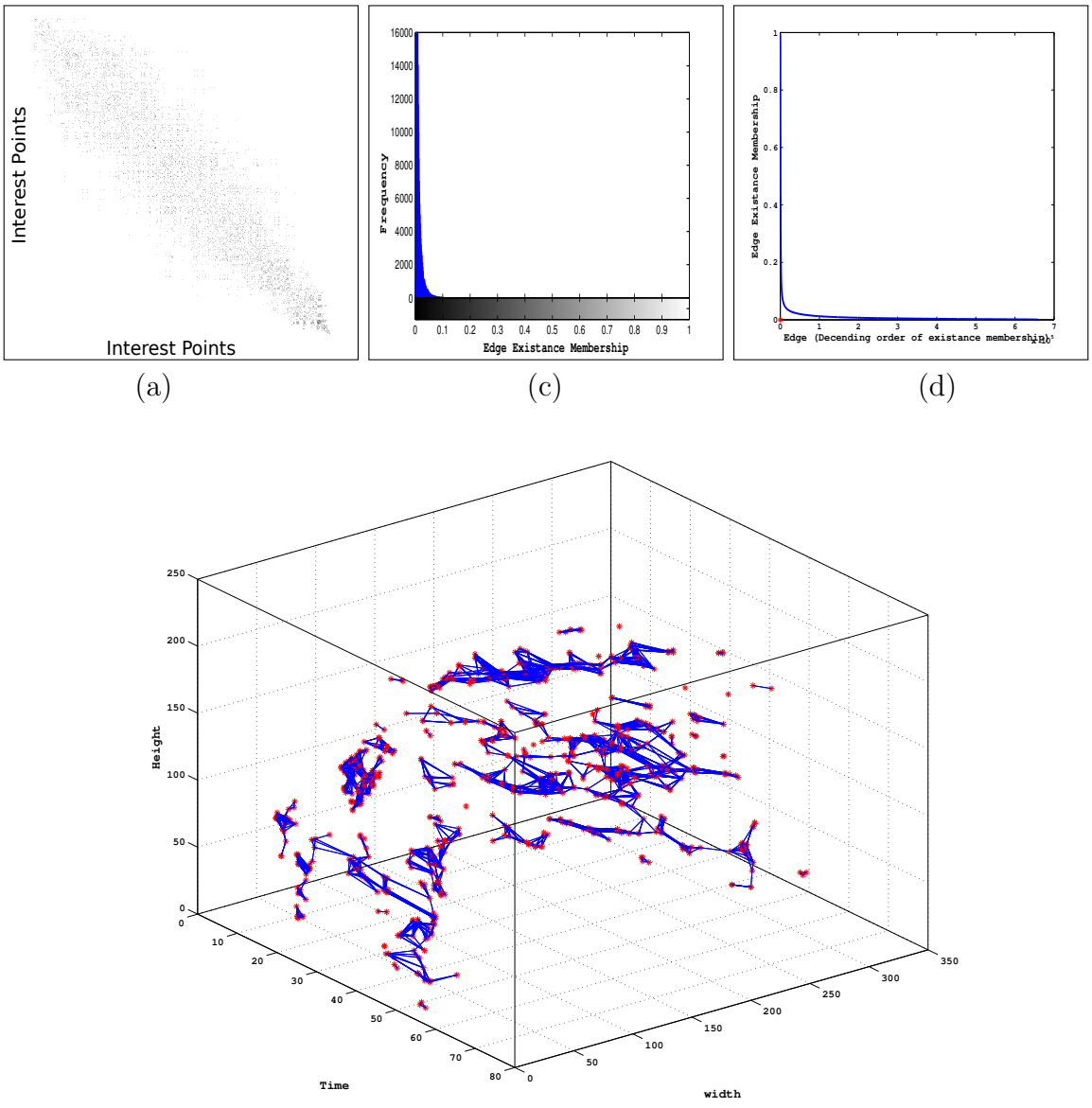
$$\min\ J = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i^T, \mathbf{x}) - \sum_{i=1}^{n}\alpha_i, \tag{6}$$

$$\text{subject to } \sum_{i=1}^{n}\alpha_i y_i = 0 \text{ and } 0 \leq \alpha_i \leq C,$$

where C is the box constraint parameter. By solving this optimization problem we get m support vectors (SV), their respective values of α_i, and the value of bias b. These SVs give a decision function of the form

$$f(x) = sign\left(\sum_{i=1}^{m}\alpha_i y_i K(\mathbf{x}_i^T, \mathbf{x}) + b\right), \tag{7}$$

where α_i are Lagrange multipliers, $\mathbf{x}$ is the test tuple and $f(\mathbf{x}) = f(-1, +1)$ is its prediction. $K(\mathbf{x}_i^T, \mathbf{x})$ is a kernel function used for computation of the similarity between two vectors [27,



(a) (c) (d)

(b)

Figure 2: A sample graph generated for a sample abnormal video from UMN anomaly dataset. (a) Adjacency matrix of the graph generated. (b) 3D-Visualization of the sample graph. (c) Edge existence membership for the sample graph. (d) Edge existence membership frequency for the sample graph.

28, 29, 30]. The similar max-margin classifier can be applied for graph classification using a graph kernel. A graph kernel $K(G_i, G_j)$ gives the similarity between two graphs G_i and G_j i.e. $K(G_i, G_j) \in [0, 1]$. A wide range of graph kernels are proposed in the literature like shortest path kernel and random walk kernel, which are the most widely used graph kernels. However, we adopt random walk kernel because it is computationally efficient than

other graph kernels. The random walk kernel [31] compares two graphs by counting number of common random walks between them. The number of common random walks of length k are calculated by taking direct product graphs because random walk on direct product graph is equivalent to simultaneous random walk in the two graphs [31]. The k^{th} power of adjacency matrix of the resultant graph after direct product gives the number of common walks. The direct product graph of two graphs is defined as given below:

Let $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$ are two graphs, then $G_\times(V_\times, E_\times)$ is the direct product graph where the node and edge set of the direct product graph are defined as

$$V_\times = (v_1^i, v_2^r) : v_1^i \in V, v_2^r \in V^{'}$$
$$E_\times = ((v_i, v_r^{'}), (v_j, v_s^{'})) : (v_i, v_j) \in E \wedge (v_r^{'}, v_s^{'}) \in E^{'}$$

Using the definition of direct product graph, Gartner *et al.* in [31] defined random walk kernel as follows:

Let G_1 and G_2 be two graphs. Then for product graph $G_\times$, let $V_\times$ be the node set of $G_\times$ and $\mathbf{A}_\times$ be the adjacency matrix for the graph product. With start probability $\mathbf{p}_\times$, end probability $\mathbf{q}_\times$, and a sequence of weights (decaying factor) $\lambda = \lambda_1, \lambda_2, ..., (\lambda_i \in \mathbb{R}, \lambda_i \geq 0 \; \forall i \in \mathbb{N})$, the random walk kernel is defined as

$$K(G_1, G_2) = \sum_{k=1}^{\infty} \lambda_k \mathbf{q}_\times^T \widetilde{\mathbf{A}}_\times^k \mathbf{p}_\times, \tag{8}$$

where, $\widetilde{\mathbf{A}} = \mathbf{A}^T[\mathbf{I}.(\mathbf{A}^T.e)]^{-1}$ is the normalized matrix. The kernel in Equation (8) is a valid positive semi definite (p.s.d.) kernel. This can be proved with the help of following technical lemma:

Lemma 1. $\forall k \in \mathbb{N} : \widetilde{\mathbf{A}}_\times^k \mathbf{p}_\times = vec[(\widetilde{A}_2^k p^{'})(\widetilde{A}_1^k p)^T]$.

Lemma 2. *If $X \in \chi^{n \times m}, Y \in \mathbb{R}^{m \times p}$, and $Z \in \chi^{p \times q}$, then*

$$vec[\widetilde{X} Y \widetilde{Z}] = [\widetilde{Z}^T \otimes \widetilde{X}] vec(Y) \in \mathbb{R}^{nq \times 1}$$

where $\otimes$ represents Kronecker product and *vec* represent the vectorization. The proofs of Lemma 1 and Lemma 2 can be found in [32]. Using Lemma 1 and Lemma 2 we can write

$$\begin{aligned}
\mathbf{q}_\times^T \widetilde{\mathbf{A}}_\times^k \mathbf{p}_\times &= \mathbf{q}_\times^T vec[(\widetilde{A}_2^k \mathbf{p}_2)(\widetilde{A}_1^k \mathbf{p}_1)^T] && \text{Using Lemma 1} \\
&= (\mathbf{q}_1 \otimes \mathbf{q}_2)^T vec[(\widetilde{A}_2^k \mathbf{p}_2)(\widetilde{A}_1^k \mathbf{p}_1)^T] && \text{Because } \mathbf{q}_\times = \mathbf{q}_1 \otimes \mathbf{q}_2 \\
&= vec[\mathbf{q}_2^T \widetilde{A}_2^k \mathbf{p}_2 (\widetilde{A}_1^k \mathbf{p}_1)^T \mathbf{q}_1] && \text{Using Lemma 2} \\
&= (\mathbf{q}_1^T \widetilde{A}_1^k \mathbf{p}_1)^T (\mathbf{q}_2^T \widetilde{A}_2^k \mathbf{p}_2) && (9)
\end{aligned}$$

Each individual term of Equation (9) equals $\Phi^k(G_1)^T \Phi^k(G_2)$ for some function Φ, and is therefore a valid p.s.d. kernel. The time complexity of computation of Equation (8) is $O(n^6)$. A fast random walk kernel is proposed by Vishwanathan *et al.* in [33] which reduces the

time complexity to $O(n^3)$ with the help of Sylvester equation and conjugate gradient (CG) methods to solve the system of equations.

$$K(G, G^{'}) = \mathbf{q}_{\times}^{T}(\mathbf{I} - \lambda \mathbf{A}_{\times})^{-1}\mathbf{p}_{\times}. \tag{10}$$

Using graph kernel, the standard support vector machine in Equation (6) can be rewritten for finding a max-margin separating line between normal and abnormal graphs as

$$\min\ J(\boldsymbol{\alpha}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(G_i, G_j) - \sum_{i=1}^{n}\alpha_i, \tag{11}$$

$$\text{subject to } \sum_{i=1}^{n}\alpha_i y_i = 0 \text{ and } 0 \leq \alpha_i \leq C.$$

And the decision function in Equation 7 for a test graph G will be

$$f(x) = sign\left(\sum_{i=1}^{m}\alpha_i y_i K(G_i, G) + b\right). \tag{12}$$

Thus solving Equation (11) for the graphs representing local activities from the training videos gives a model which is then used for making a decision of a local activity graph from test video using Equation 12.

3.3.2. Recognition of Global Abnormal Activity.

The global activities are the set of multiple local activities. The local activities in a global abnormal activity need not be abnormal. The co-occurrence of several normal local activity can lead to an abnormal behaviour. After formulating all the local activities as graphs representing geometric relations of interactions of entities, we build a high level vocabulary $\mathbf{V} = \{G_j\}_{j=1}^{n}$ of graphs using k-median clustering over the set of all graphs $\mathbf{G} = \{G_i\}_{i=1}^{n}$ by solving the objective function given below:

$$\underset{G_j \in \mathbf{V}}{\arg,\min} \sum_{G_i \in \mathbf{G}} K^{-1}(G_i, G_j). \tag{13}$$

Then vocabulary $\mathbf{V}$ of graphs of local activities is used to generate $\mathbf{x} = \{x_i\}_{i=1}^{k}$, a high level bag-of-graphs (BoG) representation for global activities. After this, a standard binary support vector machine given in Equation (6) & (7) is used to classify the global activities into normal or abnormal categories.

4. Experimental Evaluation

This section presents the experimental setup, benchmark datasets, and the outcomes of the experiments. All the simulations are conducted on a machine having 2-Intel Xeon processor with 12 core each, 2-Nvidia GPUs with 5GB device memory each, 128GB physical

memory. The programs are written in C++ and CUDA with the use of *opencv* and *armadillo* libraries. The λ in graph kernel is set to $1/d^2$, d being the largest degree in the graph dataset which is a thumb rule. The value of box constraint C in SVM is set to 1. Three datasets, namely, UCSDped1, UCSDped2, and UMN are used to validate the proposed approach. Fig 3 shows samples of one normal and one abnormal activities from each of the three datasets and their corresponding graph formulation. We compare the proposed approach with other existing state-of-the-art methods like bag-of-words using STIP/SIFT and dense trajectory based approaches. The details of the experimentation on each of the three datasets are discussed in the following subsections:

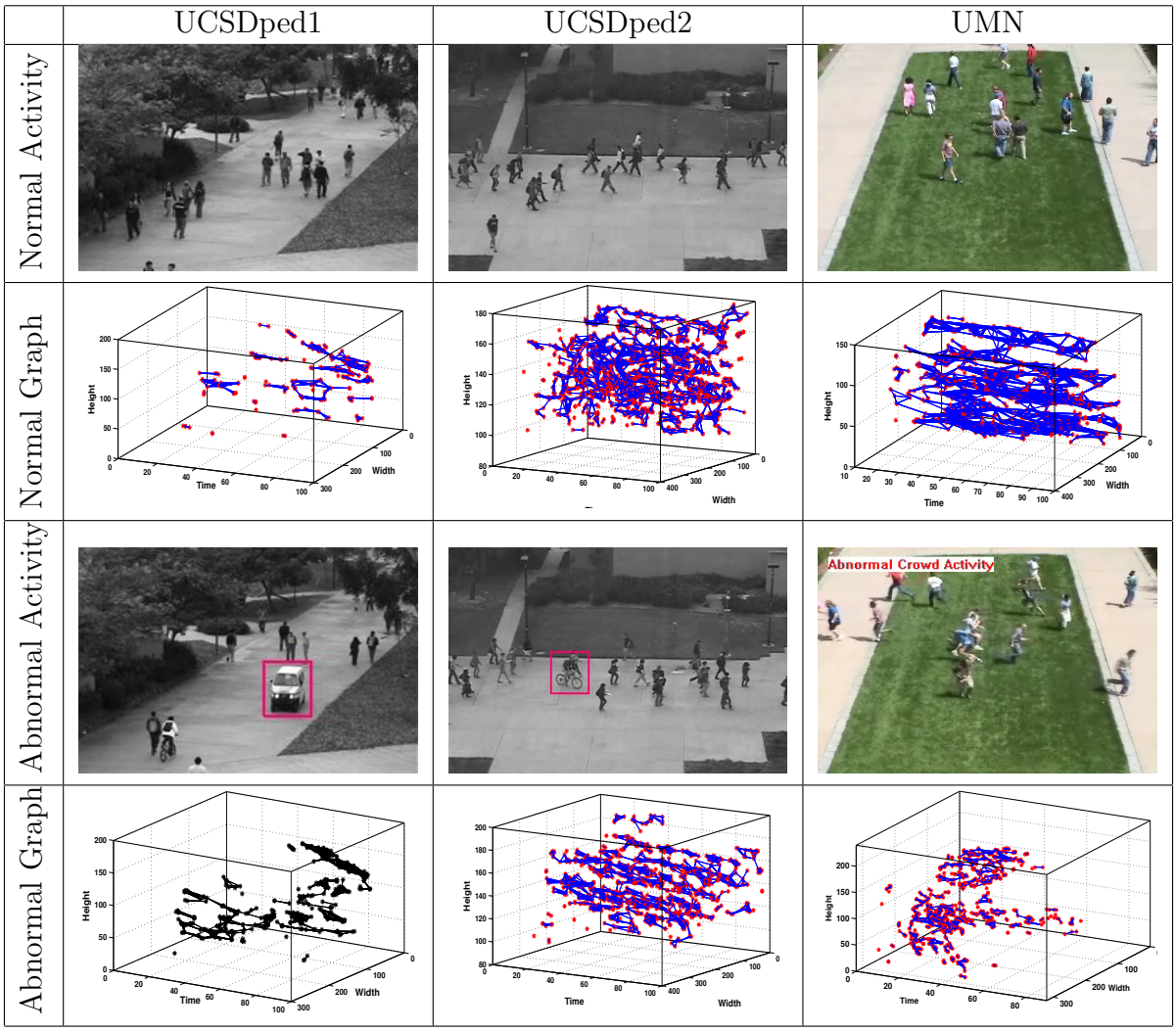


Figure 3: Illustration of normal and abnormal sample and corresponding graphs from all datasets.

4.1. Results on UCSDped1 Dataset

UCSDped1 [20]: UCSD anomaly detection dataset is a widely used standard dataset for video anomaly detection. Videos are captured with a stationary camera mounted at an

elevation, overlooking pedestrian walkways. The crowd density in the walkways ranges from sparse to very crowded. The abnormal events are caused by either the circulation of non-pedestrian entities in the walkways, or anomalous pedestrian motion patterns. This data set contains videos captured in vertical view i.e. groups of people walking towards and away from the camera, and there is some amount of perspective distortion. It contains 34 training and 36 testing videos. The dataset contains crowd of people walking normally vertical to camera. The anomaly includes fast motion, zig-zag motion, and appearance of vehicles. The performance of classification on UCSDped1 dataset using the proposed approach is 97.14%, where as the performance of existing bag-of-words approach using STIP features is 82.00% on the same dataset. The performance of existing other bag-of-words approaches using SIFT and dense trajectories give 80.00% and 85.71%, respectively, on the same dataset. There is a significant improvement in the performance of the proposed approach due to the deviation in the geometrical structure of the graphs generated during normal walking and the graphs corresponds to the fast motion, zig-zag motion, and appearance of vehicles as can be shown in fig 3. Thus the proposed approach is able to locate the evidence in order to detect abnormal activities efficiently.

4.2. Results on UCSDped2 Dataset

Table 1: Comparison of classification performance (%) of proposed approach with existing bag-of-words (BoW) approaches using STIP, SIFT and dense-trajectories (DT)

Dataset	**SIFT+BoW**	**STIP+BoW**	**DT+BoW**	**Proposed BoG**
UCSDped1	80.00	82.00	85.71	**97.14**
UCSDped2	77.62	75.82	88.86	**90.13**
UMN	85.00	85.00	81.00	**95.24**

UCSDped2 [20]: UCSDped2 dataset contains the scenes of pedestrian movement parallel to the camera plane. It contains 16 training video samples and 12 testing video samples. The dataset contains crowd of people walking normally vertical to camera. The anomaly includes fast motion, zig-zag motion, and appearance of vehicles. The proposed approach is able to extract significant evidence with discriminative ability in order to detect abnormal activities efficiently because of incorporation of geometric structure along with motion and appearance information. The geometrical structure of the graphs generated during normal walking are deviating from the graphs corresponding to fast motion, zig-zag motion, and appearance of vehicles see Fig. 3. The performance of classification on UCSDped2 dataset using the proposed approach is 90.13%, where as the performance of existing bag-of-words approach using STIP features is 75.82% on the same dataset. The performance of existing other bag-of-words approaches using SIFT and dense trajectories give 77.62% and 88.86%, respectively, on the same dataset.

4.3. Results on UMN Dataset

UMN [42]: UMN is also a publicly available dataset containing normal and abnormal crowd videos from the University of Minnesota. Each video consists of an initial part of a

Table 2: Performance comparison (%) of proposed approach with existing methods

Reference	Method	UCSDped1	UCSDped2	UMN
Adam *et al.* 2008 [34]	Adam	61.10	54.20	-
Mehran *et al.* 2009 [35]	SF	63.50	65.00	87.40
Kim *et al.* 2009 [21]	MPPCA	64.40	64.20	-
Mahadevan *et al.* 2010 [20]	MDT	75.00	75.00	96.30
Wu *et al.* 2010 [14]	Chaotic Invar.	-	-	94.70
Cong *et al.* 2011 [36]	Sparse	81.00	-	97.20
Raghavendra *et al.* 2011 [37]	PSO	79.00	-	-
Antic *et al.* 2011 [38]	BVP	82.00	-	-
Saligrama *et al.* 2012 [15]	LSA	84.00	-	96.60
Roshtkhari *et al.* 2013 [39]	Roshtkhari	85.00	-	-
Lu *et al.* 2013 [40]	150fps	85.00	-	-
Li *et al.* 2014 [41]	H-MDT	82.20	81.50	96.30
Kaltsa *et al.* 2015 [2]	Swarm	72.98	73.08	97.01
Chang *et al.* 2015 [10]	GPR	76.30	-	-
Proposed	**BoG**	**97.14**	**90.13**	**95.24**

normal behavior and ends with sequences of the abnormal behavior. The dataset contains 11 training and 11 testing video scenes in different environments where a crowd of people walking normally and after some time, they suddenly start running. Fig. 3 shows that the geometrical structure of the graphs generated during normal walking (dense and bigger graph) are deviating from the graphs generated during running (sparse and small graph). In this way, the evidences obtained using the proposed approach contains significant information in order to detect abnormal activities efficiently. The performance of classification on UMN dataset using the proposed approach is 95.24%, where as the performance of existing bag-of-words approach using STIP features is 85.00% on the same dataset. The performance of existing other bag-of-words using SIFT and dense trajectories give 85.00% and 81.00%, respectively, on the same dataset.

It is observed that the proposed approach achieves better performance when compared to other bag-of-words approached using various descriptors like STIP (HoG+HoF), SIFT, and dense trajectories on UCSDped1, UCSDped2, and UMN datasets. Table 1 gives the performance comparison of the proposed approach with existing methods.

Table 2 presents the performance comparison of proposed approach with the existing state-of-the art methods. It can be observed from the Table 2 that the proposed method achieves consistent performance on all the three datasets used. Also, the proposed approach on UCSDped1 and UCSDped2 datasets outperforms the state-of-the-art methods and achieves a comparable performance on UMN datasets. This may be due to the fact that the performance of abnormal activity recognition depends on the nature/type of anomaly present in the dataset. Overall, the proposed method achieves better performance across datasets as it is able to detect a wide variety of abnormal activities in videos.

5. Conclusion

In this paper, we present a novel framework for abnormal activity recognition in surveillance videos. The graph formulation of activities captured in surveillance videos contain significant discriminative ability to determine the behaviour of activities. The motion of the objects/entities, their co-relation, and interactions to each others is subsequently represented by graphs. Finally, the graph formulation of the video activities convert the problem of anomaly detection into a graph classification problem for this, we exploit support vector machine together with graph kernel. The use of graph kernel for measuring similarity between two graphs provides robustness to slight deformations to the topological structures due to presence of noise in data. The experimental results outperforms the existing widely used methods like dense trajectories, bag-of-visual-words etc., which proves the efficacy of the proposed approach.

References

[1] T. Abdullah, A. Anjum, M. F. Tariq, Y. Baltaci, N. Antonopoulos, Traffic Monitoring Using Video Analytics in Clouds, in: Proc. of the IEEE/ACM Int. Conf. on Utility and Cloud Computing, London, 2014, pp. 39–48.

[2] V. Kaltsa, A. Briassouli, I. Kompatsiaris, L. J. Hadjileontiadis, M. G. Strintzis, Swarm Intelligence for Detecting Interesting Events in Crowded Environments, IEEE Trans. on Image Processing 24 (7) (2015) 2153–2166.

[3] M. Bertini, A. Del Bimbo, L. Seidenari, Multi-scale and real-time non-parametric approach for anomaly detection and localization, Computer Vision and Image Understanding 116 (3) (2012) 320–329.

[4] S. Calderara, U. Heinemann, A. Prati, R. Cucchiara, N. Tishby, Detecting anomalies in people's trajectories using spectral graph analysis, Computer Vision and Image Understanding 115 (8) (2011) 1099–1111.

[5] F. Jiang, J. Yuan, S. a. Tsaftaris, A. K. Katsaggelos, Anomalous video event detection using spatiotemporal context, Computer Vision and Image Understanding 115 (3) (2011) 323–333.

[6] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, Crowded Scene Analysis : A Survey, IEEE Trans. on Circuits and Systems for Video Technology 25 (3) (2015) 367–386.

[7] O. P. Popoola, K. Wang, Video-Based Abnormal Human Behavior Recognition: A Review, IEEE Trans. on Systems, Man, and Cybernetics 42 (6) (2012) 865–878.

[8] W. Liu, H. Liu, D. Tao, Y. Wang, K. Lu, Multiview Hessian regularized logistic regression for action recognition, Signal Processing 110 (5) (2015) 101–107.

[9] W. Liu, Z. J. Zha, Y. Wang, K. Lu, D. Tao, P-Laplacian Regularized Sparse Coding for Human Activity Recognition, IEEE Transactions on Industrial Electronics 63 (8) (2016) 5120–5129.

[10] K.-w. Cheng, Y.-t. Chen, W.-h. Fang, Video Anomaly Detection and Localization Using Hierarchical Feature Representation and Gaussian Process Regression, in: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Boston, 2015, pp. 2909–2917.

[11] H. Wang, A. Kl, C. Schmid, L. Cheng-lin, H. Wang, A. Kl, C. Schmid, L. C.-l. Action, A. Kl, Action Recognition by Dense Trajectories, in: Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 2011, pp. 3169–3176.

[12] Y.-K. Wang, C.-T. Fan, J.-F. Chen, Traffic Camera Anomaly Detection, in: Proc. of the Int. Conf. on Pattern Recognition (ICPR), Stockholm, Sweden, 2014, pp. 4642–4647.

[13] M. J. Roshtkhari, M. D. Levine, An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions, Computer Vision and Image Understanding 117 (10) (2013) 1436–1452.

[14] S. Wu, B. E. Moore, M. Shah, Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 2054–2060.
[15] V. Saligrama, Z. Chen, Video Anomaly Detection Based on Local Statistical Aggregates, in: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2112–2119.
[16] H. Wang, C. Schmid, Action Recognition with Improved Trajectories, in: Proc. of the IEEE Int. Conf. on Computer Vision (ICCV), Sydney, Australia, 2013, pp. 3551–3558.
[17] Y. Yuan, S. Member, J. Fang, Q. Wang, Online Anomaly Detection in Crowd Scenes via Structure Analysis, IEEE Trans. on Cybernetics 45 (3) (2015) 562–575.
[18] V. Reddy, C. Sanderson, B. C. Lovell, Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture, in: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), Colorado Springs, CO, USA, 2011, pp. 2160–7508.
[19] C. C. Loy, T. Xiang, S. Gong, Modelling Multi-object Activity by Gaussian Processes, in: Procedings of the British Machine Vision Conf. (BMVC), London, 2009, pp. 13.1–13.11.
[20] V. Mahadevan, W. Li, V. Bhalodia, N. Vasconcelos, Anomaly detection in crowded scenes, in: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 2010, pp. 1975–1981.
[21] J. Kim, K. Grauman, Observe Locally, Infer Globally: a Space-Time MRF for Detecting Abnormal Activities with Incremental Updates, in: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), Miami, FL, 2009, pp. 2921–2928.
[22] M. Sekma, M. Mejdoub, C. B. Amar, Bag of graphs with geometric relationships among trajectories for better human action recognition, in: Proc. of the Int. Conf on Image Analysis and Processing, Genoa, Italy, 2015, pp. 85–96.
[23] I. Laptev, I. Inria, C. Beaulieu, R. Cedex, On Space-Time Interest Points, Int. Journal of Computer Vision (IJCV) 64 (2/3) (2005) 107–123.
[24] N. Dalal, B. Triggs, Histograms of Oriented Gradients for Human Detection, in: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 2005, pp. 886–893.
[25] R. Chaudhry, a. Ravichandran, G. Hager, R. Vidal, Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions, in: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Miami, FL, 2009, pp. 1932–1939.
[26] C. Cortes, V. Vapnik, Support Vector Networks, Machine Learning 20 (3) (1995) 273–297.
[27] M. Gönen, E. Alpaydn, Multiple kernel learning algorithms, Journal of Machine Learning Research 12 (2011) 2211–2268.
[28] C. Xu, D. Tao, C. Xu, A survey on multi-view learning, CoRR abs/1304.5634.
[29] C. Xu, D. Tao, C. Xu, Multi-View Learning with Incomplete Views, IEEE Transactions on Image Processing 24 (12) (2015) 5812–5825.
[30] C. Cortes, M. Mohri, A. Rostamizadeh, Multi-class classification with maximum margin multiple kernel, Proceedings of the International Conference on Machine Learning (ICML) 28 (2013) 46–54.
[31] T. Gärtner, P. A. Flach, S. Wrobel, On graph kernels: Hardness results and efficient alternatives, in: Proc. of the Computational Learning Theory and Kernel Machines, Washington, DC, USA, 2003, pp. 129–143.
[32] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, K. M. Borgwardt, Graph kernels, Journal of Machine Learning Research 11 (1) (2010) 1201–1242.
[33] S. V. N. Vishwanathan, K. M. Borgwardt, N. N. Schraudolph, Fast computation of graph kernels, in: Proc. of the Advances in Neural Information Processing Systems (NIPS), Vancouver, British Columbia, Canada, 2006, pp. 1449–1456.
[34] A. Adam, E. Rivlin, I. Shimshoni, D. Reinitz, Robust real-time unusual event detection using multiple fixed-location monitors, IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 30 (3) (2008) 555–560.
[35] R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model, in: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), Miami, FL,

2009, pp. 935–942.
[36] Y. Cong, J. Yuan, J. Liu, Sparse reconstruction cost for abnormal event detection, in: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 2011, pp. 3449–3456.
[37] R. Raghavendra, A. Del Bue, M. Cristani, V. Murino, Optimizing interaction force for global anomaly detection in crowded scenes, in: Proc. of the IEEE Int. Conf. on Computer Vision Workshops (ICCVW), Barcelona, Spain, 2011, pp. 136–143.
[38] B. Antic, B. Ommer, Video parsing for abnormality detection, in: Proc. of the IEEE Conf. on Computer Vision (ICCV), Barcelona, Spain, 2011, pp. 2415–2422.
[39] M. J. Roshtkhari, M. D. Levine, Online dominant and anomalous behavior detection in videos, in: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 2013, pp. 2611–2618.
[40] C. Lu, J. Shi, J. Jia, Abnormal event detection at 150 FPS in MATLAB, in: Proc. of the IEEE Int. Conf. on Computer Vision (ICCV), Sydney, Australia, 2013, pp. 2720–2727.
[41] W. Li, V. Mahadevan, N. Vasconcelos, Anomaly detection and localization in crowded scenes, IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 36 (1) (2014) 18–32.
[42] UMN, Dataset: unusual crowd activity dataset made available by the university of minnesota.