Anomalous entities detection and localization in pedestrian flows

Habib Ullah[a], Ahmed B. Altamimi[a], Muhammad Uzair[b], Mohib Ullah[c]

[a]*University of Hail, Hail, Saudi Arabia*
[b]*COMSATS Institute of Information Technology, Pakistan*
[c]*Norwegian University of Science and Technology, Gjovik, Norway*

Abstract

We propose a novel Gaussian kernel based integration model (GKIM) for anomalous entities detection and localization in pedestrian flows. The GKIM integrates spatio-temporal features for efficient and robust motion representation to capture the distinctive and meaningful information about the anomalous entities. We next propose a block based detection framework by training a recurrent conditional random field (R-CRF) using the GKIM features. The trained R-CRF model is then used to detect and localize the anomalous entities during the online testing stage. We conduct comprehensive experiments on three benchmark datasets and compare the performance of the proposed method with the state-of-the-art anomalous entities detection methods. Our experiments show that the proposed GKIM outperforms the compared methods in terms of equal error rate (EER) and detection rate (DR) in both frame-level and pixel-level comparisons. The frame-level analysis detects the presence of an anomalous entity in a frame regardless of its location. The pixel-level analysis localizes the anomalous entity in term of its pixels.

Keywords: Local difference binary, Anomaly detection, Optical flow, Pedestrian motion analysis.

1. Introduction

According to Davila et al. [1], the population growth and traffic congestion in urban areas are rapidly increasing which makes the safety of pedestrians a key concern. Therefore, it is important to analyze the pedestrian flows to facilitate smart video surveillance for ensuring pedestrian safety. For this purpose, an important task to ensure pedestrian safety is the detection and localization of anomalous entities in the pedestrian flows which can be used to warn against the possible risks.

The ambiguity of the term *anomalous entity* sets its own challenges in the effort to identify it. There may have different interpretations varying significantly depending on the given context. In this paper, the anomalous entity refers to the moving object exhibiting motion patterns in the pedestrian flow that do not conform to the expected behavior and may warrant special attention or action. These entities present infrequent behavior compared with all other behaviors. Similar definitions are presented by a number of papers that addressed the problem in recent years [5][6][7][8][21][9]. We consider that anomalous entities are rare in the pedestrian flow and they are different from the majority. Examples of anomalous entities include a pedestrian moving in unusual direction against general flow, a passenger avoiding payment at bus/train station, a bicycle passing through a crowd, a vehicle depicting illegal motion at an intersection, two vehicles approaching within a dangerously close vicinity of each other and an abandoned object in pedestrian flows [10][11]. Moreover, sudden changes in velocity, like an abrupt increase of magnitude and the dispersion of individuals in the pedestrian flow indicates that something unusual and potentially dangerous may have occurred. The focus of this paper is anomalous entities detection and localization associated with pedestrian flows only. We consider each moving object as part of the pedestrian flows and non-moving objects or groups of pedestrians as a background.

A number of computer visoin methods for video surveillance [4][5][6][7][8][9] have previously addressed anomalous entities detection. Most of these methods assume that the pedestrian flows are very consistent in motion. In fact, this assumption is not realistic since the pedestrian flows may be scattered and sparse. For example, the frequency and crowdedness of the pedestrians at a certain location may be higher in the official hours and lower during the weekends and later hours. Furthermore, the problem of detecting and locating anomalous entities in the pedestrian flows is very challenging due to the appearance variations of in-

dividual entities, temporal variations, and view angle changes.

To address the above mentioned challenges, we present an efficient method for anomalous entities detection and localization in pedestrian flows that does not rely on the assumption of pedestrian flow consistency. We propose a novel Gaussian kernel based integration model (GKIM). Our GKIM is based on a Gaussian kernel based integration of local difference binary patterns (LDBP) [12] and nested motion descriptor (NMD) [13]. We consider the LDBP since it is characterized by the compact representation of spatial information while the NMD is employed to encode temporal information. The distinguishing feature of NMD is the representation of motion information without requiring an explicit optical flow estimate. Furthermore, we exploit both LDBP and NMD simultaneously to integrate their strengths into a unified model. We propose a Gaussian kernel based approach to integrate the spatio-temporal features by transforming the trace and the determinant of our feature Jacobian matrix into a distinctive space. Therefore, the GKIM represents high quality description of anomalous entities in term of most distinctive information. GKIM models the evolving relative spatial relationships and captures a specific nuance of the underlying motion considering temporal variations. Due to the aforementioned properties, our proposed GKIM is independent of the scattered, sparse, and dense nature of the pedestrian flows.

The complete flow of our proposed method is shown in Fig. 1. In order to detect and localize anomalous entities, we divide each video frame into blocks of equal size where the spatio-temporal features for each block are extracted. To this end, the features are used as a-priori for recurrent conditional random field (R-CRF) [14] training which detect and localize anomalous entities during the testing stage. We propose to use the R-CRF since it can deal efficiently with the label bias problem [14] by integrating the traditional conditional random field (CRF) [15] and recurrent neural networks (RNN) [16]. The main contributions of this paper are:

1. To the best of our knowledge, we are the first to propose the GKIM model for anomalous entities detection and localization. One of the major attraction of the GKIM is its capability to model anomalous entities distinctively in pedestrian flows representing different degrees of scatteredness and sparseness. Moreover, we are the first to explore R-CRF for entities classification in pedestrian flows. The R-CRF has never been used before for pedestrian flow analysis.
2. We extensively evaluate the proposed method on three standard datasets and compared to 10 state-of-the art methods. Our results show that the proposed method significantly outperforms all 10 state-of-the-art methods.
3. We categorize state-of-the-art methods and present a comprehensive survey in this area in the next section.

To assess the proposed GKIM model, we perform extensive experiments on three benchmark datasets and compare the results with 10 state-of-the-art methods: the mixture of dynamic texture (MDT) [4], the mixture of optical flow (MPPCA) [17], the social force (SF) [3], the multiple location monitors (MLM) [18], the clustering and sparse coding (CSC) [7], the holistic features (HF) [8], hierarchical feature representation (HFR) [19], the pedestrian energy map (PEM) [20], the statistical histograms model (SHM) [21], the change detection model (CDM) [9]. Our results show that GKIM achieves superior anomalous entities detection. Moreover, our proposed GKIM outperforms the compared methods in both frame-level and pixel-level analysis in terms of equal error rate (EER) and detection rate (DR). The frame-level analysis detects the presence of an anomalous entity in a frame regardless of its location. The pixel-level analysis localizes the anomalous entity in term of its pixels.

The rest of the paper is organized as follows. In Section 2, an overview of related work is provided. The proposed method for the detection and localization of anomalous entities is presented in Section 3. Experimental results on the benchmark datasets are shown in Section 4 and the conclusion is presented in Section 5.

2. Related work

Anomaly detection and motion segmentation methods are often correlated with each other, therefore, we discuss both by dividing them into three related categories. Methods considering only segmentation are categorized under the term motion segmentation and methods considering only anomaly detection are categorized under the term anomaly detection. Similarly, methods targeting both segmentation and anomaly are categorized under the term motion segmentation and anomaly detection.

In the motion segmentation, Devanne et al. [22] analyze human behavior by decomposing the full motion into short temporal segments representing elementary motions. Lai et al. [23] integrate motion information

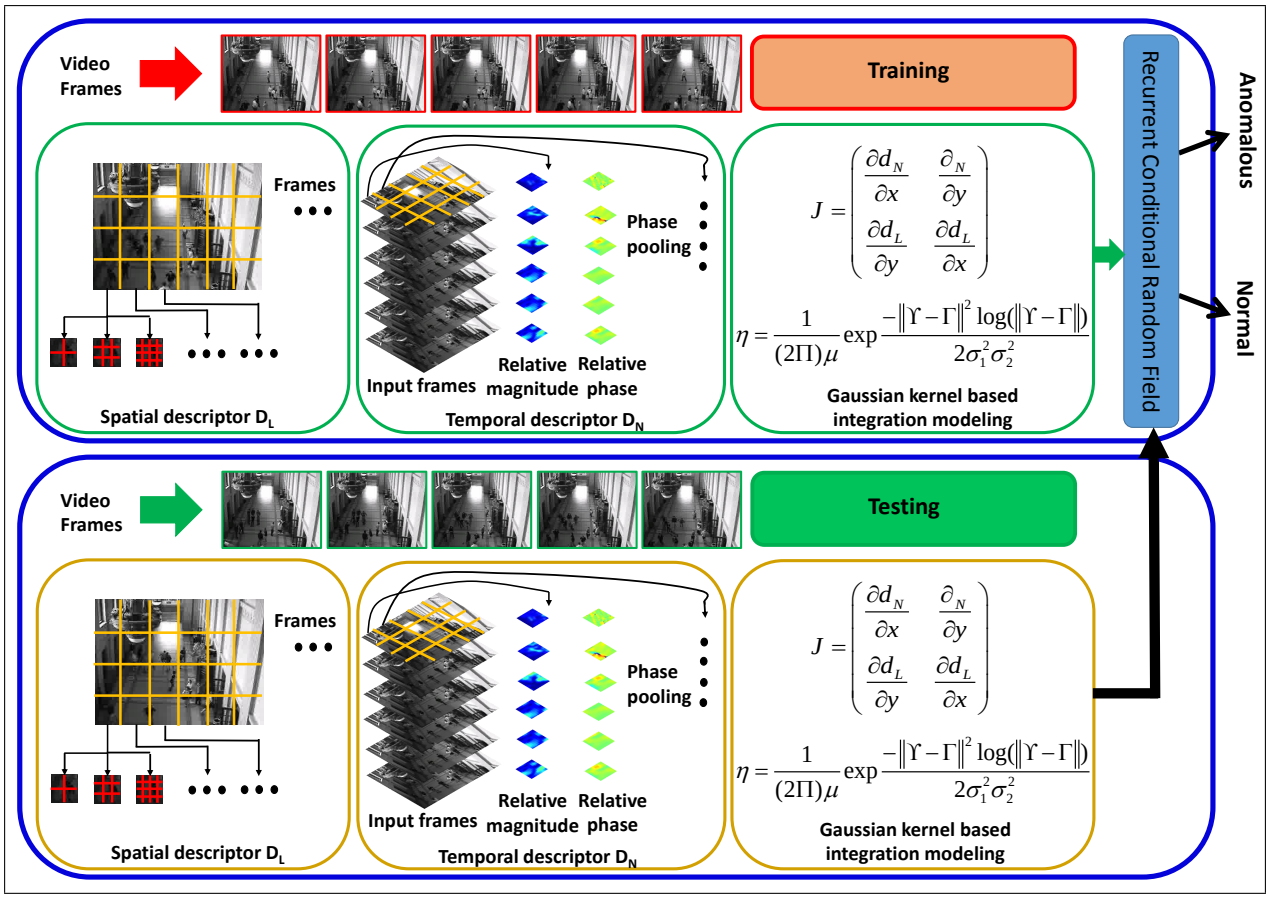

Figure 1: Illustration of the proposed GKIM method. In the training phase, we first compute the Gaussian kernel based integrated spatio-temporal features for distinctive representation of frame patches. An R-CRF model is then learned in supervised manner for separating normal entity patches from anomalous ones. In the testing phase, the learned R-CRF model is used to classify test video patches that are described by the GKIM features.

from a video sequence to construct a sparse affinity matrix. Then a spectral clustering technique is applied on the sparse affinity matrix to segment different motions. Hussain et al. [24] investigate strategies for efficient pixel wise object class segmentation of indoor scenes. They combine both pretrained semantic features and geometric features. Poling et al. [25] use nonlinear embedding of two-view point correspondences into a 9-dimensional space and identify the different motions by segmenting lower-dimensional sub-spaces. Qin et al. [26] combine the region saliency based on entropy rate super-pixel with the affinity propagation clustering algorithm to get seeds in an unsupervised manner, and use random walks method to obtain the segmentation results. Zhong et al. [27] perform moving objects segmentation and matting by integrating a background subtraction and an alpha matting technique via a heuristic seeds selection scheme. Wu et al. [28] propose a convex texture image segmentation model by extracting Gabor features and gray level co-occurrence matrix, which are fused together to effectively construct a discriminative feature space. Kumar and Bhatnagar [29] track multiple objects by detecting object head considering both colour and texture properties of videos. Li et al. [30] deal with challenges in the motion segmentation problem, including perspective effects, missing data, and unknown number of motions. The 3-D motion segmentation is first formulated from two perspective views as a subspace clustering problem. Then, they combine the point correspondence information across multiple image frames via a collaborative clustering stage. Mumtaz et al. [31] propose a motion segmentation approach

that consists of a set of location-specific dynamic texture components, for modeling local background motion, and a set of global dynamic texture components, for modeling consistent foreground motion. For this purpose, an EM algorithm is derived and spatial constraints are applied using Markov random field.

In the anomaly detection category, Marsden et al. [8] propose a set of new features for anomaly detection including crowd collectiveness and mean motion speed. Li et al. [7] propose unsupervised statistical framework for anomaly detection. A clustering and sparse coding technique is then used to learn global activity patterns and local salient behavior patterns. Mahadevan et al. [4] detect anomaly in terms of non-pedestrian entities considering mixtures of dynamic textures (MDT) that uses joint modeling of appearance and dynamics of the scene. Mehran et al. [3] detect abnormal events in terms of escape panics by exploiting the social force model (SFM). Li et al. [7] and Xu et al. [6] detect anomalies in terms of panic situation and circulation of non-pedestrian entities, by considering global and local spatio-temporal patterns. Kompatsiaris et al. [33] introduce histograms of oriented swarms combine with histograms of oriented gradients. Li et al. [34] propose a joint detector of temporal and spatial anomalies based on a video representation that accounts for both appearance and dynamics, using a set of mixture of dynamic textures models. Spatial and temporal anomaly maps are defined at multiple spatial scales that act as potentials of a conditional random field that guarantees global consistency of the anomaly judgments. Wu et al. [35] introduce the concepts of potential destinations and divergent centers to construct the corresponding class-conditional probability density functions of optical flow. The identified divergent centers indicate possible locations at which the unexpected events occur. Krausz et al. [32] detect motion patterns based on optical flow that characterize crowd behavior in stampedes. Kim and Grauman [17] exploit a mixture of probabilistic PCA models to characterize motion patterns in the local volumes. Furthermore, a global inference by incorporating a Markov random field model is applied to detect anomalies locally. Adam et al. [18] propose multiple local monitors to collect low-level statistics for anomaly detection. The measurements of all the monitors are combined together to make a final decision about the existence of an unusual event. Cheng et al. [19] detect anomaly using local feature around interest points in different scales. Yi et al. [20] exploit different energy maps to model the behavior of pedestrians. Zhang et al. [21] propose statistical histograms and support vector data description to detect anomalous entities. Almeida et al. [9] consider 2D motion histograms to identify anomalies in different crowd scenes.

In the motion segmentation and anomaly detection category, Ullah et al. [36] segment crowd motion by fusing the information from a correlation technique and a multi-label optimization technique. The orientation information on top of the segmentation process is collected to detect an abnormal situation as a deviation from what has been observed beforehand. Mehran et al. [37] propose streaklines to cluster coherent regions on the basis of their pixel similarities. Then abnormal behaviors are detected as large deviations from the expected based on the potential functions which are scalar functions calculated from the streaklines. Ali et al. [38] propose a Lagrangian coherent structure (LCS) to segment the flow using the finite time lyapunov exponent (FTLE) [39]. The FTLE is used to extract the boundaries between different flow regions in the scene to perform motion segmentation. Furthermore, the rise of a new LCS in the flow model the detection of the instability in the flow.

Table 1 presents the methods covered in this section in terms of category, features, and models used for representing pedestrian motion segmentation, anomaly detection, as well as the datasets on which these methods are tested.

3. Proposed method for anomalous entities detection and localization

The proposed GKIM method consists of two main steps namely Gaussian kernel based feature integration and R-CRF model based classification. These steps are explained in detail in the following sections.

3.1. Gaussian kernel based integration

We propose a Gaussian kernel based integration of local difference binary patterns (LDBP) [12] and nested motion descriptor (NMD) [13].

For spatial information, we exploit the local difference binary pattern (LDBP), which achieves much higher spatial distinctiveness compared to previous binary descriptors [40][41][42]. For this purpose, both average intensity I_{ave} and first-order gradients, D_x and D_y, of grid blocks are used within a frame patch as depicted in Fig. 2. The average intensity represents the direct component of a grid block. However, the average intensity is too coarse to measure the intensity changes inside a grid block. In contrast, image gradients are more resilient to photometric changes than average intensities and can also encode intensity changes inside a grid such

Table 1: State-of-the-art methods for motion segmentation and anomaly detection. The 'Type' column shows the type of the anomaly that the reference methods detect. The methods with no descriptions in the 'Type' column are targeting only motion segmentation.

Ref.	Category	Features	Model	Type	Dataset
Devanne et al. [22]	Motion segmentation	depth	Dynamic naive Bayes classifier	–	MSRC 12 Cornell activity 120
Lai et al. [23]		Trajectories	Spectral clustering	–	Hopkins 155 62 clip
Hussain et al. [24]		Semantic and geometric features	CNN	–	NYU v2
Poling et al. [25]		Subspace clustering	Global dimension minimization	–	RAS
Qin et al. [26]		Superpixel	Region saliency	–	BSD300 Free 1000
Mumtaz et al. [31]		Dynamic textures	Markov random field	–	FBDynScn
Marsden et al. [8]	Anomaly detection	Mean motion speed	GMM SVM	Escape panic violent	UMN Violent flows
Li et al. [7]		Motion	Clustering and sparse coding	Non-pedestrian entities escape panic	UCSD UMN
Mehran et al. [3]		Motion magnitudes	Social force model	Escape panic	UMN
Krausz et al. [32]		Motion and orientation	Dense optical flow	Stampedes	Loveparade video footage
Mahadevan et al. [4]		Dynamic textures	Mixture models	Non-pedestrian entities	UCSD
Li et al. [7]		Spatio-temporal patterns	Unsupervised statistical	Non-pedestrian entities escape panic	UCSD UMN
Xu et al. [6]		Spatio-temporal patterns	Unsupervised statistical	Non-pedestrian entities	UCSD
Kaltsa et al. [33]		Motion and appearance	HOS	Non-pedestrian entities escape panic	UCSD UMN
Li et al. [34]		Dynamic textures	Joint detector	Non-pedestrian entities escape panic	UCSD UMN
Adam et al. [18]		Low-level cues	Optical flow	Escape panic	UMN PETS2009
Wu et al. [35]		Motion	Bayesian	Escape panic	UMN PETS2009
Almeida et al. [9]		Motion	Optical flow	Escape panic	UMN
Ullah et al. [36]	Both	Motion orientation	Correlation graph cut	Escape panic	UCD
Mehran et al. [37]		Streaklines	Optical flow	Escape panic	UCF UMN
Ali et al. [38]		FTLE	Optical flow	Instability	UCF

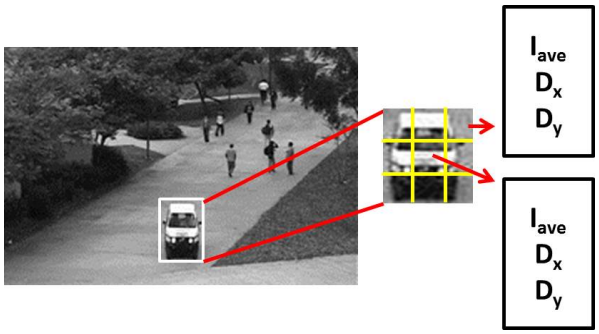

Figure 2: Spatial information extraction. A frame patch is decomposed into 3x3 equal sized blocks. The intensity average I_{ave} and gradients in both directions, D_x and D_y, of each block is computed and compared between every unique pair of blocks.

as the magnitude and direction of an edge. Therefore, we incorporate the first-order gradients also in patch description. We divide each frame patch into grid blocks and calculate I_{ave}, D_x, and D_y as

$$I_{ave}(i) = \frac{1}{P}\sum_{p=1}^{P} I(p) \tag{1}$$

$$D_x(i) = G_x(i) \tag{2}$$

$$D_y(i) = G_y(i) \tag{3}$$

where P is the total number of pixels in a grid block i. G_x and G_y are regional gradients in the x and y directions, respectively. Both regional gradients G_x and G_y represent gradients calculated in each block of a patch. For example, we calculate G_x and G_y, for a grid block of size equal to 2x2 pixels, using differential approximation: $G_x = I(x+1, y, t) - I(x, y, t)$ and $G_y = I(x, y+1, t) - I(x, y, t)$. Similarly, for a 3x3 size, using the formulations: $G_x = \frac{[I(x+1,y,t)-I(x,y,t)]+[I(x+2,y,t)-I(x,y,t)]}{2}$ and $G_y = \frac{[I(x,y+1,t)-I(x,y,t)]+[I(x,y+2,t)-I(x,y,t)]}{2}$.

We capture the spatial patterns of the frame patch through a set of binary tests, each of which compares the I_{ave}, D_x and D_y of a pair of grid blocks (i and j) as in Eq. 4,

$$\zeta(F(i), F(j)) = \begin{cases} 1, & \text{if} \quad \epsilon > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $\epsilon = F(i) - F(j)$ and $\forall(i, j), i \neq j$, ζ is defined as a tuple $\zeta(i) \in \{I_{ave}, D_x(i), D_y(i)\}$. Thus comparing the respective values for each pair of grid blocks results in 3

Figure 3: Multiple gridding. A frame patch is decomposed into three-level gridding that is 2x2, 3x3, and 4x4, where each grid is a block. The Multiple-level gridding captures information at different granularities.

bits using binary test ζ. Performing binary tests on pairwise grid blocks out of sxs grids results in a bit string of $3s^2(s^2-1)/2$. Furthermore, to achieve high robustness and distinctiveness, we use a multiple gridding strategy to capture the structure at different levels of spatial granularities, as presented in Fig. 3. Coarse-level grids can filter out high frequency noise, while fine-level grids can encode detailed local patterns. For this purpose, each frame patch is partitioned in multiple ways. The results from all the partitions are combined to create the spatial descriptor, D_L.

To obtain a regularized representation of the motion, we exploit both the spatial and temporal information. For temporal descriptor, we propose a procedure based on the nested motion descriptor (NMD)[13]. In fact, the challenging problem of anomalous entities detection is concerned with the robust representation of motion that captures the informative and meaningful properties of the anomalous entities, and discards irrelevant information associated with background and pedestrians. The distinguishing feature of our temporal descriptor is in the representation of motion associated with anomalous entities, without requiring an explicit optical flow estimate.

More specifically, we encode motion information considering both pooling the magnitude of edges, and phase gradients to capture translation of edges in a video. We exploit the complex steerable pyramid [43] to divide each block of a frame into a set of orientation and scale selective subbands. The complex steerable pyramid consists of basis filters in quadrature pairs, that estimate magnitude and phase for each subband. We calculate the relative magnitude and relative phase for each subband from consecutive frames. This captures a fixed velocity tuning for a velocity parameter v that adjusts the procedure to faster or slower motion. Additionally, phase pooling is performed by inferring the relationship between phase gradients and component velocity, such that pooling component velocity is equivalent to pooling phase gradients. A set of pooling regions

are defined to pool the component velocity in neighboring spatial and temporal regions, to provide invariance to view angle changes. Each of the pooling regions is centered at a pixel position, and the pooling regions are uniformly distributed in angle around the pixel position. Each pooling region is represented by a component velocity, and all orientations and scales are merged into a single temporal descriptor.

Relative phase or phase gradients are equivalent to the salient motion of a foreground object in a block. The motion of the anomalous entities causes pixel motion that could be the combination of translation, rotation and scale. Therefore, the motion field in a block is uniformly offset by the motion of the anomalous entities. The relative phase is also offset by the same motion. We can encode this by computing a phase difference with neighbors in position and scale. Therefore, we divide each block of a frame into orientation and scale selective subbands. The orientation subbands present an attribute that the response to an arbitrary orientation is a linear composition of basis subbands. Furthermore, a complex steerable pyramid comprises of basis filters in quadrature pairs.

From the above decomposition, it is simple to compute a phase and magnitude response at many scale and orientation subbands. The temporal descriptor D_N is formulated as

$$D_N(i,j,k) = \begin{cases} 1, & \text{if} \quad \grave{b}(i,j,k) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$\grave{b}(i,j,k) = b(i,j,k,t-2^k v) - b(i,j-1,k-1,t-2^k v) \tag{6}$$

$$b(i,j,k,t) = \frac{\sum_{q \in W_n(j,k)} \lambda^t_{ik}(q)}{\sum_{q \in W_n(j,k)} H(q)} \tag{7}$$

$D_N(i,j,k)$ represents a binarized temporal descriptor where $b(i,j,k,t)$ is the pooled component velocity for the orientation subband i, lobe j, and lobe scale k at frame t. Eq. 6 formulates the difference between component velocities at neighboring scales and positions within the same frame. Thus it represents the connection among the frame offset, pooling scale, and bandpass scale. In Eq. 7, $W_n(j,k)$, $\lambda^t_{ik}(q)$, and $H(q)$ represent the pooling regions over which the accumulation occurs, the component velocity, and the pooled phase stability constraint, respectively. q is the interest point that satisfies the phase stability constraint. The component velocity is formulated in Eq. 8

$$\lambda = \frac{-\Theta_t}{\mid \vec{\Theta} \mid} \tag{8}$$

where $\vec{\Theta} = [\Theta_x, \Theta_y]$ is the spatial phase gradient and $-\Theta_t$ is the derivative of the phase with respect to time, which comes from the modeling of the phase constancy constraint formulated in Eq. 9.

$$\Delta\Theta \bullet \vec{v} = 0 \tag{9}$$

The phase constancy constraint encodes the phase gradient $\Delta\Theta(t) = [\frac{\partial\Theta}{dx}, \frac{\partial\Theta}{dy}, \frac{\partial\Theta}{dt}]^T$ and the component velocity $\vec{v} = [\frac{\partial x_0}{dt}, \frac{\partial y_o}{dt}, 1]^T$ at a pixel position (x_0, y_0) in a block. Rearranging the terms formulates $\frac{\partial\Theta}{dx}v_x + \frac{\partial\Theta}{dy}v_y = -\frac{\partial\Theta}{dt} = -\Theta_t$. The temporal descriptor, D_N, and the spatial descriptor, D_L, are integrated to build a Jacobian matrix as formulated in Eq. 10.

$$J = \begin{pmatrix} \frac{\partial d_N}{\partial x} & \frac{\partial d_N}{\partial y} \\ \frac{\partial d_L}{\partial y} & \frac{\partial d_L}{\partial x} \end{pmatrix} \tag{10}$$

We then explore a Gaussian kernel that integrates the spatio-temporal features as formulated in Eq. 11.

$$\eta = \frac{1}{(2\pi)\mu} exp \frac{- \parallel \Upsilon - \Gamma \parallel^2 log(\parallel \Upsilon - \Gamma \parallel)}{2\sigma_1^2\sigma_2^2} \tag{11}$$

where Υ and Γ represent the determinant $\Upsilon = det(J)$ and the trace $\Gamma = Tr(J)$ of Jacobian matrix J, respectively. μ, σ_1^2, and σ_2^2 represent the mean, variance of the spatial descriptor, and the variance of the temporal descriptor, respectively. Gaussian kernel constructs integrated spatio-temporal features by transforming Jacobian matrix into a distinctive space. Thus a unified and organized description is produced from the Gaussian modeling by considering the determinant and trace of Jacobian matrix. Thus, the proposed GKIM represents high quality description of anomalous entities in term of most distinctive information. Our proposed GKIM method avoids modeling the features both at a level that is too fine or too coarse. At a level too fine, one is swamped with extraneous detail. At a level too coarse, important characteristics may be missed. Our model mixes the advantages of both spatial and temporal features to compute distinctive and unique feature reprsentations.

As the occurrences of the anomalous entities are usually sparse in the pedestrian scenes and similarities with the background are high, previous methods [4][5][6][7][8] show limited discriminative ability. A method succeeding in one pedestrian scene might fail

in another scene with different sparsity and density. In contrast, our proposed GKIM is highly discriminative and efficient in detecting anomalous entities.

3.2. Learning R-CRF model for classification

For classifying the pedestrian entities into normal and abnormal, we explore R-CRF [14]. The motivation for considering the R-CRF comes from the observation that R-CRF integrates the strengths of the Recurrent neural networks (RNN) and the conditional random field (CRF). The R-CRF exploits the CRF-like sequence-level objective function and the RNN activation as features. Thus, it takes the advantage of the sequence-level discrimination ability of the CRF and the feature learning ability of the RNN.

Recurrent neural networks (RNNs) have recently shown good performance in various applications from gesture recognition [44] to speech recognition [45]. However, the performance can be significantly enhanced by integrating elements of the conditional random field (CRF) model; specifically, the explicit modeling of output-label dependencies with transition features, and its global sequence-level objective function. An RNN maintains a representation for each feature such that similar features tend to be close with each other, and relationships between features are preserved. It is worth noting that RNN produces a sequence of locally normalized output distributions that can suffer from the label bias [15] problem.

To cope with the problem of label bias, R-CRF [14] is developed. The R-CRF is based on the RNN-LU model of Yao et al. [16]. The model consists of a layer of inputs connected to a set of hidden nodes; a fully connected set of recurrent connections amongst the hidden nodes; and a set of output nodes. Each layer represents a set of neurons, and the layers are connected with weights. The output layer produces a probability distribution over labels. The hidden layer maintains a representation of the relationship among features. The input vector has a dimension equal to the descriptor size η.

3.3. Training

We train an R-CRF in supervised manner to learn a model for classifying anomalous and normal entities. Our training data consists of labeled samples where each sample represents a frame patch of a normal or an anomalous entity. The anomalous entities include patches from cyclists, skaters, vehicles, a pedestrian walking on a lawn, a running pedestrian, a pedestrian walking in opposite direction of dominant pedestrian flow. The background patches are labeled as a normal entity. All the labeled samples can be regarded as independent samples to train the model. The details of the R-CRF model learning are given below.

The joint probability of the output label $y(t)$ given the input observation vector $x(t) = \eta$ of a traditional conditional random field [15] is formulated as

$$\begin{aligned} p\left(y(1:T)|x(1:T)\right) &\propto exp(A+B) \\ A &= \sum_{t-1}^{T}\sum_{m}\sigma_m f_m(y(t-1), y(t)) \\ B &= \sum_{t-1}^{T}\sum_{k}\omega_k r_k(y(t-1), x(t)) \end{aligned} \tag{12}$$

where $f_m(y(t-1), y(t))$ and $r_k(y(t-1), x(t))$ are the feature functions. Each feature function renders the score for any output label in terms of its relevance to the input observation vector x, representing our descriptor. σ_m and ω_k are the weight parameters associated with the feature functions f_m and r_k, respectively. These parameters encode the relative importance of feature functions.

In the R-CRF model, an RNN is used to generate the input features for a CRF . For the sake of simplicity, the input-output pair with the side feature inputs is denoted by $((x(1:t), f(1:t)), y(1:t))$. In the R-CRF, the weight ω_k is absorbed into the feature itself, transforming the objective function as

$$\frac{exp\sum_{t=1}^{T}(\psi a_{y*(t-1)y*(t)} + z_{y*(t)}t)}{\sum_{\forall y(1:T)} exp(\sum_{t=1}^{T}\psi a_{y(t-1)y(t)} + z_{y(t)}t)} \tag{13}$$

where $y*(1:T) = [y*(1)...y*(T)]$ represents the correct output labels and ψ is a real value set to 1.0. $z_y(t)$ is the element in the output layer activity before softmax. The objective function can be represented in a log-scale as

$$\begin{aligned} U &= A - \log expB \\ A &= \sum_{t=1}^{T}(\psi a_{y*(t-1)y*(t)} + z_{y*(t)}t) \\ B &= \sum_{t=1}^{T}(\psi a_{y(t-1)y(t)} + z_{y(t)}t) \end{aligned} \tag{14}$$

During the training, the R-CRF iterates between a forward pass and a backward pass to maximize the objective function formulated in Eq. 14.

Forward pass: The forward pass is computed using Eq. 15.

$$\begin{aligned}\epsilon(t,i) &= \sum_{\forall y(1:t)\bigcap y(t)=i} exp \sum_{k=1}^{t}(\psi a_{y(k-1)y(k)} + z_{y(k)}k) \\ &= \sum_{j} \epsilon(t-1,j) exp(\psi a_{ji} + z_i(t)) \\ &= exp(z_i(t)) \sum_{j} \epsilon(t-1,j) exp(\psi a_{ji})\end{aligned} \tag{15}$$

where $\epsilon(t,i)$ is the sum of partial path scores ending at position t with label i. A minor change in the forward pass results in the Viterbi algorithm represented by $\hat{\epsilon}(t,i) = exp(z_i(t)) max_{\forall j}(\hat{\epsilon}(t-1,j) exp(\psi a_{ji}))$.

Backward pass: The backward pass is computed using Eq. 16.

$$\begin{aligned}\Lambda(t-1,q) &= \sum_{\forall y(t:T)\bigcap y(t)=q} exp \sum_{k=t}(\psi a_{y(k-1)y(k)} + z_{y(k)}k) \\ &= \sum_{j} \Lambda(t,j) exp(\psi a_{qj} + z_j(t))\end{aligned} \tag{16}$$

The score of the backward pass is the sum of partial path scores starting at the position $t-1$ with the label q. The gradients with respect to the feature $z_{y(t)=k}(t)$ can be computed with the forward and backward scores as formulated in Eq. 17.

$$\begin{aligned}\frac{\partial U}{\partial z_{y(t)=k} t} &= \delta(y(t) = y*(t)) \\ - \sum_{\forall y(1:T)} \frac{exp(\sum_t \psi a_{y(t-1)y(t)} + z_{y(t)}(t)) \delta(y(t)=k)}{\sum_{\forall y(1:T)} exp(\sum_t \psi a_{y(t-1)y(t)} + z_{y(t)}(t))} &= \delta(y(t)=k) - \frac{\epsilon(t,k)\Lambda(t,k)}{\sum_j \epsilon(t,j)\Lambda(t,j)}\end{aligned} \tag{17}$$

The error signal is obtained with Eq. 17. The back-propagation procedures is subsequently reused by the model for updating the parameters. The gradients are computed to update the label transition weights according to Eq. 18.

$$\begin{aligned}\frac{\partial U}{\partial a_{ji}} &= \psi \sum_t (\delta(y(t-1)=j, y(t)=i) - \frac{A}{B}) \\ A &= \epsilon(t-1,j)\Lambda(t,i) exp(\psi a_{ji} + z_i(t)) \\ B &= \sum_t \epsilon(t,j)\Lambda(t,j)\end{aligned} \tag{18}$$

Using stochastic gradient ascent over the training data, the model parameters are updated.

3.4. Testing

In the testing phase, the learned R-CRF model is used to classify patches of the input video frames into anomalous and normal entities. More specifically, we divide each test input video frame into patches to extract spatio-temporal features which are transformed by our proposed Gaussian kernel based feature modeling. This is fed to the learned R-CRF model to determine if the patch is anomalous or normal. It is worth noticing that a background patch is considered as a normal entity. In fact, all the labeled samples can be regarded as independent samples to train the weight parameters of the R-CRF, which can be optimized by maximizing the likelihood of the training samples. Such assumption is widely used for various learning methods. Gradient descent is used for optimizing the parameters. Once the parameters of the R-CRF are determined, they are used in the objective function that classifies the patches during the testing stage. In the testing stage, the identification of an anomalous patch, irrespective of its location, in a frame represents the detection of anomalous entity. The identification of the location of an anomalous entity, in term of pixels, in a frame represents the localization of the anomalous entity.

4. Experiments

We evaluate the performance of our proposed method for anomalous entities detection on three benchmark datasets available publicly. These include UCSD [4], UMN [46], and UCD [36].

4.1. Details of datasets

UCSD dataset: The UCSD dataset consists of two subsets: ped1 and ped2. Both subsets represent surveillance videos captured by a fixed camera overlooking pedestrian walkways. In Ped1, people are moving towards and away from the camera, with some perspective distortion and ped2 contains video of people moving parallel to the camera. The resolutions of Ped1 and Ped2 are 158x238 and 240x360, respectively. The normal event appearing in the dataset is sequences of pedestrians on the walkways, with a varying density from sparse to very dense. The non-pedestrian entities include cyclists, skaters, vehicles, people walking on a lawn. The appearance of all non-pedestrian entities occurs naturally, i.e., they were not staged or synthesized for data set collection. The video footage of each scene is divided into clips of 120-200 frames. Ped1 consists of 34 training video clips and 36 testing video clips; whereas ped2

contains 16 training video clips and 12 testing video clips.

UMN dataset: The UMN dataset consists of normal and abnormal crowd videos from the university of Minnesota. It consists of three different indoor and outdoor scenes representing 11 different scenarios of escape events. There are total 7739 frames of 320x240 pixels. Each video begins with the normal behaviors of people walking and standing.

UCD dataset: The UCD dataset contains two outdoor videos of students moving across two buildings lasting for 12 and 5 minutes, respectively. Each sequence is segmented into two different subsequences with people mainly moving in a horizontal direction in the scene. This dataset defines anomaly as the deviations from what has been observed beforehand. This anomaly represents any pedestrian moving in the opposite direction of the general flow of the pedestrians.

4.2. Experimental setup

For model training, we automated the patch extraction from each dataset since the manual extraction of patches is a resource and time consuming process. We define a set of non-overlapping patches of fixed size equal to 20x20 pixels to cover all video frames of the datasets. For all the datasets, the patch size is small enough to capture anomaly location, but at the same time large enough to extract related details of appearance. For each dataset, a patch is randomly selected and then carefully labeled as an anomalous entity or normal patch.

The training clips of both Ped1 and Ped2 do not contain anomalies, therefore, we randomly select half testing clips of Ped1 and Ped2 for training the model and the rest of the clips are used as testing samples. We extract 38,540 normal frame patches and 37,728 anomalous frame patches from Ped1, and 32,854 normal patches and 31,248 anomalous patches from Ped2.

In the UMN dataset, for training, we use normal frames of one scenario from scene 1 and two scenarios from scenes 2 and 3 to model normal pedestrian behavior. We use the rest of the frames for testing. We extract 47,591 normal frame patches and 45,533 anomalous frame patches from the training frames.

For the UCD dataset, we consider the frames from subsequence 1 and subsequence 3 for the training. The subsequence 2 and subsequence 4 are used for testing. We extract 56,847 normal patches and 58,317 anomalous patches from the training frames.

We compare the results with 10 closely related state-of-the-art methods: the mixture of dynamic texture (MDT) [4], the mixture of optical flow (MPPCA) [17], the social force (SF) [3], the multiple location monitors (MLM) [18], the clustering and sparse coding (CSC) [7], the holistic features (HF) [8], hierarchical feature representation (HFR) [19], the pedestrian energy map (PEM) [20], the statistical histograms model (SHM) [21], the change detection model (CDM) [9].

For quantitative evaluation of anomalous entities detection, the equal error rate (EER) for frame-level and the detection rate (DR) for pixel-level analysis are calculated to measure the overall performance. Additionally, we compute the Receiver Operating Characteristic (ROC) curves of True-Positive Rates (TPR) versus False-Positive Rates (FPR). It is worth noting that frame-level criterion is mostly used in the literature. However, it only measures temporal localization accuracy. Therefore, it enables errors due to lucky detection of anomalous entities. For example, it allots a perfect score to a method that identifies an anomalous entity at a random location of a frame. The frame level criterion labels a frame as abnormal if it contains at least one anomalous patch, regardless of where it is localized.

In contrast, the pixel-level criterion is much reliable evaluation metric. Therefore, we consider both the temporal and spatial accuracies to rule out lucky detection. The pixel level criterion detects anomalous entity if at least 40% of the truly anomalous pixels are detected. The pixel level criterion is also used to localize the anomalous entities.

Both frame level and pixel level criteria are based on TPR and FPR. The presence and absence of anomalous entities are represented by a positive and a negative, respectively. This is compared to the frame-level ground-truth, to determine the number of true- and false-positive frames. Similarly, pixels related to the anomalous entity are compared to the pixel-level ground-truth to determine the number of true-positive and false-positive. For this purpose, we used the ground-truth of the UCSD, UMN, and UCD datasets provided by Antić et al. [47] and Ullah et al. [36][48].

4.3. Results

The qualitative performance of our proposed method on UCSD, UMN, and UCD datasets is presented in Fig. 4, Fig. 5, and Fig. 6, respectively. In each figure, first row presents the sample frames taken from the original video sequences and the second row presents the results of our proposed GKIM method.

In Fig. 4, the detection of anomalous events in terms of non-pedestrian entities in UCSD dataset is annotated in white for the purpose of visualization. Fig. 4 shows

Figure 4: Results of proposed GKIM on UCSD dataset: The detection and localization of anomalous entities are overlaid on the original frames and annotated in white for the purpose of visualization. GKIM has successfully detected and localized cyclists, skaters and vehicles as anomalous entities.

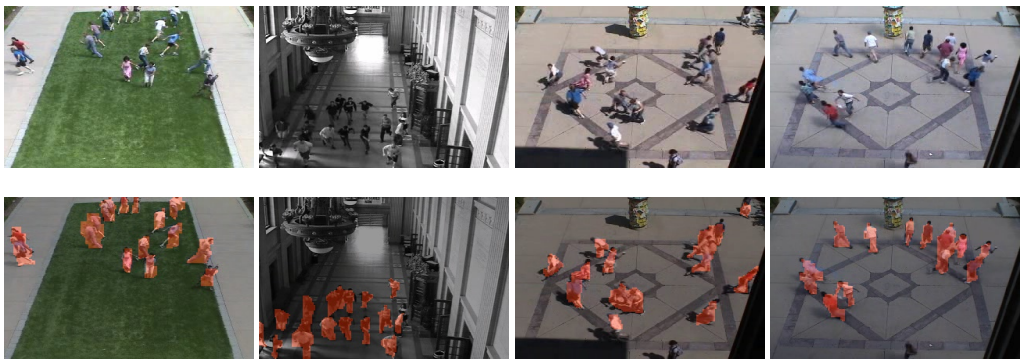

Figure 5: Results of proposed GKIM on UMN dataset. The detection and localization of anomalous entities are overlaid on the original frames and annotated in red for the purpose of visualization. GKIM has successfully detected the escape panics accurately in all the four scenes.

Figure 6: Results of proposed GKIM on UCD dataset. The detection and localization of anomalous entities are overlaid on the original frames and annotated in blue for the purpose of visualization. The pedestrian flows representing deviations from what have been observed before are detected accurately by GKIM in all the four scenes.

Table 2: Quantitative analysis. Equal error rate (EER) for frame-level criterion for the reference methods and our proposed method for both subsets, Ped1 and Ped2, are presented.

Sub.	MDT	MPPCA	SF	MLM	CSC	HF	HFR	PEM	SHM	CDM	Prop.
Ped1	25	40	31	38	20	32	25	29	19	22	16.5
Ped2	25	30	42	42	21	36	29	31	21	24	17
Avg.	25	35	36.5	40	20.5	34	27	30	20	23	16.75

Table 3: Quantitative analysis. Detection rate (DR) for pixel-level criterion for the reference methods and our proposed method for both subsets, Ped1 and Ped2.

Sub.	MDT	MPPCA	SF	MLM	CSC	HF	HFR	PEM	SHM	CDM	Prop.
Ped1	55	23.2	40.9	32.6	57	43	55	31	58	57	63.7
Ped2	60	22.4	27.6	22.4	55	40	51	27	56	55	66.8
Avg.	57.5	22.8	34.25	27.5	56	41.5	53	29	57	56	65.25

Table 4: UMN dataset. Equal error rate (EER) and detection rate (DR) for the reference methods and our proposed GKIM method are presented in the first row and the second row, respectively.

Dataset	MDT	MPPCA	SF	MLM	CSC	HF	HFR	PEM	SHM	CDM	Prop.
UMN	09	16	13	18	07	06	11	15	03	05	04
	69	55	65	50	71	75	65	49	94	86	89

Table 5: UCD dataset. Equal error rate (EER) and detection rate (DR) for the reference methods and our proposed GKIM method are presented in the first row and the second row, respectively.

Dataset	MDT	MPPCA	SF	MLM	CSC	HF	HFR	PEM	SHM	CDM	Prop.
UCD	15	25	18	21	12	16	17	22	14	12	09
	53	37	48	40	69	51	49	39	65	70	75

that our method detects and localized the cyclists accurately in the first and second columns. Two skaters and a vehicle are also detected in the third and last columns.

In Fig. 5, the detection of anomalous events in terms of escape panics in UMN dataset is annotated in red. Here, escape panics are detected properly in all the video sequences. A person in the bottom of the scene in the third column is not detected since he is walking.

Similarly, in Fig. 6, the detection and localization of anomalous events in terms of deviations from what has been observed before is annotated in blue. In Fig. 6, pedestrian flows are detected representing deviations from what has been observed before. Four persons in the bottom and three persons to the left of the third column are not detected since they are not deviating from the regular pedestrian flows seen before.

For quantitative performance analysis, we calculated the average Equal Error Rate (EER) and average Detection Rate (DR) for the three datasets.

The average EER and average DR for the Ped1 and Ped2 of UCSD dataset are reported in Table 2 and Table 3, respectively. The average EER and average DR for UMN and UCD datasets are reported in Table 4 and Table 5, respectively. In Table 4 and Table 5, the first and second rows represent the average EER and average DR, respectively.

As can be seen in the tables, our proposed method outperforms all the optical flow or tracklets based meth-

ods: the MPPCA [17], the SF [3], the MLM [18], the HF [8], the PEM [20], and the CDM [9], and the spatio-temporal volumes based methods: the MDT [4], the CSC [7], and the HFR [19]. The (SHM) [21] method performs better in case of UMN dataset in Table 4. However, our GKIM method performs better in case of other datasets: UCSD and UCD in Table 2, Table 3, and Table 5. These results show that there is a significant advantage of our proposed GKIM model that transforms the sptio-temporal features to disciminative representation bringing forth strong capabilities. The reference methods based on optical flow cannot cope with the adaptively changing sparse and dense nature of the pedestrian flows where dynamic motion and occlusions exist. Also the optical flow computes instantaneous displacement without taking into account the appearance information. Furthermore, the MDT [4] and CSC [7] fail to capture discriminative motion patterns because informative movements only occur in specific regions of the videos, that depend on the type of anomalous entity. Our proposed GKIM represents high quality description of anomalous entities with the spatial and temporal components. Therefore, we outperform the reference methods in both frame-level and pixel-level analysis. Presenting results based on both criteria reveal the robustness of our proposed method.

We also report the ROC curves for the frame-level analysis in addition to the EER and DR. The ROC curves for ped1 and ped2 of UCSD dataset are presented in the left and right columns of Fig. 7, respectively. The ROC curves for UMN and UCD datasets are reported in the left and right columns of Fig. 8, respectively. The significant improvement in the performance of our proposed GKIM method can be seen in all the figures. In fact, the previous methods generally simplify the original frames by partitioning them into volumes or tracklets. This is done for efficiency but, most importantly, for computing discriminative features with the intention that these features will thus be more robust. However, these partitioning inevitably merges the pixels of anomalous entities into background. Therefore, such features are not sufficiently discriminative for detecting anomalous entities in pedestrian flows. Moreover, the MDT and the CSC evaluate correlation of features using only video volumes. This might bring the advantage of simplicity, when it is used as a region descriptor. Nevertheless, as a generic representation, the capability of modeling feature relationship using video volumes cannot be conveniently altered to model different feature relationships in terms of anomalous entities. Our proposed GKIM method addresses these issues by modeling a Jacobian matrix as a generic feature representation. Each of its entries are evaluated by a Gaussian kernel which is discriminative, even if features are scarce. More importantly, this kernel transformation gives us unlimited opportunities to model feature relationship in an efficient manner.

Table 6: Only spatial and only temporal features, individually. Equal error rate (EER) and detection rate (DR) are provided for considering only spatial, temporal, and both information, respectively. For each dataset, the first row shows EER and the second row shows DR. For UCSD dataset, average EER and average DR for ped1 and ped2 are presented.

Dataset	Spatial	Temporal	Spatio-temporal (GKIM)
UCSD	45	42	16.75
	21	25	65.25
UMN	20	19	04
	60	68	89
UCD	24	21	09
	35	39	75

Table 7: Only RNN and only CRF classification models. Equal error rate (EER) and detection rate (DR) are provided for considering only RNN, CRF, and both R-CRF, respectively. For each dataset, the first row shows EER and the second row shows DR. For UCSD dataset, average EER and average DR for ped1 and ped2 are presented.

Dataset	RNN	CRF	R-CRF (GKIM)
UCSD	19.5	18.3	16.75
	59.8	61	65.25
UMN	06	07	04
	83	80	89
UCD	13	11	09
	68	73	75

4.4. Performance analysis of individual and GKIM features

In Table 6, we provide results considering only spatial, only temporal, and our Gaussian kernel based spatio-temporal features for all the three datasets. It is worth noticing that we are considering Gaussian kernel based modeling for individual information. For this purpose, the missing information in Jacobian matrix are replaced with 1 instead of 0 to avoid invalid calcula-

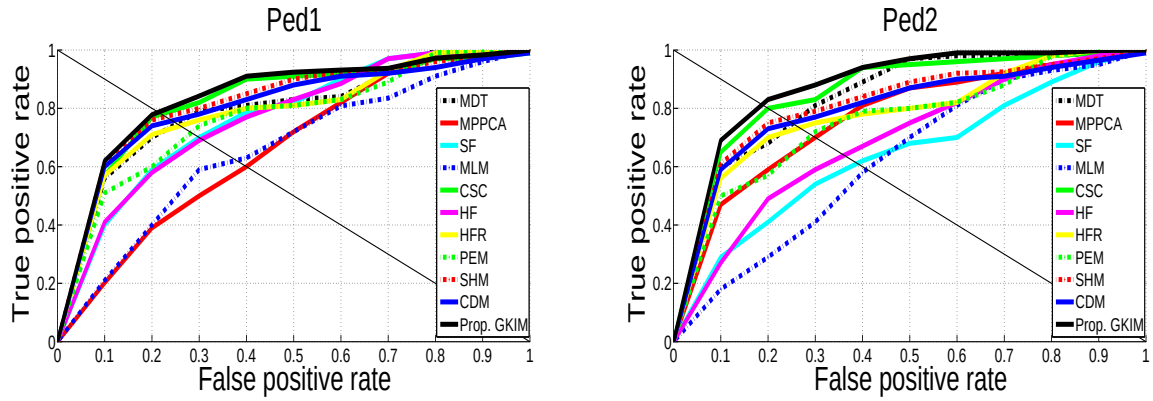

Figure 7: UCSD-ROC curves. For both UCSD subsets, Ped1 and Ped2, the ROC curves are reported in the first and second columns, respectively. In both cases, our proposed GKIM method outperforms all the reference methods.

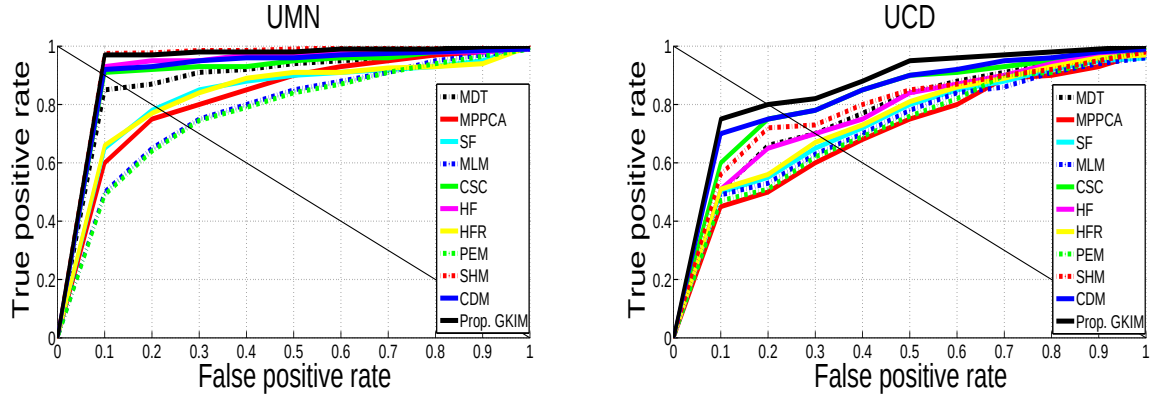

Figure 8: UMN and UCD-ROC curves. The ROC curves for both UMN and UCD datasets are reported in the first and second columns, respectively. In both cases, our proposed GKIM method outperforms all the reference methods.

tion. For example, when considering only spatial information, the temporal information in Jacobian matrix are replaced by 1. In Table 6, we can see that the impact of considering our Gaussian kernel based spatio-temporal features is significant. The performance of considering only spatial information is the lowest. However, using only temporal information shows better results than using only spatial information. In fact, the temporal information uses the spatial information implicitly during its calculations.

4.5. Performance analysis of RNN, CRF and R-CRF classification models

In Table 7, we provide results considering only recurrent neural network (RNN) [49], only conditional random field (CRF) [15], and R-CRF [14] for all the three datasets. For training the RNN and CRF individually, we use the same labeled samples to maintain consistency with the training stage of the R-CRF. In a nutshell, the same procedure is followed for training the RNN, CRF, and R-CRF. We can see in in Table 7 that the R-CRF performs better than both RNN and CRF. In fact, the R-CRF integrates the strengths of the RNN and the CRF by taking the advantage of the discrimination ability of the CRF and the feature learning ability of the RNN.

4.6. Sensitivity Analysis

To demonstrate the robust performance of our proposed GKIM method for the three datasets, we evaluated it using 25 different parameter configuration set as listed in Table 8. These configurations are encoded

Table 8: Configuration set. For sensitivity analysis for our proposed method, 25 different configurations are listed based on patch gridding, the threshold ϵ, lobe scale k, and the parameter ψ. In the patch gridding, 1x, 2x, 3x, 4x, and 5x represent 1x1, 2x2, 3x3, 4x4, and 5x5, respectively.

Param. \ **Config.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
Patch gridding	1x	2x	2x 3x	2x 3x 4x	2x 3x 4x 5x	2x 3x 4x 5x	2x 3x 4x	2x 3x	2x	1x	1x	2x	2x 3x	2x 3x 4x	2x 3x 4x 5x	1x	2x	2x 3x	2x 3x 4x	2x 3x 4x 5x	1x	2x	2x 3x	2x 3x 4x	2x 3x 4x 5x
ϵ	0	1	2	3	4	0	1	2	3	4	4	3	2	1	0	0	1	2	3	4	0	1	2	3	4
k	2	4	6	8	10	2	4	6	8	10	2	4	6	8	10	10	8	6	4	2	2	4	6	8	10
ψ	.1	.3	.6	.9	1	.1	.3	.6	.9	1	.1	.3	.6	.9	1	.1	.3	.6	.9	1	1	.9	.6	.3	.1

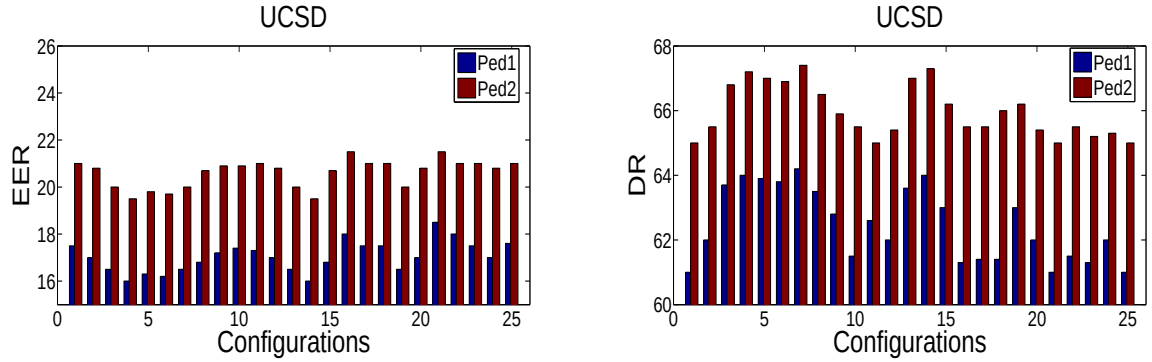

Figure 9: UCSD-equal error rate (EER) and detection rate (DR). The average EER and average DR for our proposed method for UCSD dataset are presented in the left and right columns, respectively. The variations in the results are not significant except from configurations 4 to 5, 6 to 7, 14 to 15, 15 to 16, 19 to 20, 20 to 21, and 24 to 25. These changes are due to the changes in the patch gridding.

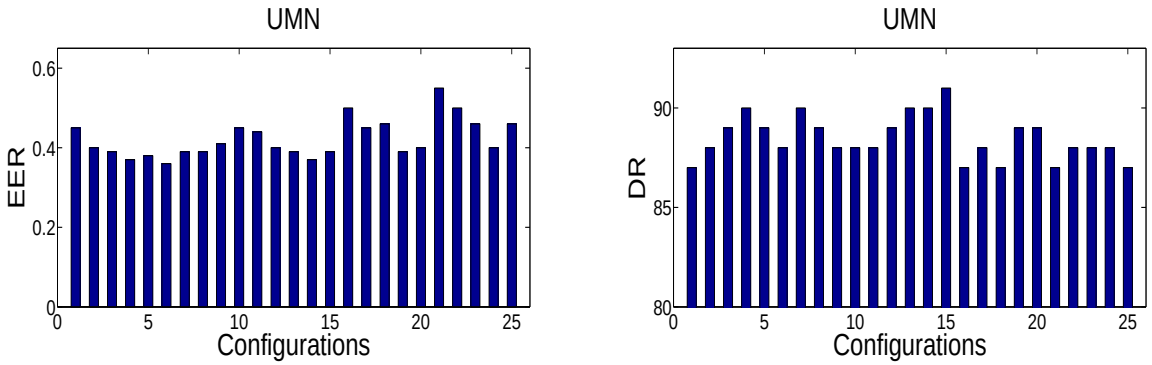

Figure 10: UMN-equal error rate (EER) and detection rate (DR). The average EER and average DR for our proposed method for UMN dataset are presented in the left and right columns, respectively. The variations in the results are not significant except from configurations 4 to 5, 6 to 7, 14 to 15, 15 to 16, 19 to 20, 20 to 21, and 24 to 25. These changes are due to the changes in the patch gridding.

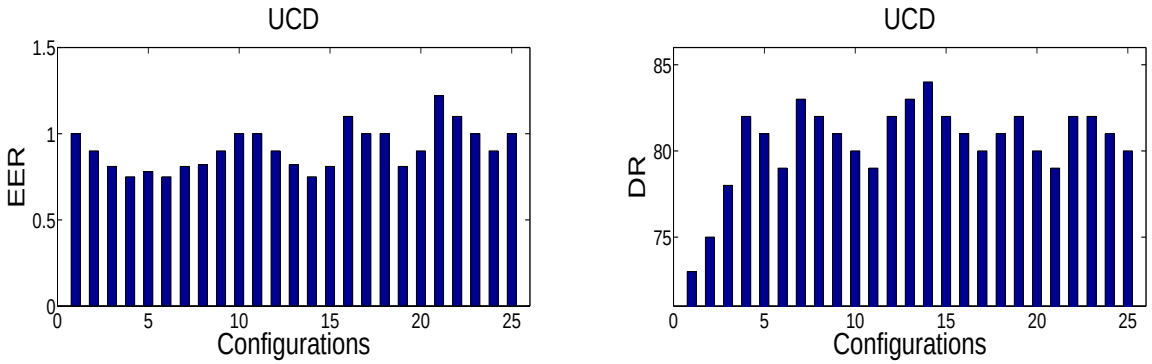

Figure 11: UCD-equal error rate (EER) and detection rate (DR). The average EER and average DR for our proposed method for UCD dataset are presented in the left and right columns, respectively. The variations in the results are not significant except from configurations 4 to 5, 6 to 7, 14 to 15, 15 to 16, 19 to 20, 20 to 21, and 24 to 25. These changes are due to the changes in the patch gridding.

in the experiments using different patch gridding, the threshold ϵ, the lobe scale k, and the parameters ψ. For this purpose, five different patch gridding, the threshold ϵ, the scale lobe k and the parameter ψ are taken into account to maintain consistency in the parameter variations. In Fig. 9 we present the results for both ped1 and ped2 of the UCSD dataset. For UMN and UCD datasets, we present the results in Fig. 10 and Fig. 11, respectively. In Figure 9, gradual changes in the performances in terms of both average EER and average DR can be noticed from configuration 1 to 4, 6 to 10, 11 to 14, 16 to 19, and 24 to 25. However, changes in the performances are significant from configurations 4 to 5, 6 to 7, 14 to 15, 15 to 16, 19 to 20, 20 to 21, and 24 to 25. In Figure 10 and Figure 11, similar changes in the performances for the same configurations can be noticed for UMN and UCD datasets, respectively. In fact, increasing the patch gridding from configuration 4 to 5 decrease the performance. Similarly, decreasing the patch gridding from configuration 6 to 8 improve the performance. The affect can be noticed in the other configurations. Hence, it is worth to increase the patch gridding from 1x1 to 4x4. However, the performance declines by considering other patch gridding. Therefore, the performance of our method does not change significantly by changing other parameters except the patch gridding.

4.7. Computational overheads

To find the computational overhead, a 16GB RAM computer with a 3.5 GHz CPU is used to carry out the experiments. It is worth noticing that the computational complexities can be further reduced since these implementations are not optimized. In Table 9, we presented the computational complexities of our GKIM method and 10 reference methods. These complexities are provided in term of average number of seconds per frame over all the datasets for the MDT [4], the MPPCA [17], the SF [3], the MLM [18], the CSC [7], the HF [8], HFR [19], PEM [20], SHM [21], the CDM [9], and our GKIM method. Comparing to other methods, our GKIM method executes a video frame in 01 second on average which is better than six reference methods including MDT [4], the CSC [7], HFR [19], PEM [20], SHM [21], and the CDM [9]. The MLM [18] and the HF [8] present the same computational complexities. The MPPCA [17] and the SF [3] present better execution times at the cost of significant declines in performances.

Table 9: Computational complexity. Time represents the complexity of each method in term of number of seconds required to process a video frame. Our GKIM method shows execution time better than six reference methods.

Methods	Time	Methods	Time
MDT [4]	25	HFR [19]	06
MPPCA [17]	.9	PEM [20]	04
SF [3]	.5	SHM [21]	07
MLM [18]	01	CDM [9]	03
CSC [7]	05	Proposed	
HF [8]	01	GKIM	01

5. Conclusion

We propose a novel GKIM method for anomalous entities detection and localization in pedestrian flows. The GKIM represents high quality description of anomalous entities in term of most distinctive information. The performance of our proposed method is tested on three datasets and compared to 10 closely related state-of-the-art methods. The performance metrics EER, DR, and ROC curves show that our method outperforms all the reference methods in both frame-level and pixel-level analysis.

As a future work, we would also extend our proposed method to detect various other anomalies.

Acknowledgements

The work described in this paper is jointly supported by the University of Hail, Saudi Arabia and COMSATS Institute of Information Technology, Pakistan.

References

[1] J. D. Dávila, Chapter three cities as innovation towards a new understanding of population growth, social inequality and urban sustainability, Cities in the 21st Century, Routledge (2016) 26.

[2] Y. Zhou, S. Yan, T. S. Huang, Detecting anomaly in videos from trajectory similarity analysis, in: International conference on multimedia and Expo, IEEE ICME, 2007, pp. 1087–1090.

[3] R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model, in: IEEE CVPR, 2009, pp. 935–942.

[4] V. Mahadevan, W. Li, V. Bhalodia, N. Vasconcelos, Anomaly detection in crowded scenes, in: IEEE CVPR, 2010, pp. 1975–1981.

[5] Y. Cong, J. Yuan, J. Liu, Abnormal event detection in crowded scenes using sparse representation, Journal of pattern recognition, Elsevier PR 46 (7) (2013) 1851–1864.

[6] D. Xu, R. Song, X. Wu, N. Li, W. Feng, H. Qian, Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts, Journal of neurocomputing, Elsevier NC 143 (2014) 144–152.

[7] N. Li, X. Wu, D. Xu, H. Guo, W. Feng, Spatio-temporal context analysis within video volumes for anomalous-event detection and localization, Journal of neurocomputing, Elsevier NC 155 (2015) 309–319.

[8] M. Marsden, K. McGuinness, S. Little, N. E. O'Connor, Holistic features for real-time crowd behaviour anomaly detection, in: International conference on image processing, IEEE ICIP, 2016, pp. 918–922.

[9] I. R. de Almeida, V. J. Cassol, N. I. Badler, S. R. Musse, C. R. Jung, Detection of global and local motion changes in human crowds, Transactions on circuits and systems for video technology, IEEE TCSVT 27 (3) (2017) 603–612.

[10] D. Ilias, M. C. EL MEZOUAR, N. Taleb, M. Elbahri, An edge-based method for effective abandoned luggage detection in complex surveillance videos, Computer vision and image understanding, Elsevier CVIU 158 (2017) 141–151.

[11] C. Cuevas, R. Martinez, D. Berjón, N. Garcia, Detection of stationary foreground objects using multiple nonparametric background-foreground models on a finite state machine, Transactions on image processing, IEEE TIP 26 (3) (2017) 1127–1142.

[12] X. Yang, K.-T. Cheng, Local difference binary for ultrafast and distinctive feature description, Transactions on pattern analysis and machine intelligence, IEEE PAMI 36 (1) (2014) 188–194.

[13] J. Byrne, Nested motion descriptors, in: International conference on computer vision and pattern recognition, IEEE CVPR, 2015, pp. 502–510.

[14] K. Yao, B. Peng, G. Zweig, D. Yu, X. Li, F. Gao, Recurrent conditional random fields, in: Proeedings of neural information processing systems, NIPS Deep Learning Workshop, 2013.

[15] J. Lafferty, A. McCallum, F. C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: International conference on machine learning, ICML, 2001.

[16] K. Yao, G. Zweig, M.-Y. Hwang, Y. Shi, D. Yu, Recurrent neural networks for language understanding., in: Interspeech, 2013, pp. 2524–2528.

[17] J. Kim, K. Grauman, Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates, in: IEEE CVPR, 2009, pp. 2921–2928.

[18] A. Adam, E. Rivlin, I. Shimshoni, D. Reinitz, Robust real-time unusual event detection using multiple fixed-location monitors, Pattern Analysis and Machine Intelligence, IEEE Transactions on 30 (3) (2008) 555–560.

[19] K.-W. Cheng, Y.-T. Chen, W.-H. Fang, Video anomaly detection and localization using hierarchical feature representation and gaussian process regression, in: International conference on computer vision and pattern recognition, IEEE CVPR, 2015, pp. 2909–2917.

[20] S. Yi, H. Li, X. Wang, Pedestrian behavior modeling from stationary crowds with applications to intelligent surveillance, Transactions on image processing, IEEE TIP 25 (9) (2016) 4354–4368.

[21] Y. Zhang, H. Lu, L. Zhang, X. Ruan, Combining motion and appearance cues for anomaly detection, Journal of pattern recognition, Elsevier PR 51 (2016) 443–452.

[22] M. Devanne, S. Berretti, P. Pala, H. Wannous, M. Daoudi, A. Del Bimbo, Motion segment decomposition of rgb-d sequences for human behavior understanding, Pattern Recognition, Elsevier PR 61 (2017) 222–233.

[23] T. Lai, H. Wang, Y. Yan, T.-J. Chin, W.-L. Zhao, Motion segmentation via a sparsity constraint, IEEE Transactions on Intelligent Transportation Systems, IEEE ITS.

[24] F. Husain, H. Schulz, B. Dellen, C. Torras, S. Behnke, Combining semantic and geometric features for object class segmentation of indoor scenes, Robotics and Automation Letters, IEEE RAL 2 (1) (2017) 49–55.

[25] B. Poling, G. Lerman, A new approach to two-view motion segmentation using global dimension minimization, International journal of computer vision, Springer IJCV 108 (3) (2014) 165–185.

[26] C. Qin, G. Zhang, Y. Zhou, W. Tao, Z. Cao, Integration of the saliency-based seed extraction and random walks for image segmentation, Journal of neurocomputing, Elsevier NC 129 (2014) 378–391.

[27] B. Zhong, Y. Chen, Y. Chen, R. Ji, Y. Chen, D. Chen, H. Wang, Background subtraction driven seeds selection for moving objects segmentation and matting, Journal of neurocomputing, Elsevier NC 103 (2013) 132–142.

[28] Q. Wu, Y. Gan, B. Lin, Q. Zhang, H. Chang, An active contour model based on fused texture features for image segmenta-

tion, Journal of neurocomputing, Elsevier NC 151 (2015) 1133–1141.

[29] M. Kumar, C. Bhatnagar, Zero-stopping constraint-based hybrid tracking model for dynamic and high-dense crowd videos, The imaging science journal, Taylor & Francis ISJ 65 (2) (2017) 75–86.

[30] Z. Li, J. Guo, L.-F. Cheong, S. Z. Zhou, Perspective motion segmentation via collaborative clustering, in: International conference on computer vision, IEEE ICCV, 2013, pp. 1369–1376.

[31] A. Mumtaz, W. Zhang, A. B. Chan, Joint motion segmentation and background estimation in dynamic scenes, in: International conference on computer vision and pattern recognition, IEEE CVPR, 2014, pp. 368–375.

[32] B. Krausz, C. Bauckhage, Loveparade 2010: Automatic video analysis of a crowd disaster, Journal of Computer Vision and Image Understanding, Elsevier CVIU 116 (3) (2012) 307–319.

[33] V. Kaltsa, A. Briassouli, I. Kompatsiaris, L. J. Hadjileontiadis, M. G. Strintzis, Swarm intelligence for detecting interesting events in crowded environments, Transactions on image processing, IEEE TIP 24 (7) (2015) 2153–2166.

[34] W. Li, V. Mahadevan, N. Vasconcelos, Anomaly detection and localization in crowded scenes, Transactions on pattern analysis and machine intelligence, IEEE PAMI 36 (1) (2014) 18–32.

[35] S. Wu, H.-S. Wong, Z. Yu, A bayesian model for crowd escape behavior detection, Transactions on circuits and systems for video technology, IEEE CSVT 24 (1) (2014) 85–98.

[36] H. Ullah, N. Conci, Crowd motion segmentation and anomaly detection via multi-label optimization, in: ICPR Workshop on pattern recognition and Crowd Analysis, 2012.

[37] R. Mehran, B. E. Moore, M. Shah, A streakline representation of flow in crowded scenes, in: European conference on computer vision, Springer ECCV, (2010), pp. 439–452.

[38] S. Ali, M. Shah, A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis, in: International conference on computer vision and pattern recognition, IEEE CVPR, (2007), pp. 1–6.

[39] S. Shadden, F. Lekien, J. Marsden, Definition and properties of lagrangian coherent structures from finite-time lyapunov exponents in two-dimensional aperiodic flows, Physica D: Nonlinear Phenomena 212 (3) (2005) 271–304.

[40] M. Calonder, V. Lepetit, C. Strecha, P. Fua, Brief: Binary robust independent elementary features, European Conference on Computer Vision, Springer ECCV (2010) 778–792.

[41] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, Orb: an efficient alternative to sift or surf, in: International Conference on Computer Vision, IEEE ICCV, 2011, pp. 2564–2571.

[42] S. Leutenegger, M. Chli, R. Y. Siegwart, Brisk: Binary robust invariant scalable keypoints, in: International Conference on Computer Vision, IEEE ICCV, 2011, pp. 2548–2555.

[43] E. P. Simoncelli, W. T. Freeman, The steerable pyramid: A flexible architecture for multi-scale derivative computation, in: International Conference on Image Processing, IEEE ICIP, 1995, p. 3444.

[44] K. Murakami, H. Taguchi, Gesture recognition using recurrent neural networks, in: SIGCHI conference on Human factors in computing systems, ACM SIGCHI, 1991, pp. 237–242.

[45] A. Graves, A.-r. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: International Conference on Acoustics, Speech and Signal Processing, IEEE ICASSP, 2013, pp. 6645–6649.

[46] Unusual crowd activity dataset of university of minnesota, available from http://mha.cs.umn.edu/movies/crowd-activity-all.avi.

[47] B. Antić, B. Ommer, Video parsing for abnormality detection, in: International conference on computer vision, IEEE ICCV, 2011, pp. 2415–2422.

[48] H. Ullah, M. Ullah, N. Conci, Real-time anomaly detection in dense crowded scenes, in: IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics, 2014, pp. 902608–902608.

[49] T. Mikolov, G. Zweig, Context dependent recurrent neural network language model., SLT 12 (2012) 234–239.