Spatial-temporal Convolutional Neural Networks for Anomaly Detection and Localization in Crowded Scenes

Shifu Zhou, Wei Shen, Dan Zeng, Mei Fang, Yuanwang Wei, Zhijiang Zhang

www.elsevier.com/locate/image

PII: S0923-5965(16)30093-5
DOI: http://dx.doi.org/10.1016/j.image.2016.06.007
Reference: IMAGE15106

To appear in: *Signal Processing : Image Communication*

Received date: 7 March 2016
Revised date: 10 May 2016
Accepted date: 17 June 2016

Cite this article as: Shifu Zhou, Wei Shen, Dan Zeng, Mei Fang, Yuanwang Wei and Zhijiang Zhang, Spatial-temporal Convolutional Neural Networks fo Anomaly Detection and Localization in Crowded Scenes, *Signal Processing Image Communication,* http://dx.doi.org/10.1016/j.image.2016.06.007

Noname manuscript No.
(will be inserted by the editor)

Spatial-temporal Convolutional Neural Networks for Anomaly Detection and Localization in Crowded Scenes

Shifu Zhou · Wei Shen · Dan Zeng · Mei Fang · Yuanwang Wei · Zhijiang Zhang

Received: date / Accepted: date

Abstract Abnormal behavior detection in crowded sc- enes is extremely challenging in the field of computer vision due to severe inter-object occlusions, varying crowd densities and the complex mechanics of a human crowd. We propose a method for detecting and locating anomalous activities in video sequences of crowded scenes. The key novelty of our method is the coupling of anomaly detection with a spatial-temporal Convolutional Neural Networks (CNN), which to the best of our knowledge has not been previously done. This architecture allows us to capture features from both spatial and temporal dimensions by performing spatial-temporal convolutions, thereby, both the appearance and motion information encoded in continuous frames are extracted. The spatial-temporal convolutions are only performed within spatial-temporal volumes of moving pixels to ensure robustness to local noise, and increase detection accuracy. We experimentally evaluate our model on benchmark datasets containing various situations with human crowds, and the results demonstrate that the proposed approach surpass state-of-the-art methods.

Keywords Spatial-temporal CNN · Anomaly detection · Crowded scene · Surveillance

1 Introduction

The widespread use of surveillance systems in railway stations, airports, roads or malls has resulted in large volumes of video data. There is an increasing need not only for recognition of objects and their behaviors, but also in particular for detecting abnormal behavior in the large body of ordinary data. Automatically detecting abnormal activities or events from long duration video sequences is

Shifu Zhou
Key Lab of Specialty Fiber Optics and Optical Access Networks, Shanghai University
E-mail: sfzhoujob@shu.edu.cn

crucial for intelligent surveillance [39], behavior analysis [37], and security applications [11]. In particular, abnormal behavior detection in crowded scenes is a challenging problem due to the large number of pedestrians in close proximity, the volatility of individual appearance, the frequent partial occlusions that they produce, and the irregular motion pattern of the crowd. In addition, there are potential dangerous activities in crowded environments, such as crowd panic, stampedes and accidents involving a large number of individuals, which make automated scene analysis in the most need.

One of the main challenges is to detect anomalies both in time and space domains [9]. This implies to find out which frames that anomalies occur (we refer to it as frame level) and to localize regions that generate the anomalies within these frames (we refer to it as pixel level) [30]. Another fundamental limitation is that there is no commonly accepted definition of anomaly. The definition of anomaly varies significantly depending on the given scenario. Conventionally, anomalies are identified as those events that display a low probability or a significant difference of occurring based on earlier observations [16,27, 44]. The existing approaches for detecting anomalies can be classified into two categories: 1) object-centric approachs [41,33], 2) holistic methods [29,32]. In a typical object-centric approach, the crowd is treated as a set of individuals. To understand the crowd activities, it needs to segment the crowd of interest into objects. However, these object-centric approaches face considerable complexity in detecting objects, tracking trajectories, and recognizing behaviors in crowded scenes. One common drawback among these methods is that they are not capable of handling crowded scenes. Once the density of objects increases, a degradation of their performance is observed.

For the holistic methods, the aim is not to detect and track each individual. Instead, the crowd is considered as a whole entity. Typical features, such as spatial-temporal gradients or optical flow data, are used in these methods to localize regions with dramatic motion. Through modeling the normal/abnormal crowd motion patterns, anomaly detection is carried out by pre-trained classifiers.

In fact, anomaly detection can be arguably defined as a binary classification problem, i.e., activities of the crowd are classified as either normal or abnormal. Recently, many works have demonstrated the power of CNN [24] in a wide variety of computer vision tasks, such as object classification and detection [23,36], text recognition [13], edge detection [38], and face recognition [40]. For video classification tasks, the CNN also show potential applicable value. Ji et al. [17] propose a 3D CNN for human action recognition based on video sequence. Wang et al. [42] develop a novel reconfigurable CNN for automatic 3D human activity recognition from RGB-D videos. Maturana et al. [28] employ a 3D CNN to detect save landing zones for autonomous helicopters from LiDAR point clouds. Encouraged by these surprising results, we are interested in anomaly detection in crowded scenes. However, motion patterns of the crowd show both spatial and temporal characters. To detect and localize anomalous events in video sequences of crowded scenes, we develop a spatial-temporal CNN model, which accesses to not only the appear-

ance information present in a single, static image, but also complex motion information extracted from continuous frames. To capture anomalous events appearing in a small part of the frame, the spatial-temporal CNN model is applied only on spatial-temporal volumes of interest (SVOI), which not only ensures the robustness to noise, but also achieves a lower computational cost. The experimental results on available benchmark datasets show that our algorithm outperforms state-of-the-art methods. Compared with previous works, our method possesses two advantages: 1) it acts directly on the raw inputs (SVOI) without any preprocessing instead of relying on hand-crafted features. 2) it does not rely on foreground segmentation results because only motion and appearance information is considered in our method. Once the crowd is moving, the motion patterns and texture of the crowd would be captured by the spatial-temporal CNN model and a reasonable anomaly detection results would be achieved.

2 Related work

In this section, we give a review to the related works in previous. A considerable amount of literature has been published on anomaly detection from static cameras [43, 10, 6]. However, most of these works are limited to sparse scenarios, where detailed visual information of each individual can be captured. Once the density of objects increases, such information cannot be easily extracted by traditional methods. Hence, a lot of algorithms have been presented to address crowded scenes.

Several physics-inspired models have been proposed for crowd representation and combined with computer vision techniques to analysis crowd activity. Mehran et al. [29] propose a novel approach to detect and localize anomalous behaviors in crowd videos by using social force model (SFM) [15]. Ali et al. [2] construct a finite time Lyapunov exponent (FTLE) field, whose boundaries vary with the changes of the crowd in terms of the dynamic behavior of the flow, to detect global anomalies in crowd scenes. Zhou et al. [47] propose a descriptor to detect the collectiveness of the crowd. Raghavendra et al. [32] propose the particle swarm optimization (PSO) method for optimizing the interaction force. The goal of their method is to drift the population of particles toward the areas of the main image motion. Xiong et al. [45] introduce a novel approach to detect two typical abnormal activities: pedestrian gathering and running. This approach is based on the potential energy model [46] and kinetic energy. The kinetic energy is determined by the crowd distribution index (CDI), which represents the dispersion of the crowd. The abnormal activities are detected through threshold analysis.

Motion based methods typically extract motion information, such as optical flow, trajectories, from spatial-temporal volumes, and then model motion patterns of the crowd. Krausz et al. [22] extract the histogram of optical flow to represent the global motion pattern of the crowd and derive statistics from it to model behaviors of pedestrians. Then, specific dangerous behav-

iors are detected by using a set of heuristic rules. Wu et al. [44] introduce an approach for crowd flow modeling and anomaly detection on both structured and unstructured scenes. Particle advection is based on optical flow, and representative trajectories of a crowd flow are obtained by clustering particle trajectories. Chaotic dynamics of representative trajectories is utilized to build a model capturing an outlier that moves in a different pattern. This approach works well in very dense scenes where a global motion pattern exists. However, it is unable to detect local anomalies that occur in a small region in a frame. In [21], a framework is proposed to model local spatial-temporal motion patterns for crowded scenes with high density. In this work, the underlying motion patterns of spatial-temporal cuboids are characterized by 3D Gaussian distributions and KL divergence is used as a distance measure to valuate similarities among cuboids in the same spatial location. For each location, the temporal relationship among local motion patterns is modeled by a Hidden Markov Model (HMM) and the spatial relationship is captured via a coupled HMM. The experimental results show that the proposed approach is suitable to extremely crowded scenes, however, many false positives can be observed in sparse scenarios. Since each location is modeled by only one HMM, the approach can work only for limited kinds of normal behaviors or specific crowded scenes. The detection rate of the abnormal behaviors decreases when the types of normal behavior are changed. Mousavi et al. [30] present a histogram of oriented tracklets to capture the motion patterns of the crowd. In their work, video sequences are divided into spatial-temporal cuboids, where the tracklets passing through are collected for statistics. Frames are identified as normal or abnormal by using Latent Dirichlet Allocation [5] and Support Vector Machines [14]. Another interesting work in the same domain, is that of Chen et al. [7] who formulate the extraction of normal interactions from training video as the problem of efficiently finding the frequent geometric relations of the nearby sparse spatial-temporal interest points. Nam et al. [31] propose a real-time abnormal situation detection method in crowded scenes based on the crowd motion characteristics, such as particle energy and motion direction.

Recently, significant researches have taken on combining motion information and appearance information to detect anomalies that move similarly to the normal motion pattern. Li et al. [26] use a set of mixture of dynamic textures model to infer multi-scale spatial and temporal anomaly maps, which act as potentials of a conditional random field that guarantees global consistency of the anomaly judgments. However, the approach takes around 23 sec per frame, which makes it impossible for many applications. The joint model of motion and context information is presented in [48] for detecting abnormal events in crowded scenes. The context pattern information is extracted through local binary patterns on three orthogonal planes, while the motion information is captured by a feature descriptor named Multi-scale Histogram of Frequency Coefficient. Then a sparse reconstruction cost from a learned event dictionary is adopted to classify local normal and abnormal events. The main drawback of the approach is that the classification of each cuboid is de-

termined by a pre-defined threshold, which makes the approach be sensitive to input video. Kaltsa et al. [19] propose an interesting work, which jointly considers motion and appearance information to distinguish different kinds of anomalies. In their work, histograms of oriented swarms (HOS), which is applied to capture the dynamics of crowded flow, together with histograms of oriented gradients (HOG) to form a descriptor. Bertini et al. [4] propose a multi-scale non-parametric approach that detects and localizes anomalies, using dense local spatial-temporal features that model both appearance and motion of persons and objects. This approach aims to handle the problem of high variability in unusual events and deal with scene changes that happen in real world by using a model updating procedure. However, the performance of this approach is unsatisfied. Another work that uses densely constructed spatial-temporal video volumes to learn video events at each pixel without supervision is [34]. In this work, a hierarchical codebook model for the dominant behaviors is construct by employing contextual graphs of video volumes. However, this work only employs HOG descriptor and omits essential information that could led to better results.

However, most previous works are relying on hand-crafted features. However, it is difficult to extract useful hand-crafted features in different scenarios. Instead, our method acts directly on the raw inputs (video sequences) and automatically extracts high-level features, which leads to remarkable performance in anomaly detection. Moreover, our method can detect different kinds of anomalous activities, and is not limited to specified behaviors.

3 Methodology

In this work, we focus on the challenge of detecting anomalies in both time and space in video sequences with crowds of varying densities. Both motion and appearance features extracted by our spatial-temporal CNN model are used to effectively and robustly capture these anomalies for a wide range of scenarios.

3.1 Potential spatial-temporal volumes extraction

To detect and localize anomalies that take place in local regions, typical methods in previous works [21, 26, 25] set a large amount of non-overlapping patches or densely sample on each frame and process all patches regardless of whether these patches carry motion information or not. Therefore, these methods lead to high computational cost and frequent false alarms. To achieve higher precision and lower computational cost, those patches that contain little motion information, should be abandoned. For instance, local regions where no pedestrians appear or move, are unnecessary to be processed. Our spatial-temporal CNN uses data derived from local SVOI instead of entire video frames, so as to only handle pixels carry rich information relevant to the event taking place.

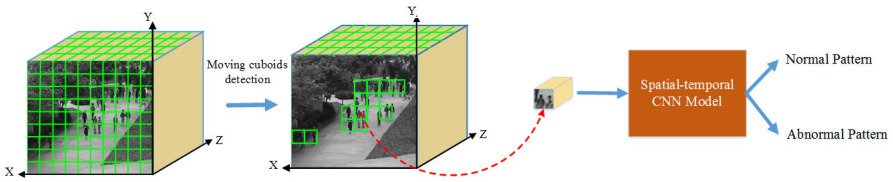

Fig. 1 Overview of the proposed approach for anomaly detection.

To extract the SVOIs, we apply optical flow, as it has been proven to be a useful low-level feature for motion detection and object tracking in many computer vision applications. We define a set of non-overlapping patches of fixed size to cover entire video frames. The size of patches needs to be set only once for each camera, or each dataset due to the static nature of surveillance cameras. For the UCSD dataset [27], a rectangular area of 15×15 pixels is chosen, as it is small enough to capture anomaly location, but at the same time large enough to extract related details of appearance. To capture motion information, continuous frames of the same patch are stacked to construct a SVOI. The temporal length of the SVOI is set to 7 frames, which provide a reasonable tradeoff between the ability to detect anomalies and storage required for anomaly detection. Hence, the size of SVOI is chosen as 15×15×7. Once SVOIs are extracted, the optical flow values of pixels within them are calculated. The resulting SVOIs and pixels within them are considered informative and are preserved if at least 65% of pixels of SVOI are moving, otherwise that pixels and its SVOI are treated as noisy and are discarded. Since the spatial size of SVOIs derived from different datasets or cameras may vary, SVOIs are resized in a fixed size before being fed into spatial-temporal CNN Model. Then, the spatial-temporal CNN model performs spatial-temporal convolution operation on SVOIs to automatically extract high-level spatial-temporal features that effectively characterize video dynamic and help identify both global and local anomalies. An overview of the procedure for detecting and localizing anomalies is depicted in Fig.1.

3.2 Spatial-temporal convolution

CNN is a local connected neural network model, which extracts local features by restricting the receptive fields of the hidden units. A typical CNN is comprised of a number of convolutional layers (often with a subsampling step) and then followed by one or more fully connected layers. A convolution layer uses a set of filters that process small local regions of the input where these filters are replicated along the whole input space. A subsampling step generates a lower resolution version of the convolution layer activations by taking the maximum filter activation from different positions within a specified window. The subsampling step adds translation invariance and tolerance to minor differences of positions of objects parts. In CNN, convolution is performed at the

convolution layers. Formally, the j-th feature map a_{ij} in i-th layer is given by:

$$a_{ij} = f(W_n * a_{(i-1)n} + b_{ij}). \tag{1}$$

Where $f(.)$ is a nonlinear function, defined as $f(x) = max(0, x)$. W is the kernel of filters, n index the set of feature maps that connected to the current feature map in the (i-1)-th layer, $*$ denotes the convolution operation, b_{ij} is the scalar bias term for current feature map. W and b need to be trained for better extracting image feature.

To effectively extract both appearance and dynamic information for observations, spatial-temporal convolutions are performed in the convolution layers of CNN. The spatial-temporal convolution is achieved by convolving a 3D kernel to the spatial-temporal volume. By these spatial-temporal convolutions, the feature maps in convolution layers are connected to contiguous frames in previous layers, thus, motion feature is captured. Formally, the spatial-temporal convolution operation between 3D kernel W_n and spatial-temporal volume $a_{(i-1)n}$ can be defined as $[W_n * a_{(i-1)n}](x, y, t)$, which is

$$\sum_{n} \sum_{u=0}^{U_i-1} \sum_{v=0}^{V_i-1} \sum_{r=0}^{R_i-1} W_n^{uvr} a_{(i-1)n}^{(x+u)(y+v)(t+r)} \tag{2}$$

Where U_i, V_i, R_i represents the height, width, and temporal length of 3D kernel, respectively, and $x \times y \times t$ is the size of spatial-temporal volume $a_{(i-1)n}$. An illustration of the spatial-temporal convolution is presented in Fig.2.

3.3 The structure of spatial-temporal CNN

In this sub-section, we describe a spatial-temporal CNN structure that we have developed for anomaly detection in crowded scenes. As shown in Fig.3, the input of the spatial-temporal CNN structure is raw SVOI. The spatial-temporal CNN structure consists of 8 layers. The input size of the spatial-temporal CNN structure is $32 \times 32 \times 7$, corresponding to 7 contiguous frames of 32×32 pixels. The SVOIs described in sub-section 3.1, are resized in fixed resolution $32 \times 32 \times 7$ and then fed into the spatial-temporal CNN model. Firstly, we apply spatial-temporal convolution with a 3D kernel size of $3 \times 3 \times 3$ (3×3 in the spatial dimension and 3 in temporal dimension) on the input data. Note that one 3D kernel can only capture one kind of feature from the SVOI, to increase the number of feature types, total 12 different 3D kernels are applied at the input data, generating 12 feature maps in the layer C_1. The size of the 12 feature maps is $30 \times 30 \times 5$. Then, a 2D kernel size of 3×3 is employed in the next convolution layer. Subsequently, we employ a subsampling operation on each spatial-temporal convolution result. In subsampling layer S_1, each feature map in the layer C_2 is performed subsampling with a factor of 2×2 in spatial domain, which leads to the same number of feature maps with reduced spatial resolution and aims to build robustness to small spatial distortions. To generate another set of feature maps on a deeper layer, we further perform

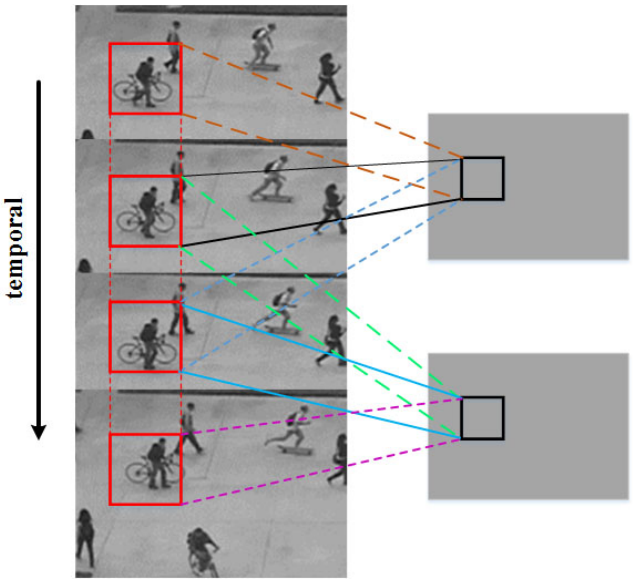

Fig. 2 Illustration of the spatial-temporal convolution across both spatial and temporal domains. In this example, there are two different 3D kernels, whose temporal dimension is both 3. That is to say, each feature map is obtained by performing spatial-temporal convolutions across 3 adjacent frames.

spatial-temporal convolutions on the feature maps. The convolution layer C_3 is obtained by applying spatial-temporal convolution with a 3D kernel size of $3 \times 3 \times 3$ on feature maps of S_1. The connection strategy between S_1 and C_3 follows the same principle described in [12], and layer C_3 consists of 48 feature maps. Layer S_2 provides the same function described above for S_1. Layer C_4 performs $3 \times 3 \times 3$ convolution operation and obtains 64 feature maps. After three layers of spatial-temporal convolution, the temporal dimension of resulting feature maps is reduced to 1. Then, layer C_4 is followed by a fully-connection layer FC_1. In layer FC_1, we further perform 2D convolution operation to capture higher level complex features. The size of the convolution kernel used is 4×4, so that the size of the output feature maps in FC_1 is reduced to 1×1, and each of them is connected to all the 64 feature maps in the layer C_4. The output feature maps in FC_1 are concatenated into a long feature vector. Finally, the fully-connection layer FC_2 is further fully connected with each unit of the feature vector. The number of its output units is 2, which is the same as the number of types of behaviors (abnormal, normal), and each of the units represents the probability of an behavior hypothesis. Let (z_i, i = 1, 2) be the output of unit i in FC_2. We apply logical regression to normalize the probabilities of the output labels. The probabilities of the output in unit 1 and 2 are given by $p_1 = 1/(1 + \exp(z_2 - z_1))$ and $p_2 = 1/(1 + \exp(z_1 - z_2))$, respectively.

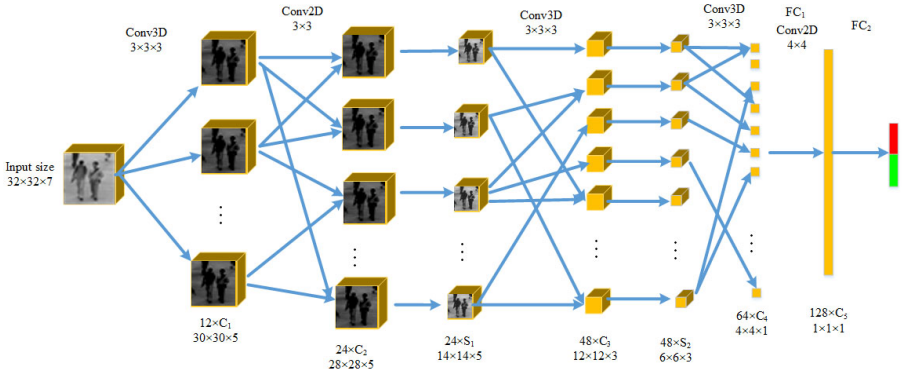

Fig. 3 The spatial-temporal CNN structure used in this paper. It contains 4 convolutional layers, 2 subsampling layers and 2 fully connection layers.

4 Experiments

In this section, we evaluate the effectiveness of our method on four benchmark datasets, i.e. UCSD [27], UMN [29], Subway [1], and U-turn [3], where different kinds of anomalies appear, and compare its performance with state-of-the-art methods. The accuracy of our algorithm both on frame level and pixel level are calculated. As mentioned in sub-section 3.1, the spatial size of SVOI should be large enough to contain useful appearance information, at the same time, small enough to capture local anomaly. Due to the static nature of surveillance cameras, the size of patches is only set once for each camera, or each dataset. Moreover, the temporal length of SVOI should be large enough to provide the perfect ability of detecting anomalies, at the same time, as small as possible to reduce the delays. The temporal length of SVOI is set as 7 empirically and then SVOIs are fed into the spatial-temporal CNN model for anomaly detection.

4.1 UCSD Dataset

The UCSD dataset contains two different scenario subsets Ped1 and Ped2, both of which are surveillance videos acquired by a stationary camera mounted at an elevation, overlooking pedestrian walkways. The crowd density changes over time and alternates between sparse and crowded. The UCSD dataset is a challenging dataset due to low resolution, many occlusions, different kinds of anomalies, and co-occurring in the same frame. Anomalies appear in the UCSD dataset include bikers, skaters, small cars, people walking across a walkway, and wheelchairs moving with different speeds, some cases of which are difficult to find out even for human observers. Ped1 contains videos of people moving towards and away from the camera, with some perspective distortion; Ped2 depicts a scene of people moving horizontally. Ped1 consists of 34 training clips

and 36 testing clips with spatial resolution of 158 × 238 pixels, while Ped2 is comprised of 16 training clips and 12 test clips of 240 × 360 pixels.

To evaluate our method on the UCSD dataset, we adopt the same criterion as state-of-the-art literature. For the UCSD dataset, two criterions are used for evaluating anomaly detection accuracy: a frame level criterion and a pixel level criterion. The frame level criterion focuses on the changes only in temporal domain, predicting which frames contain an anomaly, omitting its spatial location. The frame level criterion labels a frame as abnormal if it contains at least one abnormality, regardless of where it is localized. It should be noted that the frame level criterion only measures temporal localization accuracy, and may sometimes lead to coincidental predictions deriving from false positive appearing in frames that contain true anomalies, without these anomalies having actually been detected. On the other hand, the pixel level criterion localizes both temporal and spatial anomalies. The pixel level criterion requires that at least 40% of all ground-truth abnormal pixels are marked as abnormal, the detection is considered as successful and the frame is identified as abnormal. The pixel level criterion evaluates both the temporal and spatial accuracy of the anomaly detection, thus, it is much stricter and more detailed. By calculating the true positive rate (TPR) and false positive rate (FPR) at different detection thresholds, we obtain Receiver Operating Characteristic (ROC) curves for evaluating the performance of method. To evaluate performance of our method and compare with other state-of-the-art approaches in reliable criterions, the Equal Error Rate (EER) for the frame level criterion and the Detection Rate (DR) for the pixel level criterion are used. EER refers to the error rate of a system when the false positive and false negative rate are equal, with lower equal error rates implying a higher accuracy of the system. DR corresponds to the successful detection rate of the anomalies happening at EER. The higher the DR, the better performance of the system.

We compare quantitatively our method with state-of-the-art methods on the challenging Ped1 and Ped2 benchmarks. The size of SVOIs for UCSD dataset is set as $15 \times 15 \times 7$. Since the training clips of Ped1 and Ped2 do not contain anomalies, we randomly select half testing clips of Ped1 and Ped2 for training model and the rest clips are used as testing samples. we separately train model for Ped1 and Ped2, respectively. We extract 140248 normal samples and 35215 anomalous samples from Ped1, and 63579 normal samples and 20638 anomalous samples from Ped2. Since the number of anomalous samples is much smaller than that of normal samples, which may lead to a class-imbalance problem. To handle this problem, we resample anomalous samples to keep the number of normal and anomalous samples in balance. The number of normal and anomalous samples can ensure good performance of training model. State-of-the-art methods include the use of local low level motion histograms [1], descriptors based on social force flow dynamics [29], the mixture of optical flow observation (MPPCA) [20], the hierarchical mixture of dynamic textures (H-MDT) after applying CRF filtering [26], and the descriptor that combines histograms of oriented swarms (HOS) and histograms of oriented gradients (HOG) [19]. It should be noted that that comparisons with state-of-

Table 1 Performance of various methods by both the frame level criterion and the pixel level criterion in UCSD dataset

Method	Criterion			
	EER		DR	
	Ped1	Ped2	Ped1	Ped2
Adam[1]	38.9%	43.8%	32.6%	22.4%
SF[29]	36.5%	42%	28.3%	27.6%
MPPCA[20]	39.6%	31.1%	19.6%	22.4%
H-MDT(CRF)[26]	**17.8**%	24.7%	74.5%	70.1%
HOS-HOG[19]	27.0%	26.9%	78.9%	74.9%
Ours	24.0%	**24.4** %	**81.3**%	**81.9**%

the-art methods are performed in different number of test clips. As our method requires a part of test clips of anomalous class for model learning, while most of other methods do not require. Table 1 depicts the comparison results on the UCSD dataset under both the frame level criterion (EER) and pixel level criterion (DR). Fig.4 corresponds to ROC curves of UCSD dataset both on pixel level criterion and frame level criterion. It can be observed that our method is superior to all other existing methods for the pixel level criterion in both Ped1 and Ped2 benchmarks. Our method improves upon recently proposed powerful approaches such as H-MDT(CRF) [26] (9.1% gain for DR in Ped1 and 16.8% gain in Ped2) as well as HOS-HOG [19] (3.0% gain for DR in Ped1 and 9.3% gain in Ped2). For the frame level criterion, our method is a little weaker than H-MDT (CRF) and ranks second top. However, as stated above, the frame level criterion does not take into account "lucky co-occurrences". In contrary, the pixel level criterion is more rigorous and more reliable, and rules out these "lucky co-occurrences". Hence, it can be considered that our method achieves better performance than state-of-the-art methods.

To explore the capability of the spatial-temporal CNN model, we conduct a comparison with the frame-based 2D CNN model. The input of the 2D CNN model is a single frame, while the input of spatial-temporal CNN model is continual frames. The spatial-temporal CNN model and the 2D CNN model take the same spatial dimension of input, i.e. 32×32. The comparison results of both cases are presented in Table 2. It can be observed that the spatial-temporal CNN model outperforms the frame-based 2D CNN model. In both Ped1 and Ped2, the spatial-temporal CNN model achieves a higher DR for pixel level criterion as well as a lower EER for frame level criterion. The frame-based 2D CNN model only captures appearance information, while the spatial-temporal CNN model extracts not only appearance feature but also motion information. The results in Table 2 demonstrate that incorporating the motion information into CNN model can effectively improve the performance of method.

The examples of anomaly detection and localization in Ped1 and Ped2 are shown in Fig.5 and Fig.6, respectively. As can be observed, several anomalous events that co-occur in the same frame are correctly detected and localized by our method. However, our method maybe fail in the cases, such as "anomalies"

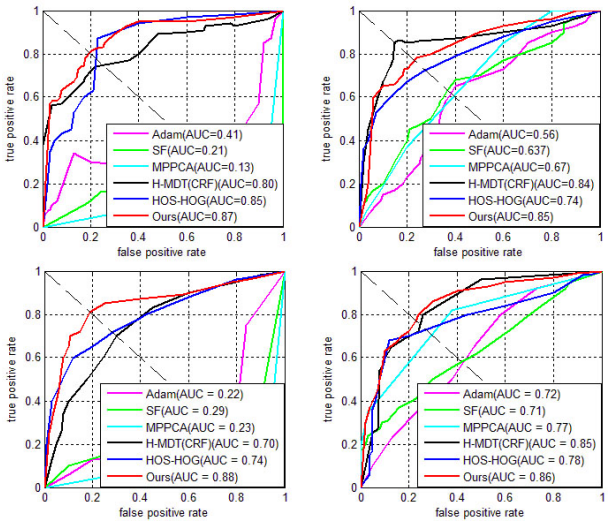

Fig. 4 ROC curves for UCSD dataset: top row : Ped1, bottom row: Ped2, left column: pixel level criterion, right column: frame level criterion.

Table 2 Performance of various CNN model

Method	Criterion			
	EER		DR	
	Ped1	Ped21	Ped1	Ped21
2D CNN	40.1%	37.9%	59.7%	68.3%
Spatial-temporal CNN	24.0%	24.4%	81.3%	81.9%

are obscured surrounding pedestrians, or "anomalies" do not appear in training set or are dissimilar with the samples of training set.

4.2 UMN dataset

The UMN dataset contains three different escape scenes (named as lawn, indoor, and plaza, respectively), in which pedestrians initially walk randomly and exhibit a sudden evacuation in different directions in the end. In normal situation, pedestrians are walking or standing in the scene, however, in abnormal cases, pedestrians escape in panic. As the dataset only provides global ground-truth abnormality maps, we follow the standard Area Under Curve (AUC) criterion, which is the area under the ROC curve. The size of SVOIs for UMN dataset is set as $20 \times 20 \times 7$. In scenes lawn, indoor, and plaza, there are 2, 6, and 3 scenarios of crowd escape events, respectively. For training, frames of 1, 3, 2 scenarios from scenes lawn, indoor, and plaza are utilized to train the spatial-temporal CNN model, while the remaining frames are used for testing. Since frames of UMN have the save resolution, we train a model for UMN. We extract 112029 normal samples and 31644 anomalous samples from

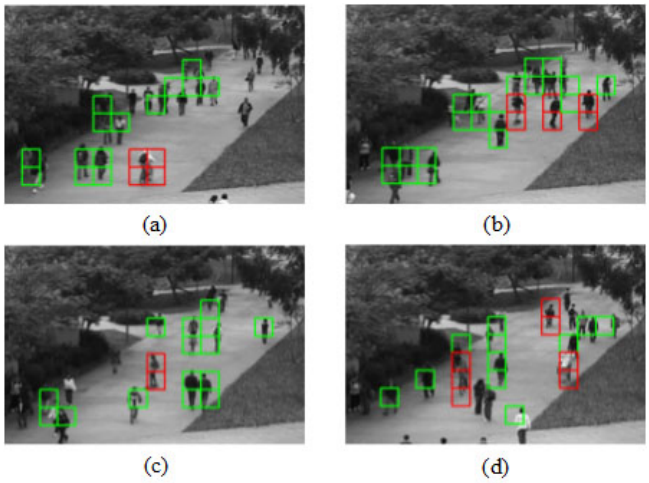

Fig. 5 Examples of anomaly detection and location in Ped1. Green rectangles show the motion areas while red rectangles denote anomalies detected by our method.

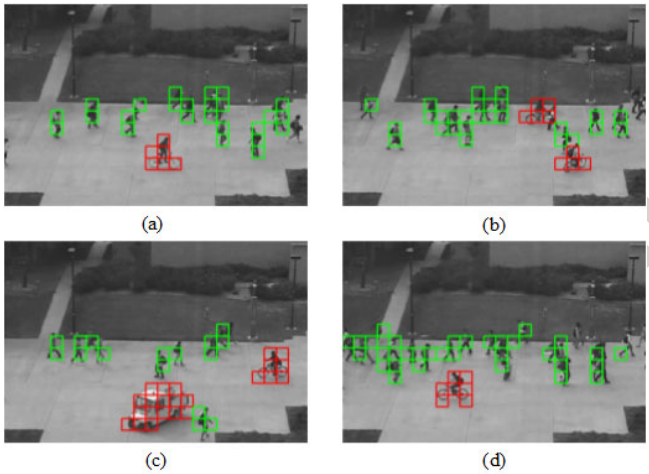

Fig. 6 Examples of anomaly location in Ped2. Green rectangles show the motion areas while red rectangles denote anomalies detected by our method.

the training frames. Resample operation is performed on the anomalous samples to make training class in balance. A comparison of the proposed method against recently published results in this dataset, is shown in Table 3 and Fig.7. As the global anomalies can be easily detected in the UMN dataset, many previous works achieve near perfect performance. Without exception, our method also achieves perfect performance and reaches 99.63% under the AUC criterion. The examples of detection results are shown in Fig.8. Our method show good performance in detecting the crowd run in high speed, however, when

Table 3 AUC performance in the UMN, Subway and U-turn dataset

Method	UMN	Subway		U-turn
		Entrance	Exit	
Chaotic invariants[44]	99.4%	-	-	-
SF[29]	94.9%	-	-	
Sparse[8]	99.6%	80.2%	83.3%	-
Local stat.aggr[35]	99.5%	88.4%	-	94.7%
H-MDT(CRF)[26]	99.5%	89.7%	90.8%	95.2%
HOS-HOG[19]	98.1%	-	-	**95.3**%
Our method	**99.63**%	**92.7**%	**91.9**%	95.2%

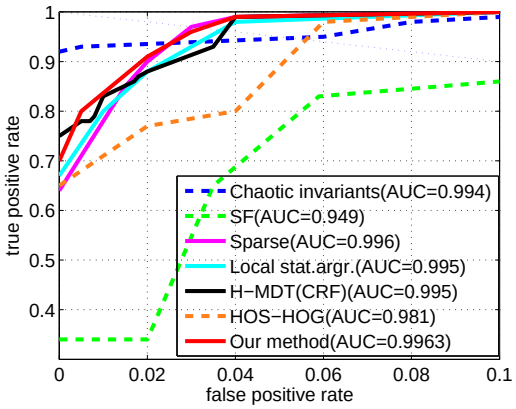

Fig. 7 ROC curves of frame level criterion in the UMN dataset.

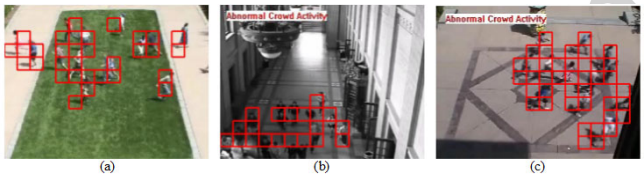

Fig. 8 Examples of anomaly location in UMN dataset. Red rectangles denote anomalies detected by our method. (a), (b), and (c) correspond to scenes lawn, indoor, and plaza, respectively.

only several pedestrians are running at the final stage of evacuation, a few false positives appear.

4.3 Subway dataset

The Subway dataset contains two video clips, recording passengers entering (1 h and 36 min, 144,249 frames) and exiting (43 min, 64,900 frames) a subway station. Normal behaviors and abnormal behaviors appear alternately in each clip. The abnormal events in this dataset include passengers go through the

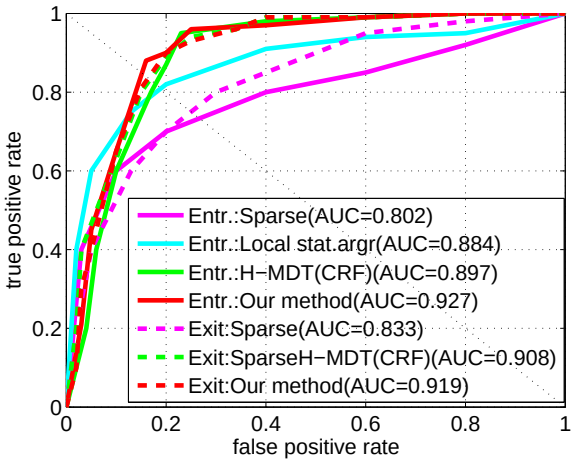

Fig. 9 ROC curves of frame level criterion in the Subway dataset.

wicket without payment and move in wrong direction (enter the exit or exit the entrance). The size of SVOIs for Subway dataset is set as $30 \times 30 \times 7$. For training, the first 50 and 25 min of video from entering sequence and exiting sequence are collected to train the spatial-temporal CNN model, respectively, while the resting are used for testing. We extract 932021 normal samples and 418298 anomalous samples from the training frames. The comparison results with previous literature are provided in Table 3 and Fig.9. The screenshots of detections in this dataset are presented in Fig.10. Most anomalies are correctly detected and localized, however, a few "anomalies" that without obvious motion are not successfully detected. Even so, our methods achieves the best result among all state-of-the-art methods in the AUC criterion, improving upon the best competitor (H-MDT (CRF) [19]) by 3.3% for Entrance and 1.2% for Exit.

4.4 U-turn dataset

The U-turn dataset depicts a sparse traffic scene at a road intersection, in which object-based techniques, for instance, tracking and analysis of object trajectories, can be applied. The dataset consists of 6117 frames of 360×240 pixels, however, it only provides frame level ground-truth. The typical abnormal events in this dataset are some vehicle making illegal U-turns. The size of SVOIs for U-turn dataset is set as $20 \times 20 \times 7$. We extract 45877 normal samples and 21409 anomalous samples from the training frames. The dataset is significant challenge to our model due to the limited number of abnormal frames and sparseness of the scenes. Even so, our method correctly distinguish normal events and abnormal behaviors, as in Fig.11. The video sequence is divided into two subsets for model training and validation. The comparisons

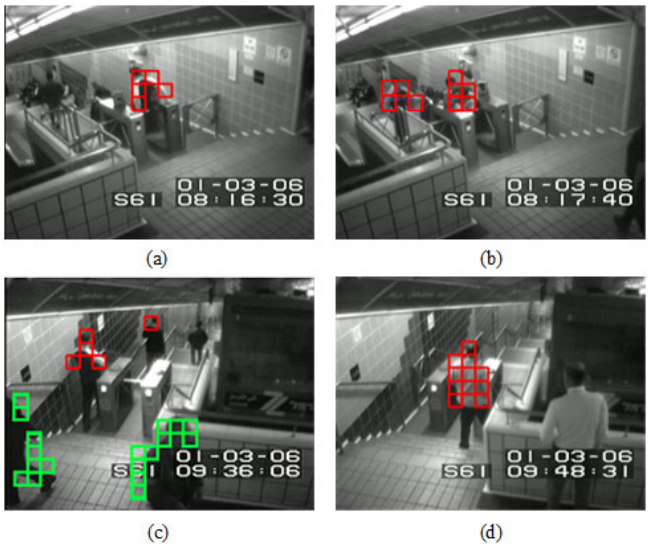

Fig. 10 Examples of anomaly location in Subway dataset. The abnormal behaviors, including moving in wrong direction, voiding payment, are successfully detected. Red rectangles denote abnormal regions while green rectangles are motion regions. (a)(b) corresponds the examples of the Entrance while (c)(d) are the screenshots of the Exit.

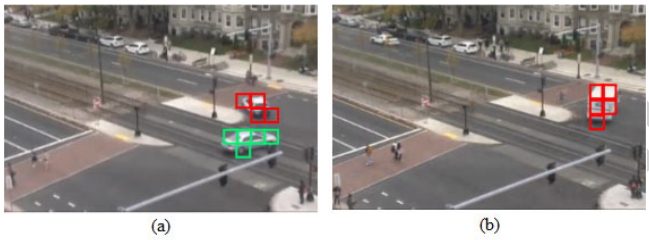

Fig. 11 Examples of anomaly location in U-turn dataset.The abnormal behaviors include illegal vehicle motion at the intersection. Red rectangles denote abnormal regions while green rectangles are motion regions.

of performance for the U-turn dataset are summarized in Table 3 and Fig.12. As shown in Table 3 and Fig.12, our method achieves a comparable result in AUC criterion, equal to 95.2%.

4.5 Computation efficiency

We implement the network training based on the efficient CNN tools: BVLC caffe [18]. All the experiments were on a standard workstation (2.8Gz Genuine CPU, 128G RAM and Ubuntu 64-bit OS). During the training stage, all of the parameters of net work are initialized randomly and trained by stochastic gradient descent based back propagation. Table 4 shows average computation

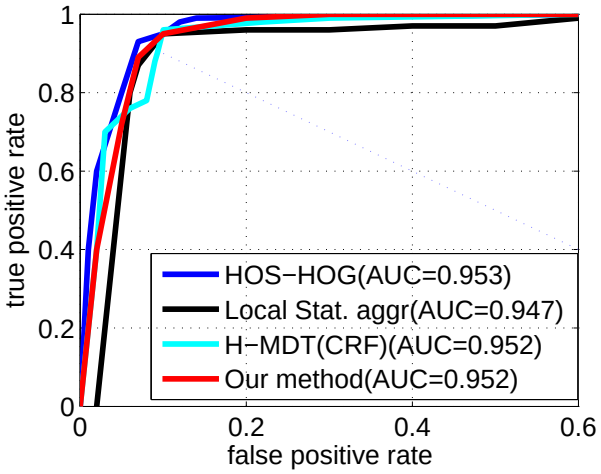

Fig. 12 ROC curves of frame level criterion in the U-turn dataset.

Table 4 Computation cost

Method	Ped1	Ped2
H-MDT(CRF)[26]	0.92 s	1.01 s
HOS-HOG[19]	0.86 s	1.13 s
Ours	0.37 s	0.39 s

cost of per frame in the UCSD dataset. As can be seen that our method achieves lower computation cost than other methods both on the Ped1 and Ped2. In the testing stage, the main computation-consuming task is computing optical flow data to localize dynamic regions and the classification task of SVOIs by the spatial-temopral CNN is quite efficient.

5 Conclusion

In this work, we develop a spatial-temporal CNN model for anomaly detection and localization in different scenes, recorded from static cameras. In our method, the spatial-temporal volumes that carry rich motion information are fed to train the spatial-temporal CNN model for anomaly detection. The spatial-temporal CNN model is designed to robustly construct features from both spatial and temporal dimensions by performing spatial-temporal convolutions, therefore, appearance feature as well as dynamic information encoded in continuous frames are captured. Experiments on 4 different kinds of benchmarks show that the performance of our method is better than state-of-the-art approaches, especially on the most challenging pixel level criterion. In the future work, we are interested in using spatial-temporal CNN model

to address spatial-temporal volumes with multi scales and variable temporal length.

References

1. Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D.: Robust real-time unusual event detection using multiple fixed-location monitors. Pattern Analysis and Machine Intelligence, IEEE Transactions on **30**(3), 555–560 (2008)
2. Ali, S., Shah, M.: A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, pp. 1–6. IEEE (2007)
3. Benezeth, Y., Jodoin, P.M., Saligrama, V., Rosenberger, C.: Abnormal events detection based on spatio-temporal co-occurences. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 2458–2465. IEEE (2009)
4. Bertini, M., Del Bimbo, A., Seidenari, L.: Multi-scale and real-time non-parametric approach for anomaly detection and localization. Computer Vision and Image Understanding **116**(3), 320–329 (2012)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research **3**, 993–1022 (2003)
6. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM computing surveys (CSUR) **41**(3), 15 (2009)
7. Cheng, K.W., Chen, Y.T., Fang, W.H.: Video anomaly detection and localization using hierarchical feature representation and gaussian process regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2909–2917 (2015)
8. Cong, Y., Yuan, J., Liu, J.: Sparse reconstruction cost for abnormal event detection. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 3449–3456. IEEE (2011)
9. Cong, Y., Yuan, J., Liu, J.: Abnormal event detection in crowded scenes using sparse representation. Pattern Recognition **46**(7), 1851–1864 (2013)
10. Cuntoor, N.P., Yegnanarayana, B., Chellappa, R.: Activity modeling using event probability sequences. Image Processing, IEEE Transactions on **17**(4), 594–607 (2008)
11. Gandhi, T., Trivedi, M.M.: Pedestrian protection systems: Issues, survey, and challenges. Intelligent Transportation Systems, IEEE Transactions on **8**(3), 413–430 (2007)
12. Garcia, C., Delakis, M.: Convolutional face finder: A neural architecture for fast and robust face detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on **26**(11), 1408–1423 (2004)
13. Goodfellow, I.J., Bulatov, Y., Ibarz, J., Arnoud, S., Shet, V.: Multi-digit number recognition from street view imagery using deep convolutional neural networks. arXiv preprint arXiv:1312.6082 (2013)
14. Hearst, M.A., Dumais, S.T., Osman, E., Platt, J., Scholkopf, B.: Support vector machines. Intelligent Systems and their Applications, IEEE **13**(4), 18–28 (1998)
15. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. Physical review E **51**(5), 4282 (1995)
16. Hospedales, T., Gong, S., Xiang, T.: A markov clustering topic model for mining behaviour in video. In: Computer Vision, 2009 IEEE 12th International Conference on, pp. 1165–1172. IEEE (2009)
17. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on **35**(1), 221–231 (2013)
18. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
19. Kaltsa, V., Briassouli, A., Kompatsiaris, I., Hadjileontiadis, L.J., Strintzis, M.G.: Swarm intelligence for detecting interesting events in crowded environments. Image Processing, IEEE Transactions on **24**(7), 2153–2166 (2015)

20. Kim, J., Grauman, K.: Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 2921–2928. IEEE (2009)
21. Kratz, L., Nishino, K.: Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 1446–1453. IEEE (2009)
22. Krausz, B., Bauckhage, C.: Analyzing pedestrian behavior in crowds for automatic detection of congestions. In: Computer vision workshops (ICCV workshops), 2011 IEEE international conference on, pp. 144–149. IEEE (2011)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012)
24. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
25. Li, N., Wu, X., Xu, D., Guo, H., Feng, W.: Spatio-temporal context analysis within video volumes for anomalous-event detection and localization. Neurocomputing **155**, 309–319 (2015)
26. Li, W., Mahadevan, V., Vasconcelos, N.: Anomaly detection and localization in crowded scenes. Pattern Analysis and Machine Intelligence, IEEE Transactions on **36**(1), 18–32 (2014)
27. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 1975–1981. IEEE (2010)
28. Maturana, D., Scherer, S.: 3d convolutional neural networks for landing zone detection from lidar. In: Robotics and Automation (ICRA), 2015 IEEE International Conference on, pp. 3471–3478. IEEE (2015)
29. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 935–942. IEEE (2009)
30. Mousavi, H., Mohammadi, S., Perina, A., Chellali, R., Murino, V.: Analyzing tracklets for the detection of abnormal crowd behavior. In: Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on, pp. 148–155. IEEE (2015)
31. Nam, Y., Hong, S.: Real-time abnormal situation detection based on particle advection in crowded scenes. Journal of Real-Time Image Processing pp. 1–14 (2014)
32. Raghavendra, R., Bue, A.D., Cristani, M., Murino, V.: Optimizing interaction force for global anomaly detection in crowded scenes. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, pp. 136–143. IEEE (2011)
33. Rodriguez, M., Laptev, I., Sivic, J., Audibert, J.Y.: Density-aware person detection and tracking in crowds. In: Computer Vision (ICCV), 2011 IEEE International Conference on, pp. 2423–2430. IEEE (2011)
34. Roshtkhari, M.J., Levine, M.D.: Online dominant and anomalous behavior detection in videos. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pp. 2611–2618. IEEE (2013)
35. Saligrama, V., Chen, Z.: Video anomaly detection based on local statistical aggregates. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 2112–2119. IEEE (2012)
36. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 (2013)
37. Shao, L., Ji, L., Liu, Y., Zhang, J.: Human action segmentation and recognition via motion and shape analysis. Pattern Recognition Letters **33**(4), 438–445 (2012)
38. Shen, W., Wang, X., Wang, Y., Bai, X., Zhang, Z.: Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3982–3991 (2015)
39. Sodemann, A., Ross, M.P., Borghetti, B.J., et al.: A review of anomaly detection in automated surveillance. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on **42**(6), 1257–1272 (2012)

40. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pp. 1701–1708. IEEE (2014)
41. Tu, P., Sebastian, T., Doretto, G., Krahnstoever, N., Rittscher, J., Yu, T.: Unified crowd segmentation. In: Computer Vision–ECCV 2008, pp. 691–704. Springer (2008)
42. Wang, K., Wang, X., Lin, L., Wang, M., Zuo, W.: 3d human activity recognition with reconfigurable convolutional neural networks. In: Proceedings of the ACM International Conference on Multimedia, pp. 97–106. ACM (2014)
43. Wang, L., Qiao, Y., Tang, X.: Latent hierarchical model of temporal structure for complex activity classification. Image Processing, IEEE Transactions on **23**(2), 810–822 (2014)
44. Wu, S., Moore, B.E., Shah, M.: Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 2054–2060. IEEE (2010)
45. Xiong, G., Wu, X., Chen, Y.l., Ou, Y.: Abnormal crowd behavior detection based on the energy model. In: Information and Automation (ICIA), 2011 IEEE International Conference on, pp. 495–500. IEEE (2011)
46. Xiong, G., Wu, X., Cheng, J., Chen, Y.L., Ou, Y., Liu, Y.: Crowd density estimation based on image potential energy model. In: Robotics and Biomimetics (ROBIO), 2011 IEEE International Conference on, pp. 538–543. IEEE (2011)
47. Zhou, B., Tang, X., Wang, X.: Measuring crowd collectiveness. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3049–3056 (2013)
48. Zhu, X., Jin, X., Zhang, X., Li, C., He, F., Wang, L.: Context-aware local abnormality detection in crowded scene. Science China Information Sciences **58**(5), 1–11 (2015)