Machine Vision and Applications

Exploration of STFD Descriptor for Online Crowd Abnormal Behavior Detection and Deep-Learning-based CNN Classification

--Manuscript Draft--

Manuscript Number:	
Full Title:	Exploration of STFD Descriptor for Online Crowd Abnormal Behavior Detection and Deep-Learning-based CNN Classification
Article Type:	S.I. : Human Abnormal Behavioural Analysis
Corresponding Author:	Yuanping Xu Chengdu University of Information Technology CHINA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Chengdu University of Information Technology
Corresponding Author's Secondary Institution:	
First Author:	Yuanping Xu
First Author Secondary Information:	
Order of Authors:	Yuanping Xu
	Li Lu
	Zhijie Xu
	Jia He
	Jiliu Zhou
	Chaolong Zhang
	Jing Wang
Order of Authors Secondary Information:	
Funding Information:	National Natural Science Foundation of China (61203172) — Prof Yuanping Xu Department of Science and Technology of Sichuan Province (2018YYJC0994) — Prof Yuanping Xu Department of Science and Technology of Sichuan Province (2017JY0011) — Prof Jia He
Abstract:	Exploration of STFD Descriptor for Online Crowd Abnormal Behavior Detection and Deep-Learning-based CNN Classification
Suggested Reviewers:	Dongbing Gu, PhD Professor, University of Essex dgu@essex.ac.uk Dongbing Gu is a professor in School of Computer Science and Electronic Engineering, University of Essex, UK. His current research interests include distributed control algorithms, distributed information fusion, cooperative control, model predictive control, and machine learning. He has published more than 180 papers in international conferences and journals. His research has been supported by Royal Society, EPSRC, EU FP7, British Council, and industries. He is a board member of International Journal of Model, Identification, and Control, Cognitive Computations, Intelligent Industrial Systems, and Frontiers Robotics and AI. He served as a member of organizing committee and programme committee for many international conferences. Prof. Gu is a senior member of IEEE.

	Hui Yu, PhD Reader in Visual Computing, University of Portsmouth hui.yu@port.ac.uk Dr Yu is a Reader in the School of Creative Technologies. He previously held an appointment with the University of Glasgow. He has won prizes for his study and research include Excellent Undergraduate Prize (provincial level), the Best PhD Thesis Prize, EPSRC DHPA Awards (PhD) and Vice Chancellor Travel Prize. Dr Yu is an Associate Editor of IEEE Transactions on Human-Machine Systems. He is a member of the Peer Review College, the Engineering and Physical Sciences Research Council (EPSRC), UK.
	Hongji Yang, PhD Professor, Bath Spa University h.yang@bathspa.ac.uk Hongji Yang currently works at the Centre for Creative Computing, Bath Spa University. Hongji does research in Software Engineering, Distributed Computing and Computing in Social science, Arts and Humanities. Their current project is 'SmileMouth Project: Medical Big Data Services for Dentistry. He is Deputy Director of Creative Computing Centre and Editor in Chief of International Journal of Creative Computing

Exploration of STFD Descriptor for Online Crowd Abnormal Behavior Detection and Deep-Learning-based CNN Classification

Yuanping Xu[1], Li Lu[1], Zhijie Xu[2,3], Jia He[2], Jiliu Zhou[2], Chaolong Zhang[1], Jing Wang[3]
[1]School of Software Engineering, Chengdu University of Information Technology, Chengdu, China
[2]School of Computer, Chengdu University of Information Technology, Chengdu, China
[3]School of Computing & Engineering, University of Huddersfield, Queensgate, Huddersfield, UK
ypxu@cuit.edu.cn

Abstract With the rapidly increasing demands from real-time and intelligent applications of Closed-Circuit Television (CCTV) in security industries, crowd abnormal behavior detection and classification has become a hot research topic in aspect of crowd monitoring and management in public areas. This research has investigated an automatic online crowd anomaly detection model by exploring a novel compound image descriptor from video streams and the training of a deep learning network based on these descriptor instances. In this paper, the work reported has focused on: 1) creating spatio-temporal cuboids in online (or near real-time) manner through extracting local feature tracks in temporal space and the foreground blocks (i.e., moving objects) based on Gaussian Mixture Model (GMM) in spatial space, such that the extracted foreground blocks can effectively remove the irrelevant backgrounds for reducing the computational costs in the subsequent processing stages; 2) integrating a rich image descriptor (named spatio-temporal feature descriptor – STFD), and its instances for registering the crowd attributes (namely, collectiveness, stability, conflict and density) in the generated spatio-temporal cuboids. STFD can not only reflects the dynamic variation of the target crowd over time using its local feature tracks, but also enhances the interactive information of neighborhood pedestrians in a crowd (e.g. the interaction force) through the K-nearest neighbor (K-NN) graph modelling; 3) inputting generated STFD descriptor instances into the devised convolutional neural network (CNN) to detect suspicious crowd behaviors in a supervised manner. The proposed model has been tested and evaluated on two benchmarking databases - UMN and PETS - as the primary experimental data sets. Results compared with the benchmarking ground truths have shown substantial improvements of the devised model in terms of accuracy and efficiency for online crowd abnormal behavior detection.

Keywords crowd abnormal behaviors, Gaussian Mixture Model, spatio-temporal feature descriptor, crowd attributes, convolutional neural network (CNN),

1 Introduction

Automated crowd abnormal behavior detection and classification has become one of the most popular research topics in video analysis and intelligent applications. In particular, the study is receiving heightened attention to detect potential dangerous situations in real-time (or near real-time) in public areas (e.g., football stadiums, railway stations, open air concerts and busy streets) under current geopolitical atmosphere and public security demands. It is widely perceived that potential security threats in public areas are embedded and encapsulated in the consisting individual and crowd behaviors subjecting to the surrounding environments and group entity interaction patterns. Hence, the study and abstract of crowd behaviors using rich and high-level descriptors inherited from actions and interactions among crowd entities over spatial and time domains yields promising direction of the proposed work.

The ultimate goal of crowd behavior analysis is to detect or recognize crowd anomaly events automatically and intelligently in video clips within various complex contexts. The vision-based approaches for crowd behavior analysis can be classified into two major categories, 1) object-based (or individual based) methods and 2) holistic methods [1]. The former infers both behaviors and their associated individual entities, which has been applied to analyze and monitor behaviors of crowds with the low to medium density. However, these object-based methods face considerable difficulties to recognize individual activities in crowds with high density since it is inherently difficult to acquire accurate individual information in a crowded scene infested with occlusions and shadowing. Moreover, it is very difficult to distinguish individuals in crowd scenes having similar appearances. Therefore, the relatively matured individual segmentation and tracking techniques are not feasible for crowd studies [2].

Compared with the object-based cases, the holistic approach treats the crowd as a single entity, and it focuses on the entire scene rather than analyzing the specific actions of each individual [3]. It is based on the assumption that individual objects in a crowded scene are often too small to be identified or of any major values for crowd monitoring purpose. Therefore, most of the holistic approaches had been designed to treat a crowd as an integral whole for behavior pattern analysis. For example, Mehran extracted optical flows calculated based on the social-force theory, and then K-means clustering has been applied on the flow area to obtain several clusters [4]. With a corpus of clusters, Latent Dirichlet Allocation (LDA) was then deployed to discover the topics in the normal crowd behaviors, and Expectation Maximization (EM) algorithm with the Bag of Words (BoW) model were later used to maximize the likelihood of a corpus [5]. Both the LDA training and EM approximation algorithms need a great deal of computation, which is challenging for the process to be deployed for online and automated monitoring [5]. Moreover, Wang proposed a method based on optical flow histograms to represent the global motion in the scene [6], and the generated histograms are used to detect panic events. In these holistic approaches, crowd dynamic models are often adopted to estimate the behavior patterns as a whole, such that local behaviors in unstructured scenes cannot be handled well.

Another taxonomy for classifying crowd behaviors

follows a more classical image processing route through studying the pixel-value-based crowd trajectories and motion features. Motion features in a video can be divided into global features and local features. The global features are extracted using background removal and target tracking methods. Local features are interest points of individual video frames that are combined to describe crowd information. Although the relevant processes often require high volume preprocessing, local features contribute to the forming of the so-called image descriptors that are the core technique for object recognition, e.g. Dalal applied the HOG (Histograms of Oriented Gradient) descriptor that is computed by the gradients of an image [7]. These methods have been successfully applied in individual action recognition. The optimization of image descriptors based on automatic partitions of crowd behavioral patterns and the corresponding feature extraction is a challenging task for detecting abnormal crowd behaviors in complicated scenes (e.g. high-density crowds, occlusion and shadowing, and low resolutions). Recently, Mousavi et al. proposed the HOT (Histogram of Oriented Tracklets) descriptor that merges orientation and magnitude into 2D histograms [8], and it is mainly used to recognize abnormal behaviors (e.g. panic and violence). But the HOT descriptor and other similar studies have discarded the interaction force among individuals in a crowd, and unfortunately, most crowd abnormal behaviors contain significant interactive forces among pedestrians. To engaging the interactive forces, Shao et al. proposed subgroup descriptors to quantify interactive information [9], and then each subgroup tracked by the Kanade-Lucas-Tomasi (KLT) feature point is represented as a K-NN graph, such that the complex interaction between pedestrians can be described by using the intra-group and inter-group descriptors. Although the group descriptors have shown promising results for crowd behavior analysis, they still have some weaknesses: 1) the local information (e.g. the attributes of collectiveness and stability) has to be discarded during group segmentation; 2) the computational output of group descriptors are strongly depended on the group segmentation algorithms, e.g. a group may has a large number of individuals having same velocity and motion orientation, and each big group has one and only one STFD value or instance (four factors) for the later deep learning model that is insufficient. In contrast, this study proposes a novel model that bypasses the troublesome group segmentation step for integrating the extraction and representation of the crowd motion information in one step, such that it preserves the integrity of local information and reduces the computational complexity caused by the group segmentation processes.

Compared with existing algorithms and models, the main improvements of the proposed integrated model in this study are summarized as below:

1) Inspirational analogy between an individual in a crowd and a feature point in the Spatio-Temporal (ST) space. Separately, in temporal space, feature points in foreground blocks that can be extracted by Gaussian Mixture Model (GMM) are tracked by KLT. In spatial space, neighborhood feature points are concatenated by K-NN graphs at every frame, so as to the dynamic variation of the target crowd can be reflected and represented based on the motion information generated by local feature tracks in defined spatio-temporal cuboids.
2) A novel image descriptor - spatio-temporal feature descriptor (STFD) is defined by calculating the attributes (i.e., collectiveness, stability, conflict and density) of feature points with respect to their corresponding cliques (this study defines that the K^{th} nearest neighbors of a feature point organizes its clique for calculating the corresponding four attributes), and the STFD has been devised in this study for extracting and representing the interactive information in the constructed ST cuboids.
3) The STFD and a CNN (Convolutional Neural Network) have been integrated to explore a deep learning model for classifying crowd abnormality types. The STFD instances drawn from a crowd video form a feature map (2D image) that holds all extracted information of global movements and local interactive information of a crowd. The generated feature maps are then inputted into the devised CNN model for detecting and classifying abnormal crowd behaviors.

The rest of this paper is organized as follows. Section 2 introduces the related studies on the KLT based feature point tracking, trajectory tracking, crowd feature extractions and CNN that are key techniques to define and learn image descriptors. Section 3 presents the framework of devised anomaly detection and classification model. Section 4 provides detailed discussions and explanations of the novel technique for generating the STFD descriptor based on ST cuboids. It also aids a deeper insight of feature vectors, feature maps and the CNN training and prediction mechanisms. Applications and experimental results on real-world video scenes are analyzed and evaluated in the Section5. Section 6 concludes the proposed model with merits and areas to improve for future study.

2 Related Studies

2.1 KLT Based Feature Point Tracking

KLT is a tracking algorithm based on the concept of optical flow [10]. KLT tracker is the most robust and real-time solution for local feature point tracking [11-13]. The KLT combines feature point extraction and tracking, and it only extracts and tracks the good feature points (i.e., good texture in fixed-size feature windows), such that the results of feature point tracks are much more reliable. Moreover, KLT was used in the researches [14,15] to recognize crowd behaviors and the derived results provide reliable detections in crowded regions. In our study, the pyramidal optical flow algorithm is used to avoid the loss of the local feature points due to their fast movements [10], and then KLT is used to select the good feature points on the first frame of the video clips. As KLT automatically detects new feature points to replace lost points in the subsequent frames, the feature points in each subsequent frame can be identified and refreshed dynamically according the feature points extracted from the first frame. Each feature point should be connected from its location on the first frame to the location in its current frame to form a trajectory, see detail explanations in Section 2.2.

2.2 Trajectory Tracking

In a crowd of high density, conventional tracking

algorithms are faced with significant challenges: 1) the size of a pedestrian is too small to be tracked in a crowd having a large number of pedestrians of close vicinity; 2) the number of pixels of a pedestrian would be decreased rapidly due to the occlusions caused by pedestrian interactions; 3) it is difficult to differentiate the different individuals when several individuals are completely overlapping due to constant cross-over among pedestrians; 4) there are objects having similar appearances and discontinuous trajectories since the target object often exits the field of view and re-appears later again in the scene. These problems often lead to the loss of target objects, and difficulty of trajectory tracking. To address these difficulties, alternative solutions (e.g., KLT algorithms) that tracks pixel based particles or local feature points of an image instead of individual pedestrians are explored. However, traditional KLT algorithms can only support a short period tracking (e.g. several frames) and it still struggled to track feature points (i.e., corner points) of pedestrians over a long period of time, especially in high density situations. The difficulty in obtaining complete trajectories can be alleviated by putting together a set of fragments of motion features (named tracklets) tracked within a short period of time continuously to form a longer trajectory. In order to balance the tracking time and the computational performance of traditional KLT operations, tracklets are usually extracted from dense feature points (corner points) using specific mechanisms to enforce the spatio-temporal coherence between tracklets [16].

2.3 Motion Feature Extraction

Some holistic attributes (e.g. crowd density and movement flow) can be extracted from the modelled trajectories in temporal and spatial space, such that these attributes can be used to quantify the crowd behaviors. For instance, Zhang et al. defined the social attribute-aware force model (SAFM) to extract crowd motion features, such as disorder, congestion, interaction force, etc [17]. Dahrendorf et al. indicated that social conflict was one of the central themes in social research [18], and Wheelan et al. stated that conflict can be caused by the competitions for resources [19]. However, there features (e.g. disorder and social conflict) representing interactive information among individuals are very difficult to be extracted from high density crowds. Mehran et al. explored a holistic approach to measure the interaction force according to desired and actual velocities [4]. Inspired by the early works, this research devised a robust image descriptor (named STFD) by calculating the crowd attributes (i.e., collectiveness, stability, conflict and group density) of feature points modelled in K-NN graphs for representing and analyzing complex crowd behaviors.

2.4 Convolutional Neural Network (CNN)

Recently, deep learning has gained outstanding achievements in machine visual recognition applications such as image classification, localization and detection [20,21]. Recent developments in CNN – a core model of deep learning, have greatly advanced the performance of these state-of-the-art visual recognition systems [21]. Contrast with traditional methods that have to extract features from input data in a rigid and almost mechanical manner, CNN based models can automatically find features from video image. One of successful applications relating to CNN is that Krizhevsky presented the AlexNet to classify images in ILSVRC2012 (the ImageNet Large-scale Visual Recognition Challenge) and achieved a winning performance with the test error rate at 15.3% [22]. AlexNet is considered as the first successful deep learning model. Later, in 2015, he presented a new CNN model, ResNet, that won the ILSVRC 2015 with an incredible error rate at 3.6% [23]. Inspired by the CNN for image recognition, Shao et al. started to develop a multi-task deep model to jointly learn and combine appearance and motion features for crowded scene understanding on 2015 [24]. It provided a fundamental idea for crowd feature extractions and a feasible deep learning model for jointly learning the extracted features. However, there are still a lot of works should be done in Shao's research to develop practical systems for online crowd abnormal behavior detection and classification based on camera clusters.

3 Theoretical Approaches of the Integrated Anomaly Detection and Classification Model

The overall workflow of the proposed integrated model for online detection and classification of crowd anomaly events, especially for the extraction and computation of STFD is elaborated in this section (refer to Fig.1).

1) **ST segmentation:** in defined ST cuboids, the motion information of a crowd is obtained through extracting and creating local feature tracks in the temporal space and connecting K-NN graphs of local features (corner points or feature points) in the spatial space. In this study, the devised model tracks corner points of the target image (i.e., KLT tracks the points whose brightness have significant changes in an image) instead of corner points of a pedestrian.
2) **Feature Extraction:** A novel image descriptor - spatio-temporal feature descriptor (STFD) is defined by calculating the attributes of feature points with respect to their corresponding cliques, such that it can avoid the complex group segmentation steps. The STFD has been devised in this study for extracting and encapsulating the interactive information in the constructed ST cuboids.
3) **CNN-based Recognition:** The calculated STFD instances from a frame drawn from video streams can form a feature map consisted of certain number of feature vectors that holds all extracted information of global movements and local interactive information of a crowd. Taking STFD instances as the feature vectors on ST cuboids, each feature vector has been labeled with 10 different labels to be served as inputs for the devised CNN model for classification.

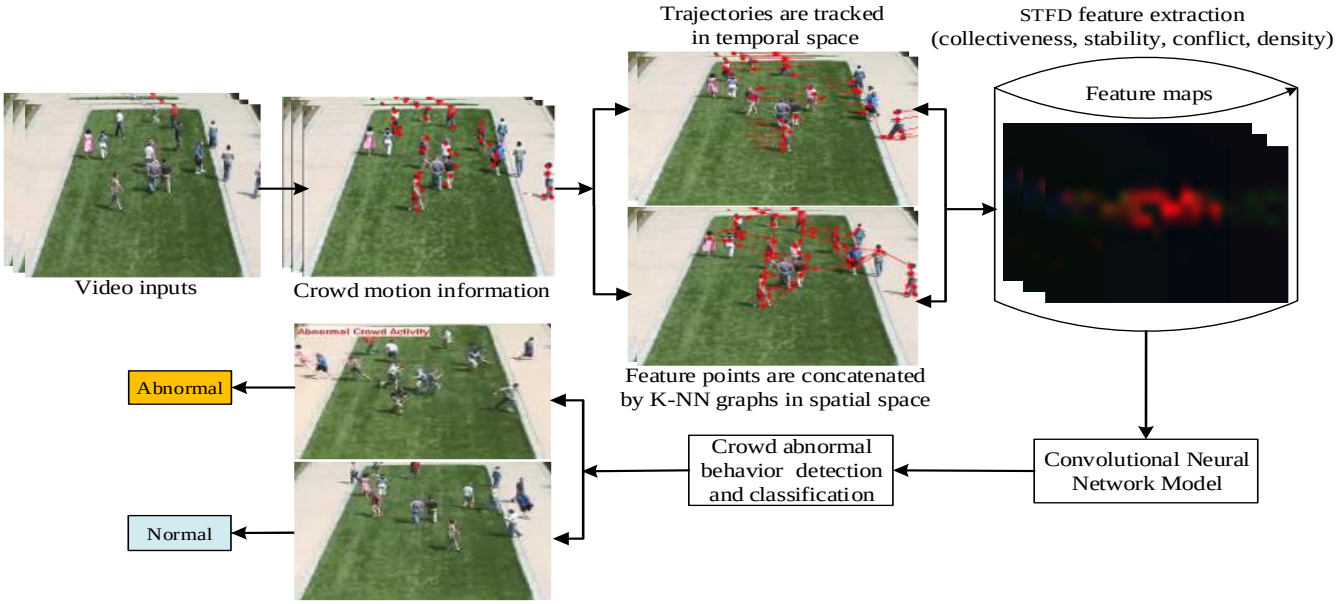


Fig. 1 The general framework of crowd anomaly detection and classification

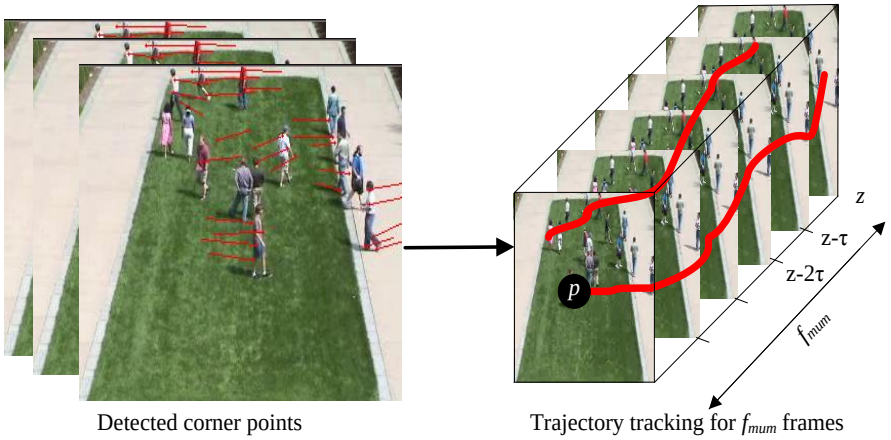


Fig. 2 The trajectories from the frame z-f_{num} to z

4 Realization and Operative Flow

4.1 Hybrid Spatio-temporal Segmentation

(1) Trajectory tracking in the temporal space

As shown in Fig.2, the m feature points (corner points) were tracked to obtain m trajectories in the period of f_{num} (f_{num}=25) frames by using KLT, and the generated trajectories that remain static or have not been updated for a long period time will be filtered out to reduce the computational cost. Moreover, the filtering step is key to alleviating the negative impact of irrelevant feature points drawn from a crowd. The trajectories (T_r^z) of the current frame z are defined as the following:

$$\begin{cases} Tr^z = \left\{T_1^z, T_2^z, ..., T_m^z \mid T_j^z = \left\{x_j(z - F\tau), ..., x_j(z)\right\}\right\} \\ \tau = f_{num} / F \end{cases} \quad (1)$$

where f_{num} is the frame rate of the video sequence (i.e., the number of frames per second), F is the number of segments (e.g. 5 in this study) in ST cuboid, and τ is the temporal interval for every F on a trajectory T_j^z, and x_j^z value is the coordinate of the feature point on frame z. The fragmentation of ST trajectories has two significant advantages: 1) it can save substantial computational time than the conventional method that relying on computations performed on every two adjacent frames; 2) the trajectory segments still contain the complete motion information (e.g. location, velocity, orientation) so as to ensure the accuracy of the following STFD calculation.

(2) Foreground segmentation in the spatial space

Gaussian Mixture Model (GMM) is one of the most powerful models for segmenting moving objects in a video sequence. In order to obtain spatial information of the target crowd, GMM is applied to get foreground blocks that indicates prominent moving regions in a scene, and the extracted foreground blocks can reduce the computational time in the subsequent processes of feature tracking. To facilitate modelling, we made an analogy between a feature point and a pedestrian. The feature points that are extracted according to corner points of the target image in foreground blocks are connected by K-NN graphs. In the z frame, this model defines K nearest neighbors (3 to 5 in this study) of a feature point p as a clique C_p^z. ($C_p^z = N_p^z, \ldots, N_{p+K}^z$). Thus, motion information in the z frame can be represented by the movement of C_p^z.

At this stage, the motion information of each local feature point is represented by trajectories in the temporal space and connected by K-NN graphs in the spatial space, such that crowd interactions can be quantified by the crowd attributes extracted from ST cuboids in the form of STFD Descriptor instances.

4.2 The Novel STFD Descriptor for Crowd Behaviors

At this stage, the generated image descriptor (STFD) can be quantified by using the four selected crowd attributes – collectiveness, stability, conflict and density and then a set of STFD instances can be calculated by using equation (5) – (8). More importantly, the STFD instances are calculated on basis of each individual feature point and the corresponding clique rather than the filtering techniques based on the complicated group segmentation, such that the computational complexity of quantification of crowd motion information can be greatly simplified. These instances are comprehensive ST features, serving as inputs for training and testing on the devised CNN structure. The selected attributes can provide quantified expressions of the interactive motion information of pedestrians in the flow, see Table 1.

Table 1 the crowd attributes for representing the crowd behaviors

Attributes	Descriptions	Equations
Collectiveness	*coll*	equation (5)
Stability	*stab*	equation (6)
Conflict	*conf*	equation (7)
Density	*density*	equation (8)

(1) Flow direction variations:

In the proposed model, the motion vectors are generated from a 2D trajectory on a set of successive frames (from *z*-τ to *z*). The motion vector $\overrightarrow{p^z}$ of a feature point *p* on frame *z* is defined as:

$$\overrightarrow{p^z} = \left\{ \overrightarrow{x_j^{(z-\tau)} x_j^{(z)}} \right\}_{1 \le j \le n} \tag{2}$$

where *n* is the number of frames in the time interval τ.

The flow direction variation is calculated by averaging the angular differences (*angle*) across all segments of the target trajectory on the *z* frame, which is defined as the following:

$$angle = \frac{1}{F} \sum_{f=0}^{F-2} d_\theta \left(\overrightarrow{p^{z-f\tau}}, \overrightarrow{p^{z-(f+1)\tau}} \right) \tag{3}$$

where $\overrightarrow{p^{z-f\tau}}$ is the motion vector of *p* in the *z*-$f\tau$ frame, and d_θ is defined as the angular difference of two vectors $\vec{a}$ and $\vec{b}$:

$$\begin{cases} \Delta\theta = \arccos\left(\dfrac{\vec{a} \cdot \vec{b}}{|\vec{a}| \times |\vec{b}|} \right) \\ d_\theta\left(\vec{a}, \vec{b}\right) = \begin{cases} \Delta\theta, if\ (\Delta\theta < \pi) \\ 2\pi - \Delta\theta, \text{otherwise} \end{cases} \end{cases} \tag{4}$$

(2) Collectiveness attribute

The collectiveness attribute indicates the degree of individuals acting as a clique in the collective motion [9]. In the flow direction, the motion vector of a feature point *p* with respect to its K-NN neighbor points can be obtained through calculating the motion direction of its clique. Thus, the collectiveness at *p* can be defined as the following:

$$\begin{cases} coll(p) = \sum_{i \in C_p^z} h\left(\overrightarrow{p^{z-\tau} p^z}, \overrightarrow{i^{z-\tau} i^z} \right) \\ h\left(\overrightarrow{p^z}, \overrightarrow{i^z} \right) = \begin{cases} d_\theta, if\ (d_\theta < \pi/2) \\ 0, \text{otherwise} \end{cases} \end{cases} \tag{5}$$

where C_p^z ($C_p^z = N_{p+1}^z, \ldots, N_{p+K}^z$) denotes the feature points in the clique of *p*, and $\overrightarrow{p^z}$ is the motion vector of *p* on the *z* frame.

(3) Stability attribute

The stability is the number of the invariant neighbors of each feature point in a clique. Thus, the stability definition (*stab*) of a feature point (*p*) can be given as the following:

$$stab(p) = \sum_{p \in C_p^z} \left(K - | N_p^{z-f_{num}} \setminus N_p^z | \right) \tag{6}$$

where $|N_p^1 \setminus N_p^\tau| = |\{p : p \in N_p^{z\text{-}fnum} \text{ and } p \notin N_p^z\}|$, $N_p^{z\text{-}fnum}$ is the invariant neighbors of *p* in the *z*-f_{num} frame (i.e., an entire spatio-temporal cuboid).

(4) Conflict attribute

The conflict attribute characterizes interaction between two feature points when they are approaching to each other. In this study, the conflict degree of a feature point *p* is calculated by the angular difference and distance between *p* and every other feature point in the corresponding clique. The conflict (*conf*) of a feature points (*p*) in its clique is defined as:

$$conf(p) = \sum_{i \in C_p^z} \frac{d_\theta \left(\overrightarrow{p^{z-\tau} p^z}, \overrightarrow{i^{z-\tau} i^z} \right)}{\left\| p^z i^z \right\|_2} \tag{7}$$

where p^z,i^z denotes the distance between *p* and *i* in the same clique on the *z* frame.

(5) Density attribute

The density attribute (*den*) is the spatial distribution of feature points in the foreground blocks, which is the measure of how close the local feature points are. The local density of each feature point *p* can be considered as a kernel density estimation based on the positions of its neighborhood set. Thus, the corresponding density map of *p* is defined as the following:

$$den(p) = \frac{1}{\sqrt{2\pi}\sigma} \sum_{i \in C_p^z} \exp^{-\frac{\left\| p^z i^z \right\|_2}{2\sigma^2}} \tag{8}$$

where *p* and *i* represent a pair of the neighbor feature points in a clique, and σ is the bandwidth of 2D Gaussian kernel.

The proposed model integrates the collectiveness, stability, conflict and density into a new spatial and temporal feature descriptor (STFD), such that it can hold not only the interaction behaviors of a clique but also it presents the overall spatial distribution of feature points. Thus, this comprehensive descriptor can describe the complete global motion information of pedestrians, and in the meantime, preserve both local and global features while ensuring the overall computational performance. Thereby an instance of the STFD descriptor (*stfd*) of *p* becomes a ST feature vector:

$$stfd(p) = \frac{1}{f_{num}} \begin{pmatrix} coll(p), stab(p), \\ conf(p), den(p) \end{pmatrix} \tag{9}$$

4.3 CNN-based Detection and Classification

The structure of the devised CNN model is illustrated in Fig.3, which has an upper (appearance) and a lower (motion) level. The upper and lower levels are concatenated into a network structure that contains a data layer, five convolution layers (Conv), three pooling layers (Pool), two normal layers (Norm) and two full connections layers (FC).

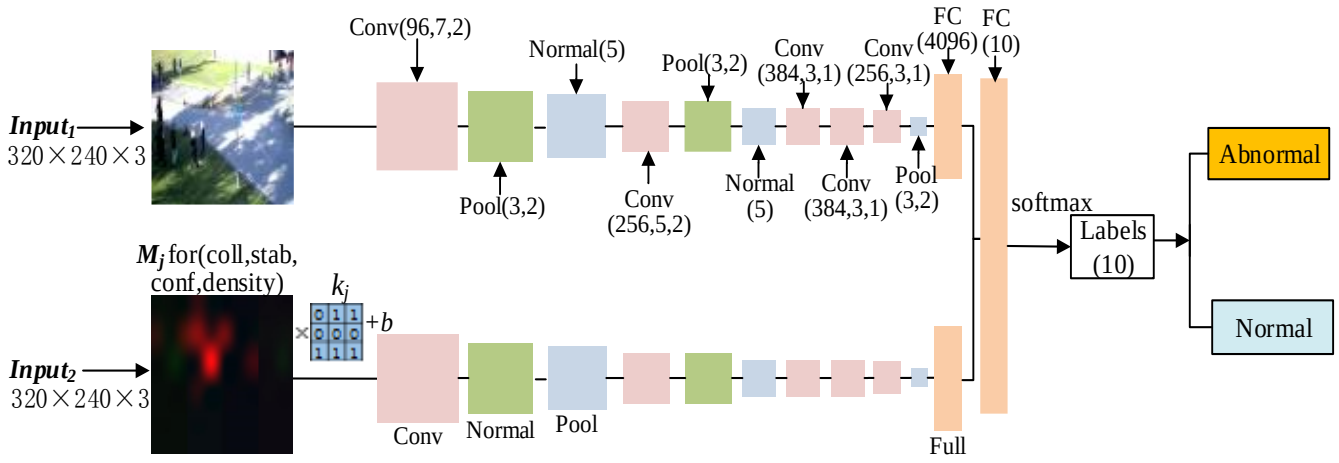


Fig. 3 The structure of the devised CNN

Conv (*H*, *L*, *S*) defines convolutional layers with output *H*, kernel size *L* and stride size *S*, respectively. The parameters of upper or lower network level are Conv(96,7,2)->ReLU->Pool(3,2)->Norm(5)->Conv(256,5,2)->ReLU->Pool(3,2)->Norm(5)->Conv(384,3,1)->ReLU->Conv(384,3,1)->ReLU->Conv(256,3,1)->ReLU->Pool(3,2)-> FC(4096) ->FC(10). So the instances of STFD descriptor of feature points are encapsulated into feature vectors which further form a feature map for CNN inputs by using equation (9). Customized number of instances of STFD become training samples labeled with 10 labels (e.g. indoor, runner and panic, etc.). The labeled training data sets are inputs of the devised CNN model to determine whether the current video frame is abnormal or not in the online detection phase.

1. During the training phase:
   - ✓ ***INPUT₁***: original video images (I_{rgb}) from frame 1 to 13064 serve as input data sets for ***INPUT₁***, which has been separated into a training data set (I_{tain}) (about 70% of image data for training), a testing data set (I_{test}) (about 20% of image data for testing) and a validating data set ($I_{validate}$) (about 10% of image data for validating).
   - ✓ ***INPUT₂***: The instances of STFD descriptor of I_{tain} from frame 1 to 13064 have been encapsulated into feature maps (each frame has a feature map) that further become a training data set (D_{tain}) and a validating data set ($D_{validate}$).
   - ✓ Every data item (i.e., a feature map) in training data set (I_{tain}, D_{tain}) and validating data set ($I_{validate}$, $D_{validate}$) have been marked with 10 labels, including where (indoor, Lawn, Plaza, park), who (pedestrian, runner) and action (walk, run, panic， gather).
   - ✓ Training data set (I_{tain}, D_{tain}) and validating set ($I_{validate}$, $D_{validate}$) have been unified into the size of 320×240 and persisted into files with the format of Lightning Memory-Mapped Database (LMDB).
   - ✓ The labeled video images (I_{tain}, $I_{validate}$) and training data (D_{tain}, $D_{validate}$) serve as input training maps (***INPUT₁*** and ***INPUT₂***) for this CNN model:
     1) **Data layer:** The M_{Itrain} can be obtained through applying every data item of I_{tain} to minus the mean value of I_{tain} ($mean_{Itain}$) on the upper level, and the similar operation has also been applied on D_{tain} to get M_{Dtrain} on the lower level. The results of M_{Itrain} and M_{Dtrain} are inputted to the subsequent upper and lower level of the devised CNN model respectively, see equation (10):

$$M_{Itrain} = I_{train} - mean_{Itrain}$$
$$M_{Dtrain} = D_{train} - mean_{Dtrain} \quad (10)$$

     2) **Convolution layer:** feature maps of the upper level neural network use the convolution kernels to extract various image characteristics (color, texture, contour et.al) on both global and local regions of the target video clip. This layer multiplies the l^{th} map of the current layer by the weight (k_{ij}^{l}) of a convolution kernel and the output map (M_j) of i^{th} (x_i^{l-1}) of the previous layer before adding the corresponding bias ($b_j l$). Then a Rectified Linear Unit (RELU) function (f) is applied to get the feature map (x_j^l) of l^{th} through the nonlinear mapping, see equation (11):

$$x_j^l = f\left(\sum_{i \in M_j} x_i^{l-1} \times k_{ij}^l + b_j^l\right) \quad (11)$$

     3) **Pooling layer:** The convoluted features from convolution layer are inputted into the pooling layer by using max-pooling to reduce their dimensions.
     4) **Full layer:** All sampling features from pooling layer are inputted into the full layer where different crowd abnormal behaviors (e.g. crowd panics and gatherings) will be classified by using the in-built softmax classifier. The loss function of entity CNN can be defined as the following:

$$E = -\frac{1}{N}\sum_{n=1}^{N} lab_n \log o_n + (1-t_n)\log(1-o_n) \quad (12)$$

     where N denotes the number of classification, lab_n denotes labels of classification, and o_n is the predicted values of the probability.

2. During the testing phase:
   - ✓ ***INPUT₁***: 20% of original video images (I_{rgb}) from the frame 1 to 13064 has become a testing data set (I_{test})
   - ✓ ***INPUT₂***: The instances of STFD descriptor of I_{test} from frame 1 to 13064 have been encapsulated into feature maps that further become a testing data set (D_{test}).
   - ✓ Testing data set (I_{test}, D_{test}) has been unified into the size of 320×240 and persisted into LMDB files.
   - ✓ The video images (I_{test}) and testing data (D_{test}) serve as inputs (***INPUT₁*** and ***INPUT₂***) for classification.
   - ✓ The optimum weights and mean files ($mean_{Itain}$, $mean_{Dtain}$) can be obtained after the devised CNN model

has been well trained during the training phase, and then it can classify the testing data set (I_{test}, D_{test}) into the normal or abnormal frames. For example, the classification results of having "runner", "run" or "panic" labels are classified as an abnormal frame.

5 Applications and Experimental Results

The UMN and PETS S3 databases had been selected to test and evaluate the devised integrated model. The UMN was created by the University of Minnesota [25], and the statistical information of UMN has been listed in Table 2. It consists of 11 video clips of different scenes: lawn (video 1-2), indoor (video 3-8) and plaza (video 9-11) respectively, and these clips contain 7739 video frames with frame rate of 25 frames per second holding both normal and abnormal crowd behaviors. These abnormal crowd behaviors depict human escaping from dangerous zone or running away with panics. PETS S3 database has four video clips for each crowd activity on park scenes with a resolution of 768×576, and these clips have 5325 video frames holding both normal and abnormal crowd behaviors, e.g., congregation and walking [26].

Table 2 the statistical information of UMN

Scene	Video name	The number of total frames	The start frame of an abnormal behavior
Lawn (1453)	video1	625	484
	video2	828	665
Indoor (4144)	video3	549	303
	video4	685	563
	vdeo5	769	492
	video6	579	450
	video7	895	734
	vdeo8	667	454
Plaza (2142)	vdeo9	658	551
	vdeo10	677	570
	vdeo11	807	717

The experiments were implemented by using Visual Studio 2013 and MATLAB software running on pure CPU (i7, 4 threads). In these experiments, human walking or loitering are defined as normal events while crowd running or gathering on sidewalks are considered as abnormalities.

The accuracy and recall ratio are used in these experiments to evaluate the efficiency and validity of the integrated model. In Equation(13):

- ✓ True Positive (*TP*) is an abnormal sample that is correctly classified as an abnormal one by the CNN;
- ✓ True Negative (*TN*) is a normal sample that is correctly classified as a normal one by the CNN;
- ✓ False Positive (*FP*) is a normal sample that is improperly classified as an abnormal one;
- ✓ False Negative (*FN*) is an abnormal sample that is improperly classified as normal one.
- ✓ Precision is the proportion of *TP* in the abnormal samples which are classified.
- ✓ Recall is the proportion of *TP* in real abnormal samples.

$$precision = \frac{FP}{TP + FP}$$
$$\text{Re}call = \frac{TP}{TP + FN} \quad (13)$$
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

1) Case 1: The validity test of the crowd density factor

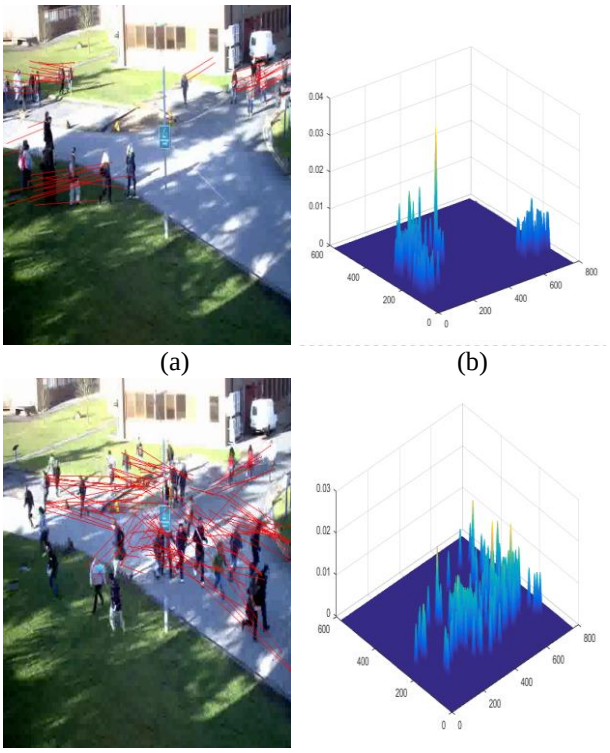


Fig. 4(a) motion vectors of the 40th frame, (b) the density map of the 40th frame, (c) motion vectors of the 350th frame, (d) the density map of the 350th frame

The first experiment of this study has focused on testing the validity of the integrated crowd characteristics indexed in STFD. Fig.4(a) shows the video image from the 40th frame, and in which the pedestrians are moving towards the same location; Fig.4(b) shows the corresponding density map of Fig.4(a), and the area of dense regions is augmenting over time as the feature points are increasing in the video image from 1st to 40th frames (e.g. Fig.4(a) reaches 70 feature points); Fig.4(c) shows the video image from the 350th frame, and in which the pedestrians are suddenly dispersed to different directions; Fig.4(d) shows the corresponding density map of Fig.4(c). In Fig.4(d), the motion area is decreasing, and it has 140 feature points on this frame. Table 3 shows the comparison of the computational time and the classification accuracy between STFD and the three crowd attribute integration (i.e., collectiveness, stability, conflict). With data sets from PETS S3, STFD has obtained 85% classification accuracy rate, whereas only 80.25% is achieved by using only three crowd attributes (i.e., collectiveness, stability, conflict), and there is no major difference between the computational times of STFD and the three attribute approach when dealing with the 5325 total frames of data sets in PETS S3 database. Thus, the proposed STFD improved the classification accuracy while ensuring that the computational performance when addressing the crowd density factor.

Table 3 experimental results of classification accuracy and computational time of crowd abnormal behavior detection by using STFD and the other three crowd attributes when dealing with data sets in PETS S3 database

Descriptors	Accuracy (%)	Time consuming (s)
STFD	85	252s

STFD without density	80.25	251s

2) Case 2: The comparison of different feature descriptor combinations

To assist the definition of the optimum weights (W_{ij}) for different image descriptor factors (collectiveness-*coll*, conflict-*conf*, stability-*stab and* density-*den*), this study elaborately designed an experiment to test the accuracy for all three-factor combinations from the four descriptor factors in STFD by using data sets selected from both the two databases (UMN, PETS S3). In Fig.5, the R_{stab} descriptor is defined as the feature vectors that integrate all other three descriptor factors except *stab*, which tests the classification accuracy for the combination of *coll*, *conf* and *den* by using the same process, and similar tests had been carried out with R_{coll}, R_{conf} and R_{den}. The lowest classification accuracy with UMN data sets is R_{stab} (see Fig.5), which indicates the *stab* is highly involved in the classification process of the devised model, and it also proves that *stab* is a key descriptor for detecting panic events, so the W_{ij} of *stab* should be setup with relatively higher value in the devised CNN model. The lowest classification accuracy with PETS S3 data sets is R_{conf}, which justifies that the *conf* is highly involved in the classification process of the devised model, and it also proves that *conf* is a key factor in crowd gathering event detection. Thus, based on this experiment, we can observe the most relevant descriptor factor for different crowd abnormal behaviors with R_{coll}, R_{conf}, R_{stab} and R_{den} respectively. Moreover, according to the experimental results in Fig.5, the classification accuracy of STFD is the highest with both UMN and PETS S3 data sets, and it proves the validity and efficiency of the proposed STFD.

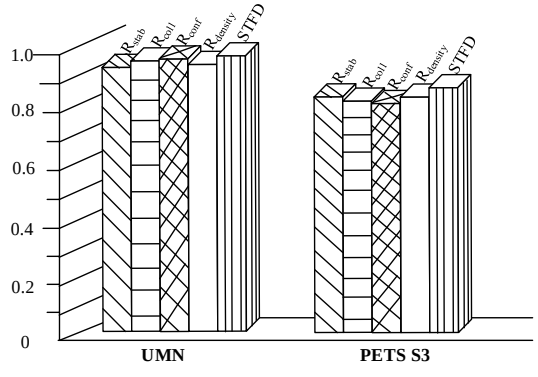


Fig. 5 experimental results for classification accuracy of crowd abnormal behaviors by using STFD and the other three-factor combinations from the four factors with both UMN (averaged the results for three scenes of UMN data sets) and PETS S3 databases.

3) Case 3: Benchmarking of the crowd abnormal behavior detection model

Three scenes (i.e., lawn, indoor and plaza) from the UMN database were selected to conduct the experiments. Table 4 presents the experimental results of the proposed model with the selected scenes that serving as input samples (i.e., normal and abnormal data sets). In Table 4, the area under the ROC (Receiver Operating Characteristic Curve, ROC) curve (AUC - the classification recognition rate) of the three scenes is 98.61, 98.64 and 97.5 respectively, and the larger area of AUC, the better classification recognition rate, and the error rate of recognition for all the three scenes are lower by using this integrated model.

Table 4 Experimental results for the integrated model

Scene	Actual /Recognized Normal frames	Actual /Recognized Abnormal frames	Total Accuracy	AUC
lawn	1298/1290	155/153	99.31	98.61
indoor	3393/3385	751/750	99.78	98.64
plaza	1942/1940	200/188	99.34	97.5

This experimental case is to test the AUC (average value of the three scenes) and benchmarking the computational performance of the devised integrated model against the classic optical flow algorithm, social force model (SFM), and the histogram of optical flow orientation (HOFO) method with testing data sets from lawn, indoor and plaza scenes. Table 5 shows the comparison of AUC and computational performance (i.e., the total computational time for processing 7739 frames of the three scenes in UMN) between the proposed model of this research and the other three traditional methods ([4], [9], [6]). It can be found that this model has better AUC than other classic methods while ensuring the high computational performance. Moreover, the experimental results also prove that STFD are more effective than the group-level descriptors in [9]

Table 5 The accuracy and computational performance comparison among different methods.

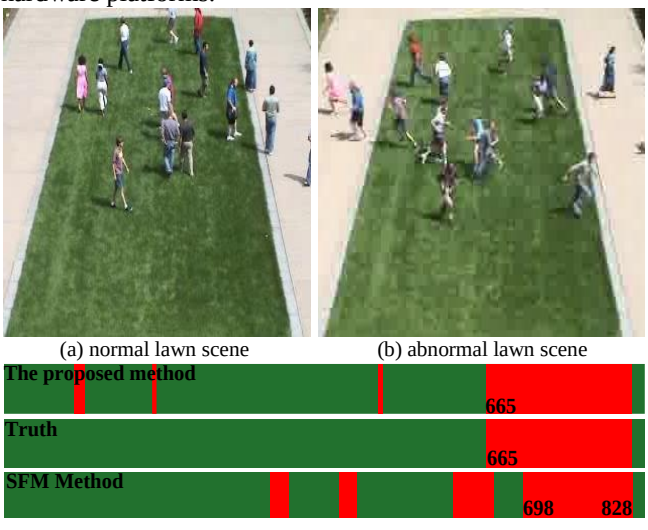
Method	AUC (%)	Time consuming (s)
Optical Flow [4]	84	429
Social Force [4]	94.9	455
Goup descriptors [9]	85.5	483
HOFO [6].	95.6	456
The proposed model	98.25	389

Fig.6, Fig.7 and Fig.8 provide frame-level quantitative representations of the start frames of the detected crowd panic behavior, where the green and red bar represents the normal frames and abnormal frames respectively. The "Truth" bar indicates the actual start and end frames (red blocks) that the crowd panic event had been happening [4]. According to Fig.6, the start frame (665) for the proposed method is the same as the "Truth" and much quicker than the SFM in the lawn scene. In general, the AUC of proposed model is closer to the truth and better than the classic STM methods.

6 Conclusions

The proposed integrated model for crowd abnormal behavior detection and deep-learning-based classification starts with the construction of ST cuboids in online (or near real-time) manner through extracting local feature tracks in the temporal space and the foreground blocks (i.e., moving objects) based on Gaussian Mixture Model (GMM) in the spatial space. The extracted foreground blocks can effectively remove the irrelevant backgrounds for reducing the computational costs in the subsequent processing stages. Based on the created ST cuboids, this study proposed a novel integrated image descriptor (STFD) for supporting online crowd abnormal behavior detection in CCTV video clips in an intelligent manner. The STFD can not only

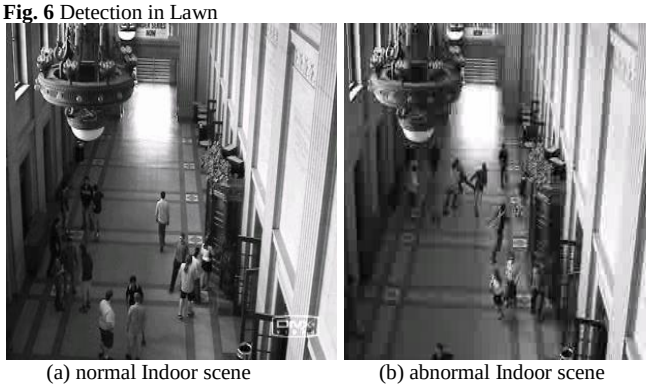
consider image pixels as local feature points to preserve interaction behaviors among pedestrians but also characterizes the spatial distribution of pedestrians in the scene. A new descriptor factor – crowd density has been integrated into STFD, and this integration can significantly improve the classification accuracy without enhancing the computational cost (see experimental Case 1). During the process of generating STFD instances, the four factors of STFD are calculated by using the clique concept, so as to the computational complexity has been greatly reduced and the interactive information of local feature points can be kept intact. By comparison of the different combinations of the four feature descriptor factors with data sets from both UMN and PEST databases (see experimental Case 2), the optimum weights for different feature descriptor factors on the convolution layer of CNN had also been justified. Based on the experimental Case 3 of comparisons of AUC and computational performance for the devised model, the experimental results show that this integrated model has good performance in terms of both classification accuracy and computational cost, and it can be readily transferred to the actual scenes of real world application conditions. There are two future extensions of this study: 1) because crowd abnormal events occur at some local regions rather than entire frame, STFD can be extracted from local blocks that can be defined through the grid partition of an image, such that the early crowd abnormal event detection can be achieved since the crowd abnormal behaviors are often started at some local regions rather than the entire area of a frame; 2) another potential improvement would be the optimization of the computational performance by using cluster-based HPCs and/or heterogeneous multi-GPU hardware platforms.



(a) normal lawn scene (b) abnormal lawn scene

(c) Lawn scene results

Fig. 6 Detection in Lawn

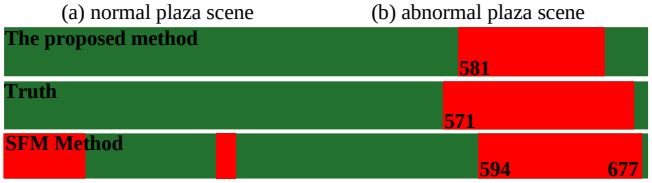


(a) normal Indoor scene (b) abnormal Indoor scene

(c) Indoor scene results

Fig. 7 Detection in Indoor

(a) normal plaza scene (b) abnormal plaza scene



(c) Plaza scene results

Fig. 8 Detection in Plaza

7 Acknowledgments

This work is supported by the NSFC (61203172), the STD of Sichuan (2018YYJC0994, 2017JY0011 and 2014GZ0007), and Shenzhen STPP (GJHZ20160301164521358).

References

1. Li, T., Chang, H., Wang, M., Ni, B., Hong, R., Yan, S.: Crowded Scene Analysis: A Survey. IEEE Transactions on Circuits & Systems for Video Technology 25(3), 367-386 (2015).
2. Loy, C.C., Xiang, T., Gong, S.: Detecting and discriminating behavioural anomalies. Pattern Recognition 44(1), 117-132 (2011).
3. Jacques Junior, J.C.S., Raupp Musse, S., Jung, C.R.: Crowd Analysis Using Computer Vision Techniques. Signal Processing Magazine IEEE 27(5), 66-77 (2010).
4. 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 2009, pp. 935-942
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993-1022 (2003).
6. Wang, T., Snoussi, H.: Histograms of Optical Flow Orientation for Visual Abnormal Events Detection. In: IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance 2012, pp. 13-18
7. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on 2005, pp. 886-893
8. Mousavi, H., Galoogahi, H.K., Perina, A., Murino, V.: Detecting Abnormal Behavioral Patterns in Crowd Scenarios. (2016).
9. Shao, J., Chen, C.L., Wang, X.: Learning Scene-Independent Group Descriptors for Crowd Understanding. IEEE Transactions on Circuits & Systems for Video Technology 27(6), 1290-1303 (2017).
10. Bouguet, J.Y.: Pyramidal implementation of the Lucas Kanade feature tracker description of the algorithm. Opencv Documents 22(2), 363-381 (1999).
11. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. Proc. 7th IJCAI, 1981 73(3), 674-679 (1981).
12. Shi, J.: Good features to track. In: Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on 2002, pp. 593-600
13. Tomasi, C.: Detection and tracking of point features. Technical Report 91(21), 9795-9802 (1991).
14. Saxena, S., Mond, F., ois, Thonnat, M., Ma, R.: Crowd Behavior Recognition for Video. In: Advanced Concepts for Intelligent Vision Systems, International Conference, Acivs 2008, Juan-Les-Pins, France, October 20-24, 2008. Proceedings 2010, pp. 970-981
15. Chaker, R., Aghbari, Z.A., Junejo, I.N.: Social Network Model for

Crowd Anomaly Detection and Localization. Pattern Recognition 61, 266-281 (2016).
16. Mousavi, H., Galoogahi, H.K., Perina, A., Murino, V.: Detecting Abnormal Behavioral Patterns in Crowd Scenarios. Springer International Publishing, (2016)
17. Zhang, Y., Qin, L., Ji, R., Yao, H., Huang, Q.: Social Attribute-Aware Force Model: Exploiting Richness of Interaction for Abnormal Crowd Detection. IEEE Transactions on Circuits & Systems for Video Technology 25(7), 1231-1245 (2015).
18. Dahrendorf, R.: Toward a Theory of Social Conflict. Journal of Conflict Resolution 2(2), 170-183 (1958).
19. Wheelan, S.A.: The handbook of group research and practice. SAGE Publications, (2005)
20. Schmidhuber, J., rgen: Deep learning in neural networks. Elsevier Science Ltd., (2015)
21. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Li, F.F.: Large-Scale Video Classification with Convolutional Neural Networks. In: IEEE Conference on Computer Vision and Pattern Recognition 2014, pp. 1725-1732
22. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: International Conference on Neural Information Processing Systems 2012, pp. 1097-1105
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. 770-778 (2015).
24. Shao, J., Kang, K., Chen, C.L., Wang, X.: Deeply learned attributes for crowded scene understanding. In: Computer Vision and Pattern Recognition 2015, pp. 4657-4666
25. Unusual crowd activity dataset of university of Minnesota, available from http://mha.cs.umn.edu/movies/crowd-activity-all.avi, Std
26. PETS 2009 Benchmark Data. multisensor sequences containing different crowd activities. http://www.cvg.rdg.ac.uk/PETS2009/a.html, Std.

Cover letter