IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 2, 2021, accepted September 30, 2021, date of publication October 5, 2021, date of current version October 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3118009

Efficient Anomaly Detection in Crowd Videos Using Pre-Trained 2D Convolutional Neural Networks

ABID MEHMOOD
Department of Management Information Systems, College of Business Administration, King Faisal University, Al-Ahsa 31982, Saudi Arabia
e-mail: aafzal@kfu.edu.sa

This work was supported by the Deanship of Scientific Research at King Faisal University through Nasher Track under Grant 206054.

ABSTRACT Surveillance of crowded places can benefit from improved techniques of anomaly detection in crowd videos. Several existing methods have detected various types of crowd abnormal behaviors by using spatial and temporal information got from videos. So far as real-time detection of anomalies is concerned, special attention must be given to reducing the model complexity that leads to computational and memory loads. This paper proposes a low computational cost approach to detect crowd anomalies. The proposed approach avoids the expensive optical flow calculations by adopting a pre-trained 2D convolutional neural network (CNN) for motion information and implements a lighter form of the 2D CNN to achieve high recognition accuracy at low computational cost. Experiments on public datasets show that the proposed model outperforms the existing approaches in terms of detection accuracy alongside providing better performance in generating input frames.

INDEX TERMS Spatial-temporal CNN, anomaly detection, crowd abnormal behaviors, violence detection, crowd surveillance.

I. INTRODUCTION

Surveillance applications are becoming more important for effective monitoring of crowded places. Video surveillance systems must observe abnormal activities involving unusual crowd activity such as crowd chaos, e.g., crowd running in one direction or dispersing from a central point, as well as violent interactions at crowded places, e.g., assault and fighting. The continuous tracking of surveillance videos is not workable for humans. Therefore, recent research has focused on proposing efficient methods for autonomous monitoring systems. There are two broader types of difficulties in developing such systems. First, there are difficulties in identifying the abnormal behavior itself in scenes containing many people in proximity, as the individuals often appear with high volatility and there is frequent occlusion. The challenges are further augmented by the irregular motion patterns found in a crowded scene. Crowd unusual activity is identified based on different parameters, such as its movement pattern and speed, as well as emerging point. Similarly, an action apparently looking as aggressive may actually be normal and vice versa. In addition, there are many types of essentially different abnormal behaviors. The second type of difficulty originates from the high computational complexity involved in most of the abnormal behavior recognition algorithms, which makes it unfeasible to use them in the real-world where detection of anomalies in the real-time is required.

To address the first difficulty, more inclusive methods need to be developed based on various abnormal events in a crowded environment, focusing specifically on distinct possibilities related to each abnormal behavior. Several works in the literature have focused on different abnormal events in crowded scene. In particular, methods have been proposed to detect the unexpected presence of a non-pedestrian within a crowd (e.g., [1]–[3]), anomalous motion patterns of pedestrians (e.g., [1], [4]), escape panics (e.g., [5], [6], [7]), violent behaviors (e.g., [8], [9]), and traffic accidents (e.g., [8]). As it is not workable to analyze the behavior of individuals in a crowded scene, many studies such as [4], [10] work towards obtaining an overall view of the scene for the detection of abnormal behavior. So far as the modalities are concerned, one common requirement for all behavior recognition methods developed for videos is that they must incorporate both spatial and temporal features, since both types of features

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

characterize actions. Many methods, such as [10]–[14] have been proposed to represent temporal information. The recent developments in deep learning techniques have motivated researchers to exploit their strength to augment the traditional methods in anomaly detection problems. Some examples include developing the deep Gaussian mixture model to learn normal events [15], [16], applying adversarial learning to identify abnormal frames [17], [18], and performing recurrent neural network-based sparsity learning for anomaly detection [6].

Real-time detection of anomalies has been the focus of several classical [1], [19] as well as deep learning-based methods [2], [20]. The latter type of methods has generally achieved superior performance with high computational costs, whereas the methods of the former type execute with low computational resources but provide lower performance. For enabling the real-time detection, the complexity of models must be reduced. Therefore, many efforts have aimed at low complexity such as cascading the local and global descriptors [21], combining the low complexity features only rather than high-level semantic features [22], using a spatiotemporal auto-encoder network to extract abnormal behaviors automatically [23], and creating spatiotemporal cuboids by dynamically extracting temporal features [24]. Among the deep learning methods, an interesting approach proposed by Kim *et al.* [13] attempts to eliminate the need for training 3D CNNs using a video dataset. In particular, it suggests to exploiting 2D CNNs pre-trained on images for learning both spatial and temporal information. This approach enables the use of many readily available image-trained 2D CNNs, and also significantly reduces the requirements for computational and memory resources. The current paper attempts to exploit the strength of the aforementioned approach by extending it to the specific problem of crowd abnormalities detection, and contributes to the state-of-the-art by improving the detection accuracy while limiting the resource loads.

Following are the major contributions of this study.

- The paper reviews the recent developments in crowd abnormality detection to determine a common set of abnormalities to cover a wider set of abnormal behaviors in crowded scenes. It focuses on two common categories, i.e., escape panics and violent interactions. Some examples of the former type of crowd unusual activities include crowd running in one direction, crowd dispersing from a central point, and activities involving crowd chaos or evacuation. Similarly, common examples of violent interactions found in crowded scenes include assault, fighting, trampling.
- The study combines the detection of various commonly found abnormal behaviors in a crowded scene. In this way, after identifying the specifics of the selected abnormal events, most relevant datasets are used to guide the training process of an abnormal event detection model.
- The paper makes a few contributions to improve specifically the efficiency of the proposed detection system. The study makes use of advanced deep learning architectures to incorporate both spatial and temporal information simultaneously. Thus, a new model is proposed that achieves the best performance with relatively shallower network depth.
- To further enhance the efficiency, the approach employs an advanced method of motion representation within the temporal stream, thus avoiding the use of expensive optical flow, yet utilizing the motion features to recognize the behavior with high accuracy. For this purpose, an existing method of general behavior recognition is extended in the specialized context of crowd abnormalities detection.
- Similarly, the use of only the interesting parts of frames is proposed to avoid processing of irrelevant volumes of pixels.

In the rest of this paper, related studies are outlined in the next section. Section III discusses the proposed approach, whereas Section IV provides details of experiments and evaluation of the approach. Finally, Section V concludes the paper.

II. RELATED WORK

This section reviews the recent research focused on proposing efficient methods for autonomous monitoring systems for crowd abnormal events. A summary of the state-of-the-art anomaly detection methods is provided in Table 1.

A. TRADITIONAL METHODS

Several supervised methods, which require data labeled as both normal and abnormal classes, have been proposed to detect abnormal events and behaviors. Ullah *et al.* [1] developed a Gaussian kernel-based integration model that integrates spatiotemporal features to capture anomalous entities. A recurrent conditional random field (R-CRF) is then trained using these features to detect anomalies. In [2], object-centric convolutional auto-encoders are used to learn motion and appearance information. Training samples are divided into clusters, each containing a certain kind of normality. A binary classifier then trains by separating the positively labeled data points in a cluster from those negatively labeled in all other clusters. Biswas and Gupta [3] represent motion using feature matrices that are decomposed into sparse components corresponding to abnormal activities. Khan *et al.* [5] proposed a method based on outlier rejection [5] that excludes the pixels containing the direction of motion unconforming with the dominant motion direction. They carried out classification using uni-variate Gaussian discriminant analysis with the K-means algorithm. Zhou *et al.* [6] proposed AnomalyNet that uses three enhanced neural processing blocks jointly for feature learning, sparse representation, and dictionary learning.

Another category known as semi-supervised methods learns a normalcy model only (one-class classification) from a provided normal training set. For example, the texture

TABLE 1. Categorization of anomaly detection methods.

Reference	Supervision category	Feature/ Model	Type(s) of anomaly	Dataset(s)
Traditional Methods				
Ullah et al. [1]	Fully supervised	Spatiotemporal, R-CRF	NP, AP, EP, DV	UCSD, UMN, UCD
Ionescu et al. [2]	Fully supervised	Motion gradient, object-centric AE	NP, EP, DV	UCSD, Avenue, UMN, ShanghaiTech
Biswas and Gupta [3]	Fully supervised	Motion, sparse component tracking	NP, EP	UCSD, UMN
Khan et al. [5]	Fully supervised	Histogram of super-pixel, Gaussian discriminant analysis	NP, EP, TA	UCSD, UMN, LV
Zhou et al. [6]	Fully supervised	Fused motion, AnomalyNet	NP, EP, DV	UCSD, Avenue, UMN
Hao et al. [19]	Semi supervised	Spatiotemporal, Texture classification	EP	UMN
Hatirnaz et al. [20]	Semi supervised	Motion, semantic search	EP, DV	UMN, PETS2009
Chaker et al. [4]	Semi supervised	Motion tracklets, Social network model	NP, AP	UCSD
Zhang et al. [10]	Semi supervised	Image appearance, fluid forces features	NP, EP	UCSD, UMN
Singh et al. [14]	Semi supervised	OF magnitude and directions, KNN, Kmeans	EP	UMN, PETS2009, Avenue
Li et al. [21]	Unsupervised	Spatial-temporal, dictionary learning	NP, EP	UCSD, UMN
Hu et al. [22]	Unsupervised	Motion, dictionary learning	NP, EP, DV	UCSD, UMN, Avenue, PETS2009
Bansod et al. [23]	Unsupervised	HoMM, K-means clustering	NP, EP	UCSD, UMN
Yuan et al. [24]	Training-less	Structural motion, 3D-DCT	NP, EP	UCSD, UMN
Sikdar and Chowdhury [25]	Training-less	Motion, 3D-DCT	NP, EP, DV	UCSD, UMN, CUHK-Avenue, ShanghaiTech
CNN-Based Method				
Sabokrou et al. [18]	Fully supervised	CNN, Sparse AE	NP, DV	UCSD, Subway
Singh et al.[7]	Fully supervised	CNN, Ensemble of CNNs	NP, DV	UCSD, Avenue
Shao et al. [8]	Fully supervised	MC-CNN, MIR-OF	EP, TA, VL	IR-Flying
Lin et al. [9]	Fully supervised	CNN, C3DGAN	VL, DV	SHADE
Farooq et al. [26]	Fully supervised	FTLE field, CNN	DV	UCF, UMN, PETS2009, NGSIM
Xu et al. [27]	Fully supervised	Local feature tracklets, GMM, CNN	EP	UMN, PETS S3
Direkoglu [28]	Fully supervised	MII, CNN	EP	UMN, PETS2009
Hasan et al. [29]	Semi supervised	CNN, motion, Convolutional AE	NP, DV	UCSD, Avenue
Hu et al. [30]	Semi supervised	HLSOF, Faster R-CNN, SVM	NP, EP, DV	UCSD, UMN, Avenue, Subway exit
Ramchandran and Sangaiah [31]	Unsupervised	Edge images, Canny edge detector	NP, EP	UCSD, Avenue
Li et al. [32]	Unsupervised	3D gradient, ST-AAE, ST-CAE	NP, EP	UCSD, UMN, Avenue
Wang et al. [33]	Unsupervised	FCN, VAE	NP, EP	UCSD, UMN, Avenue, PETS
Nawaratne et al. [34]	Training-less	CNN, Active Learning with Fuzzy Aggregation	NP, DV	UCSD, Avenue
Fan et al. [35]	Supervised	Spatiotemporal auto-encoder CNN	NP, EP	UCSD, UMN
Li et al. [36]	Supervised	Attention-based spatial stream, optical flow, CNN	VL	Hockey fights, Movies, Violent flows
Asad et al. [37]	Supervised	Multi-level future fusion, wide-dense residual block	VL	Hockey fights, Movies, Violent flows, BEHAVE
Ullah et al. [38]	Supervised	MobileNet CNN, 3D CNN	VL	Hockey fights, Movies, Violent flows
Song et al. [39]	Supervised	Key frames sampling, 3D CNN	VL	Hockey fights, Movies, Violent flows

NP: Non pedestrians EP: Escape panics AP: Anomalous pedestrian motion patterns
VL: Violence DV: Deviation from observed behavior TA: Traffic accidents/ abnormalities

extraction method proposed by Hao *et al.* [19] employs Gabor-filtered textures with the highest information entropy values to improve extraction of textures with rich details of crowd motion. They identified abnormal behaviors based on gray level co-occurrence matrix model. Recently, Hatirnaz *et al.* [20] developed a concept-based search interface based on optical flow features that annotating videos using a semantic metadata model. Chaker *et al.* [4] use a social network model to capture the crowd interaction using an unsupervised framework. They first create spatiotemporal cuboids by partitioning a video scene and then model the crowd behavior in each cuboid using local social networks (LSN), which are then used to build a global social network for a scene that detects and localizes abnormal behaviors. Zhang *et al.* [10] adopted the image appearance and fluid forces features to represent the crowd motion in a consistency group. They define each consistency group to contain pedestrians with similar spatial information, which provides a scene perception. A one-class SVM is used to detect anomalies. Singh *et al.* [14] detect abnormalities in crowd motion patterns based on motion magnitude and direction determined by optical flow.

Unsupervised methods work based on an assumption that majority of data in unlabeled dataset follow a trend, so the behaviors found in the least fit data are considered as anomalous. Several approaches have been proposed in the literature that distinguish data based on inter-data relationships. Li *et al.* [21] proposed an approach that represents activity patterns within a video cube using a codebook and then uses a reconstruction-cost criterion to detect anomalies. Hu *et al.* [22] used contextual gradients for small regions within the events to construct a histogram descriptor for abnormal event detection. Bansod *et al.* [23] employed histogram of magnitude and momentum (HoMM) features to incorporate appearance and motion characteristics and learned the behavior of objects using a clustering technique. Methods that do not require training data are often referred to as training-less approaches. These methods rely only on the test video inferences and learn spontaneously by parameter adjustments and adaptation. Yuan *et al.* [24] developed a structural context descriptor (SCD) based on particle interaction to represent crowd anomaly. It tracks the targets in different frames using 3D discrete cosine transform (DCT). Recently, Sikdar and Chowdhury [25] proposed a framework that detects objects under motion using a multi-object association-based mechanism. A temporal saliency guided optical flow map then described local motions in the video frames. Once the local descriptors are constructed, the magnitude of local disruption is calculated. A frame is anomalous if the local changes are beyond a threshold value.

B. DEEP LEARNING-BASED METHODS

CNN-based supervised methods have been proposed to detect various anomalous events and behavior. Sabokrou *et al.* [18] take a pixel-wise average of video frames to create a sequence to be passed on to a pre-trained fully convolutional network, which generates multiple grids (represented as feature vectors set) describing a specific region of the input. The feature vectors farther from the reference model are detected as describing an anomaly. Singh *et al.* [7] employed an aggregation of several pre-trained CNN models for anomaly detection. An ensemble of different fine-tuned CNN architectures allows extraction of distinctive crowd features to identify the characteristics of normal and anomalous events in the crowded scenes. Shao *et al.* [8] built an unmanned aerial vehicles (UAV) system that adopts crowd density and velocity estimation techniques based on multitask cascading CNN (MC-CNN) and multi-scale infrared optical flow (MIR-OF), respectively, and compares the value of fused descriptors with respective threshold to determine abnormal crowd behavior. Lin *et al.* [9] collected a large synthetic dataset to simulate commonly found abnormal events and behaviors. They adopted 3D CNN as backbone and applied a self-attention mechanism to further improve the abnormality detection performance. To reduce the gap between the synthetic data and real-world situations, they designed a Cyclic 3D Generative Adversarial Network (C3DGAN) that transforms the synthetic videos into realistic ones. Farooq *et al.* [26] proposed the use of finite-time Lyapunov exponent (FTLE) field to represent crowd-dominant motion and used a CNN to detect divergence behavior in crowded scenes. Xu *et al.* [27] proposed an approach that detects crowd escape panic behavior by using a dual-channel CNN. It performs spatiotemporal segmentation, and a feature descriptor is defined based on attributes of feature points. Direkoglu [28] proposed an approach based on a new Motion Information Image (MII) model, and used a CNN to detect abnormal events containing panic and escape behaviors. A spatiotemporal auto-encoder was proposed [35] that automatically learns and extracts the examples of abnormal behavior from datasets to be used by a classification CNN. Li *et al.* [36] fuse attention based spatial RGB stream with commonly used temporal and spatial streams to propose a multi-mode fusion method to detect violence. Asad *et al.* [37] detect violence by combining a CNN and a wide-dense residual block to learn spatial features and LSTM units to learn temporal features. Ullah *et al.* [38] use a lightweight CNN to identify frames containing persons and pass those frames to a 3D CNN for detection of abnormalities. Similarly, a 3D CNN is used in [39] that samples the key frames based on gray centroid prior to passing them for classification.

Semi-supervised approaches have used CNNs to learn normal events and behavior and detected abnormalities based on deviation from the normalcy. Hasan *et al.* [29] utilized an auto-encoder to adopt handcrafted features containing motion information and to learn temporal uniformity in videos. The fundamental idea is that the auto-encoder will reconstruct the motion sequences found in normal videos with low error as compared with those found in anomalous videos. So, if a frame results in high reconstruction error, it is considered as containing anomaly. Hu *et al.* [30] used Faster Regional CNN (Faster R-CNN) to detect objects and Histogram of Large-Scale Optical Flow (HLSOF) to define their action. A multiple instance SVM is trained and used to detect abnormal behavior.

Unsupervised and training-less methods have mainly focused on detection of non-pedestrians and escape panic behaviors. Ramchandran and Sangaiah [31] developed an

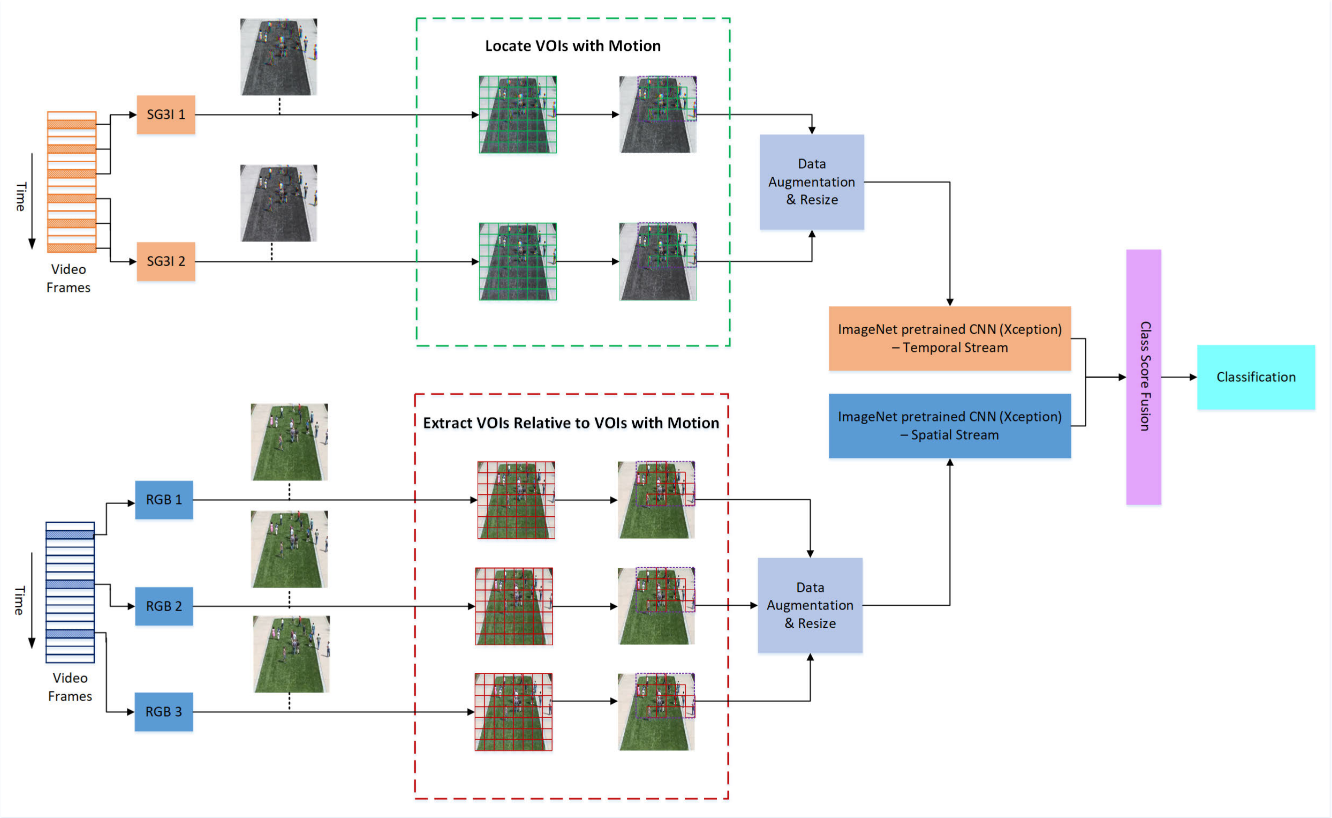
Time
Video Frames
SG3I 1
SG3I 2
Locate VOIs with Motion
Data Augmentation & Resize
ImageNet pretrained CNN (Xception) – Temporal Stream
ImageNet pretrained CNN (Xception) – Spatial Stream
Class Score Fusion
Classification
Extract VOIs Relative to VOIs with Motion
RGB 1
RGB 2
RGB 3
Time
Video Frames
Data Augmentation & Resize

FIGURE 1. The overall structure of the crowd abnormal behavior recognition approach.

unsupervised learning framework that generates reconstructed frames after making a combined input of raw image sequences and edge image sequences to the convolutional auto encoder LSTM model. The model detects anomaly by calculating reconstruction error. Li *et al.* [32] used a spatial-temporal adversarial auto-encoder (ST-AAE) and a spatial-temporal convolutional auto-encoder (ST-CAE) that apply 3D convolution and devolution, respectively, to detect patterns from temporal dimensions. Anomalies are detected in two steps. First, they apply ST-AAE to filter out clearly normal cuboids and classify apparently suspicious cuboids from raw 3D video cuboids. Next, ST-CAE is employed to identify the specific anomalous patches from suspicious cuboids by determining the reconstruction error. Wang *et al.* [33] designed new generative models to detect abnormal objects. They extract the foreground using a pre-trained FCN and calculate the optical flow for motion features, which are then used as input to two neural network models that filter normal samples. Finally, the reconstruction error of the input with a preset threshold determined during training is used to detect the abnormalities. Nawaratne *et al.* [34] have proposed a training-less incremental spatiotemporal learner that makes use of active learning with fuzzy aggregation of CNN features to keep track of new anomalies by continuously developing the definitions of normal behavior.

III. THE PROPOSED APPROACH

This work proposes an approach to detect abnormal behaviors in videos containing crowded scenes. To achieve high performance in detection, this approach makes use of both the spatial appearance information and the complex motion features to detect and localize anomalies in crowded scenes. For this purpose, a two-stream CNN structure [40] consisting of a spatial and a temporal stream is used. The former stream of the CNN learns the spatial characteristics of objects in the scene, whereas the latter trains on indispensable temporal features. In pursuit of reduced computational complexity, the method makes some improvements in how the temporal information is extracted. First, unlike previous works (e.g., [21], [41]) that process densely samples from each frame, leading to high computational costs, this method discards the patches that contain little motion information. So, it uses data from only selected volumes of interest (VOIs) carrying rich information pertaining to motion. Therefore, for example, regions where no pedestrians appear are discarded. Second, to extract and represent the temporal information, a low computational cost method is applied instead of optical flow used by several similar approaches in the past. Specifically, we adopt the stacked grayscale 3-channel image (SG3I) of Kim and Won [13] for this purpose. Hence, to extract the VOIs, our approach applies SG3Is and keeps them in training the temporal stream of the CNN. A 2D CNN pre-trained

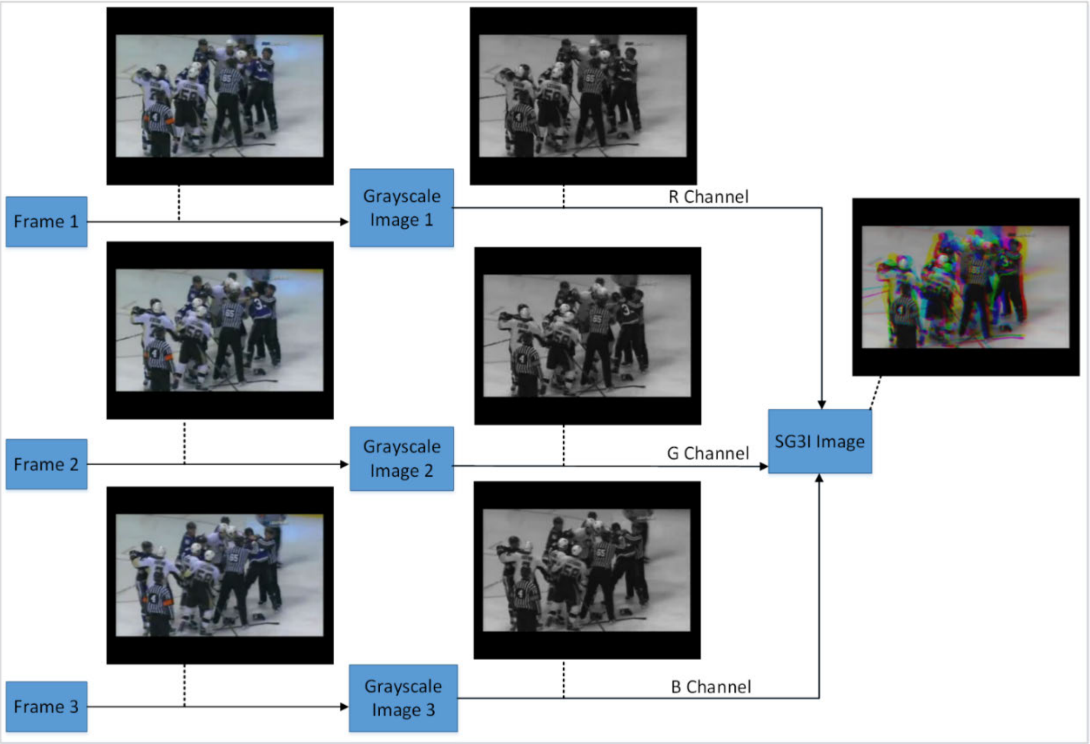

FIGURE 2. Conversion of sequential video frames into SG3I format.

on still images is adopted by feeding it the SG3Is extracted from video shots. Figure 1 shows the overall structure of the proposed approach. The details of generating the input images and their classification using the pre-trained network are discussed in the following subsections.

A. INPUT IMAGE FORMATION

The approach uses a two-stream architecture where the spatial stream learns the spatial appearance of objects in the scene, whereas the temporal stream learns the motion features found in consecutive frames of video. Since a pre-trained 2D CNN is adopted for both spatial and temporal streams, input images must be created to train each stream. Specifically, representative RGB images are formed to be input to the spatial stream, whereas SG3I images are provided as input to the temporal stream. The following subsections provide more details on the generation and processing of 2D images from video frames for both streams.

1) SG3I IMAGES FOR TEMPORAL STREAM

As mentioned previously, our approach aims to eliminate the regions with little motion information. To this end, we use the stacked grayscale 3-channel image (SG3I) format [13], as it has been shown to detect motion effectively while reducing the computational expense involved in other approaches, such as optical flow. The VOIs containing informative details of the event are obtained in two steps: (i) converting to the SG3I format, and (ii) determining if the VOIs contain useful information or noise.

The conversion to SG3I involves creating a single 3-channel RGB color image from three sequential frames extracted from the video. For this purpose, first, a grayscale image is generated from each of the three selected frames, and then a new image (SG3I) is created that incorporates each of the grayscale images in its RGB channels. Thus, the resultant SG3I image is a single-color image in which the colored regions (i.e., hue) pertain to brightness variations of the corresponding pixels along the time-axis through the chosen frames. Figure 2 shows the steps involved in forming the SG3I images from selected sequential frames. Notice that the SG3I image shows grayscale output with no hue for a pixel with an identical value for all three RGB channels. On the other hand, it shows a color to indicate a displacement in brightness in case the RGB values differ at a pixel in the selected frames. In this way, the colored regions resulted by the pixels with different RGB values effectively represent the motion patterns. An important configuration to be made here involves the selection of the three image frames to create an SG3I. Specifically, on the one hand, the frames need to be selected uniformly ensuring that the time interval between them is short enough to avoid noisy motions. On the other hand, the interval must not be made too short to make it hard to absorb meaningful information about the action. To ensure a balance, a technique similar to [13] was applied. The entire video clip is divided into several sub-clips and one SG3I is generated for each sub-clip to ensure selecting the representative frames. Figure 3 shows a sample of frame sequences taken from various datasets adopted in this study and a visualization of their corresponding SG3I images.

Further, to locate the VOIs with useful information, a set of non-overlapping patches of fixed size are defined to cover the entire SG3I image. The size of patches is kept as adjustable to be set once for each dataset in a way that it is small enough to capture the details of behavior but large enough to cover the related details of appearance. The VOIs and the respective pixels are considered informative if at least 70% of pixels contain motion (i.e., color or hue as determined by the difference of values among RGB channels in the same pixel of SG3I) and thus they are preserved by taking an aggregate of all such VOIs as shown in Figure 1. Other VOIs and their pixels are considered noise and are discarded. It is important to note that the spatial size of VOIs extracted from different datasets may vary since we adjust the size for each dataset. Therefore, the VOIs are resized continually to make them consistent with the input requirements of the CNN model. After processing, the sets of SG3Is obtained from sub-shots of the video are used to fine-tune the temporal stream of the CNN.

2) RGB IMAGES FOR SPATIAL STREAM

To enable the model to recognize the spatial appearance of objects accurately, it is important to select a representative frame from the video. To ease the difficulty in selecting the best frame with the object, a technique similar to the frames of the temporal stream is applied. Therefore, a video clip is divided into several sub-clips, thus selecting one representative frame for each sub-clip. However, there is one important detail here. Recall that the image inputs (SG3Is) for the temporal stream were resized to pass only the VOIs containing motion to the CNN. Therefore, the VOIs from both types of images, i.e., SG3Is and RGBs, must be correlated to ensure a collective meaning for the CNN. Thus, the method keeps a track of the frames and the pixels containing motion information selected while creating input for the temporal stream. Specifically, while forming RGB image input, it selects one of the three frames selected previously to create SG3I. Then, within the selected frame, it locates and extracts the VOIs from the same location where the moving object was previously detected. In this way, after obtaining multiple RGB images from the video shot, they are used to finetuning the spatial stream of the CNN.

B. CNN ARCHITECTURE

Xception [42] is adopted after some changes in its architecture to be used as the pre-trained 2D CNN in each of the two streams. Based on Inception [43], the Xception network uses a modified depth-wise separable convolution instead of inception modules to reduce the number of parameters. Also, to enhance the efficiency and performance, it uses residual connections originally proposed by ResNet [44] for all flows. The overall architecture of Xception network contains three main flows, i.e., entry, middle, and exit flows. For space reasons, this section will only focus on the middle flow as this work makes some modifications to this part. The middle flow is sometimes referred to as the core structure part, and it comprises a 9-layer structure that repeats 8 times. Within the 9-layer structure, there are 3 layers each of Relu, separable Conv2D, and batch normalization. It is important to note here that the core structure contains a huge number of convolution layers. Deepening the network is often appropriate in complex scene classification, however, reducing the number of convolution layers also greatly affects the network performance, as previously reported in the literature [45], [46]. Therefore, inspired by the work of Shi *et al.* [47], it was deemed worthwhile to explore the relationship between the depth of the network and its performance specific to the current dataset. For this purpose, we preserved the non-core (non-repeating) structure of the network and tried with different network structures within the core part. Specifically, recall that each repetition in the core contains 3 layers each of Relu, separable Conv2D, and batch normalization. Let us call this combination of layers as RCB. Thus, there are 8 repetitions of RCBs within the core structure. So, by varying the number of RCBs, 7 different lightweight network structures (let us call *LWXception*) were created such that LWXception1 contains 1 RCB, LWXception2 contains 2 RCBs, and so on, LWXception7 contains 7 RCBs. The performance of each core structure was monitored and finally the core structure containing 3 RCB layers shown in Figure 4 was adopted and used within both spatial and temporal streams, because of its high score in terms of accuracy and number of parameters. More details of the results of experiments with each configuration are given in Section 4. Note that both types of images used by this work (RGB and SG3I) are compatible with the input format of Xception, and thus are directly fed to fine-tune the individual streams after resizing and augmentation as previously shown in Figure 1.

C. FINE-TUNING AND TESTING

For the fine-tuning on the selected datasets, RGB and SG3I images were obtained using the techniques explained in Section 3.1. To generate the images during the training and testing phases, the respective split of each dataset was used as detailed later in Section 4. Here, the batch size was set as 32, and some data augmentation techniques were applied for both types of images. Specifically, the system first performed a random cropping to the center for both RGBs and SG3Is, followed by a horizontal flip, and finally resized them to 229×229. Further, the model performance was analyzed with various network configurations by applying different optimizers and various values of the learning rate and momentum. In general, whenever the loss reduction was not observed for over 10 epochs after setting a specific learning rate, we reduced the learning rate by 1/10. The specifications of each of the three optimizers used by the system are as given in the following. Stochastic gradient descent (SGD) was used with a learning rate of 0.001, momentum 0.9, weight decay

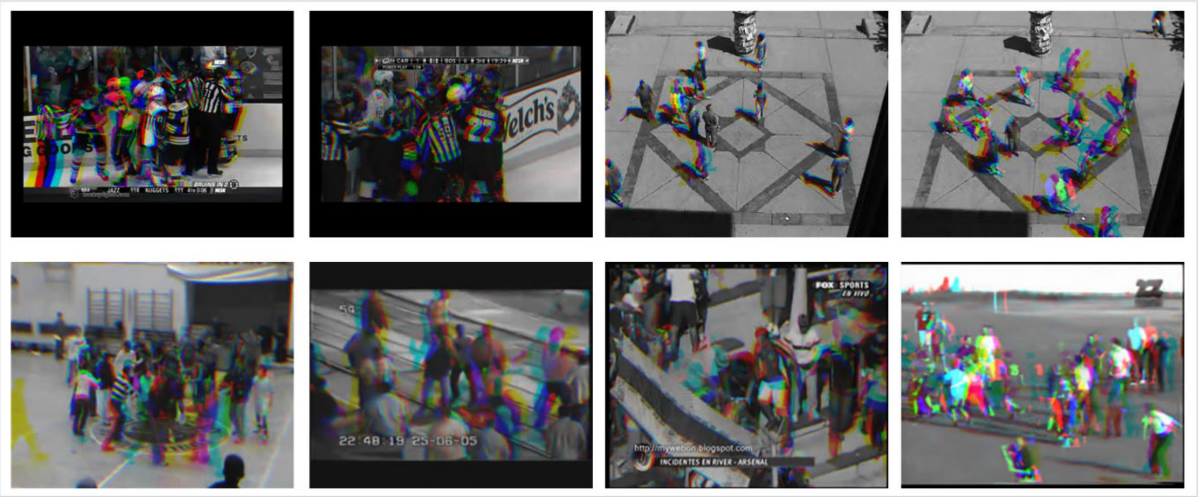

FIGURE 3. Sample of SG3I images generated for hockey fights, UMN, and violent flows.

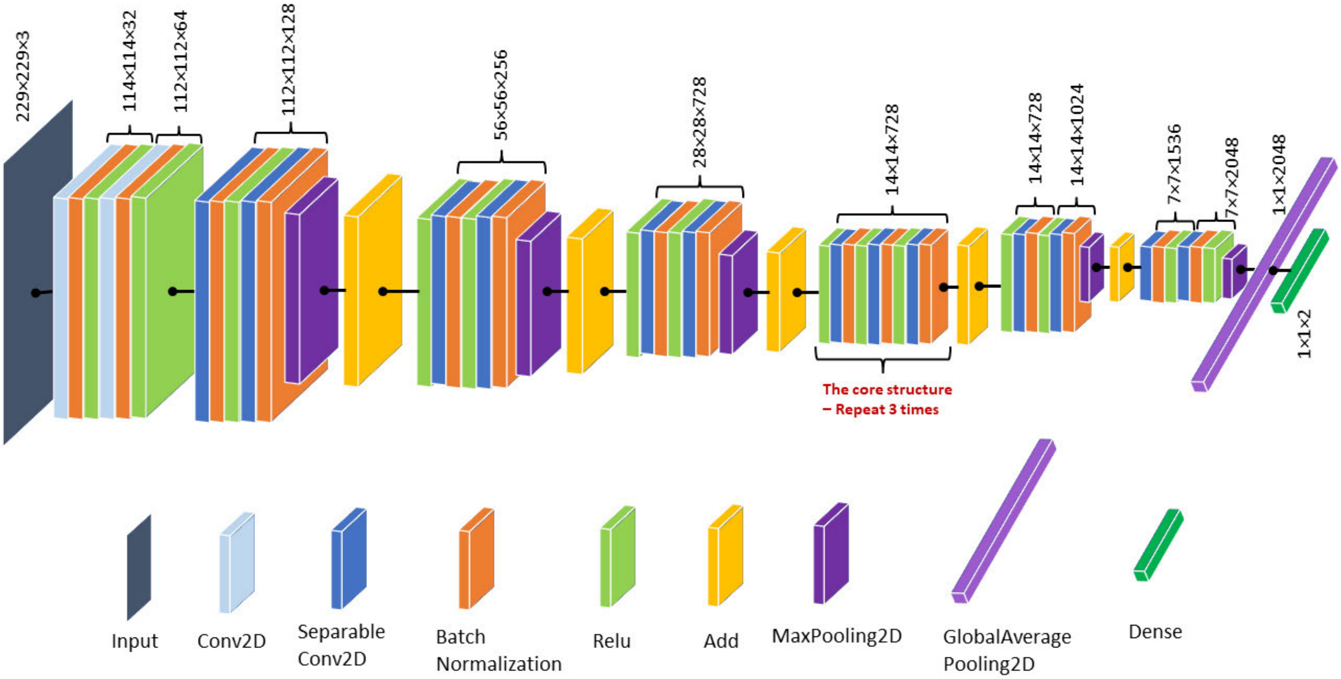

FIGURE 4. Proposed structure of Xception network used within each stream of CNN.

of 0.0005, and a nesterov value of False. RMSProp was used with a learning rate of 0.001, rho value of 0.9, and epsilon value of 1×10^{-7}. Adam was used with a learning rate of 0.001, beta1 value of 0.9, beta2 value of 0.999, epsilon value of 1×10^{-8}, and amsgrad value of false. Since the last network configuration yielded the best results, the same was adopted for testing and the rest of experiments discussed in the next section.

IV. EXPERIMENTS

We evaluated the proposed approach using several experiments designed to assess its capability in detecting and localizing abnormalities in crowded scenes, as well as to compare it with the state-of-the-art methods proposed in the literature. The system was implemented in Python using Keras with TensorFlow 2.0 on Ubuntu 20.04 OS. As detailed later in this section, during training and experiments, both CPU-only (Intel i7-8650 @2.11 GHz 32GB RAM) and GPU (NVIDIA GTX 1080Ti 11 GB) configurations were used. In general, three main types of experiments were performed. First, the best network configuration was determined in terms of various parameters and the architecture of the Xception network and layers in the core structural part. Here, we also determine the effectiveness of the use of the volumes of interest instead of the entire frame. Second, the abnormality detection capability of the proposed model was evaluated by using various metrics. Third, the performance of the approach is compared with the state-of-the-art crowd abnormality detection methods.

A. DATASETS

The performance of the proposed approach was evaluated by adopting three public datasets. Specifically, UMN [48], Hockey Fights [49], and Violent Flows [50] datasets were used. Note that we selected these datasets because of their close relevance to the two general categories of crowd abnormalities addressed in this study, i.e., escape panic and violent interactions.

UMN dataset pertaining to unusual crowd activity comprises 11 videos with a resolution of 240 × 320 each containing both normal and abnormal crowd behaviors. The videos are shot in three different scene settings—a lawn, interior, and plaza. The number of videos in the first scene, second scene, and third scene are 2, 6, and 3, respectively. To enable the network to learn the specifics of abnormal behavior, a uniform strategy for obtaining and using the train/test splits was adopted. Specifically, while testing a particular video in a scene, the video being tested was left out and the remaining videos in that scene were used to train the model.

The Hockey Fights dataset comprises 1000 video clips with a resolution of 360 × 280 divided into two groups, i.e., fights and non-fights. The videos are shot from various angles and both normal and violent activities occur in a comparably dynamic setting. To meet the requirements of a crowded scene violence, 240 video clips involving several players (crowded scene) were selected manually for training. Further, these videos were divided into 3 separate groups, each containing 80 videos (40 each from the fights and non-fights groups). Here, we created a separate split of 20 videos to be used for testing.

The Violent Flows dataset includes 246 training and 21 testing videos of resolution 320 × 240 obtained from YouTube that contain real-world crowds engaging in violent activities. All the videos are produced in an uncontrolled environment, thus providing a wide range of scene and action types. The videos in the training set are divided into 5 distinct sets. Each of the first 3 sets contains 25 videos, each of violence and non-violence type. The remaining 2 sets contain 24 each of violence and non-violence. Here, we used the testing split of 21 videos separately provided by the authors of the dataset.

B. PERFORMANCE EVALUATION OF THE APPROACH

In this section, before presenting the results of overall system performance in various settings, we first compare the results achieved by unique structures of Xception network. Recall that the current study intended to explore the relationship between the depth of the network and its performance after fine-tuning on the SG3I images. Therefore, as mentioned in Section 3.2, seven different lightweight network structures (LWXception1 to LWXception7) were created with a different number of repetitions of the core structure. All the seven network structures and the original Xception were trained under the same conditions and the accuracy in each dataset and the number of parameters were recorded. The results are shown in Table 2. The structures LWXception3, LWXception4, and LWXception5 comprising 3, 4, and 5 combinations of RCB (relu, separable Conv2D, batch normalization) layers perform equal and provide the best accuracy as compared with the other structures. However, considering the number of parameters, the use of LWXception3 seems more appropriate as it provides the highest accuracy with a relatively shallower network structure. Hence, this structure was adopted to be used within both spatial and temporal streams. The same structure was used in all other experiments. In the rest of this section, Xception refers to the LWXception3 lightweight structure of the network.

TABLE 2. Comparison (classification accuracy and number of parameters) of unique structures of Xception using SG3I.

Method	Accuracy			Parameters
	UMN	Hockey Fights	Violent Flows	
Xception	0.9905	0.9952	0.9859	20,873,774
LWXception7	0.9905	0.9952	0.9859	19,255,430
LWXception6	0.9905	0.9952	0.9859	17,637,086
LWXception5	0.9912	0.9971	0.9881	16,018,742
LWXception4	0.9912	0.9971	0.9881	14,400,398
LWXception3	0.9912	0.9971	0.9881	12,782,054
LWXception2	0.9901	0.9958	0.9866	11,163,710
LWXception1	0.9901	0.9958	0.9866	9,545,366

In the next step, we tested the model using the test split of each dataset discussed previously. The performance of the model was measured in terms of the number of correct predictions made against both abnormal and normal classes. The confusion matrix in Figure 5 shows the results. Among all the abnormal test cases, 99.26%, 99.82%, and 99.01% are correctly detected for UMN, hockey fights, and violent flows datasets, respectively. Among all the normal test cases, 98.98%, 99.59%, and 98.61 are correctly classified for UMN, hockey fights, and violent flows datasets, respectively. One can see from the results that the model is slightly more likely to mis-classify normal behavior as abnormal. However, as given by the nature of the problem, the correct detection of abnormalities has priority over the other case. Therefore, it was deemed acceptable to have some false positives rather than mis-classifying the abnormal behavior has normal. So far as the comparative results on different datasets are concerned, the misclassifications have the highest number in case of violent flows dataset. One probable reason for this could be that, among the three datasets, violent flows dataset contains the most complex scenes and also involves a vast variety in terms of perspectives and backgrounds. To further evaluate the true positive and false positive rates of the model for each dataset, the ROC curves are obtained. Figure 6 shows

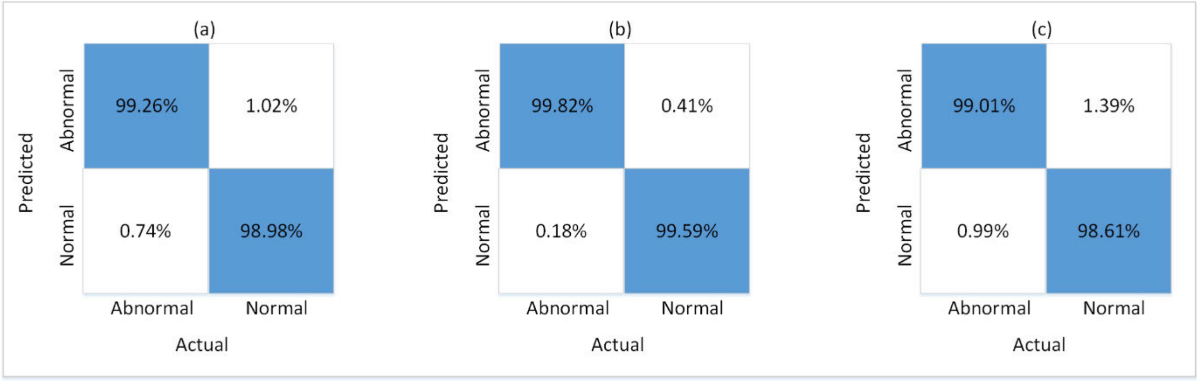

FIGURE 5. Confusion matrix of the approach on (a) UMN dataset (b) Hockey fights dataset (c) Violent flows dataset.

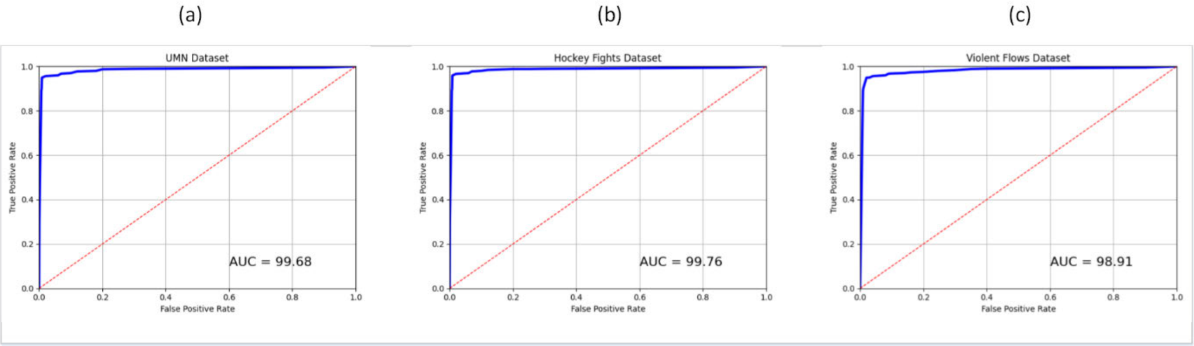

FIGURE 6. ROC curve and AUC value for: (a) UMN dataset (b) Hockey fights dataset (c) Violent flows dataset.

TABLE 3. Classification results of the proposed approach on UMN dataset.

	Spatial	Temporal	Fused Two Stream (Full Frame)	Fused Two Stream (VOIs)
Recall	0.9799	0.9487	0.9899	0.9926
FP Rate	0.0276	0.0555	0.0198	0.0102
Precision	0.9726	0.9447	0.9804	0.9898
Accuracy	0.9762	0.9466	0.9851	0.9912

TABLE 4. Classification results of the proposed approach on Hockey fights dataset.

	Spatial	Temporal	Fused Two Stream (Full Frame)	Fused Two Stream (VOIs)
Recall	0.9751	0.9482	0.9900	0.9982
FP Rate	0.0299	0.0567	0.0144	0.0041
Precision	0.9702	0.9436	0.9857	0.9959
Accuracy	0.9726	0.9458	0.9878	0.9971

the ROC curves and AUC values for each dataset. The model achieves the AUC values of 99.68, 99.76, and 98.91 on UMN, hockey fights, and violent flows datasets, respectively.

As the approach presented here uses two separate networks in the two-stream architecture, experiments are performed to study the role of each stream. The abnormality detection accuracies between the two streams are determined. Here, 5 RGB frames per video were used to fine-tune the spatial stream and 30 RGB frames to test it. On the other hand, 10 SG3I images per video were used for fine-tuning the temporal stream and 10 SG3Is to test it. The results are shown in Table 3, Table 4, and Table 5 for UMN, hockey fights, and violent flows datasets, respectively. One can see that the spatial stream CNN consistently yields better results comparative to the temporal stream on all three datasets. However, both streams augment each other when fused, as clear from the results. Further, the results show that the proposed method of using the volumes of interest (VOIs) instead of the entire frame enables the model to learn the features of abnormal behavior better. In this way, an improvement of 0.61%, 0.93%, and 1.72% can be seen in average accuracies on UMN, hockey fights, and violent flows datasets, respectively, on top of the fused two-stream accuracies yielded by a full-frame equivalent of the model. A comparison of the performance results achieved with the two-stream architecture using VOIs on different splits of the three datasets is shown in Figure 7. The comparison shows that the overall performance of the model remains almost equal in various splits of the datasets.

Furthermore, recall that one of the major goals for commencing this study was to recognize abnormal behavior

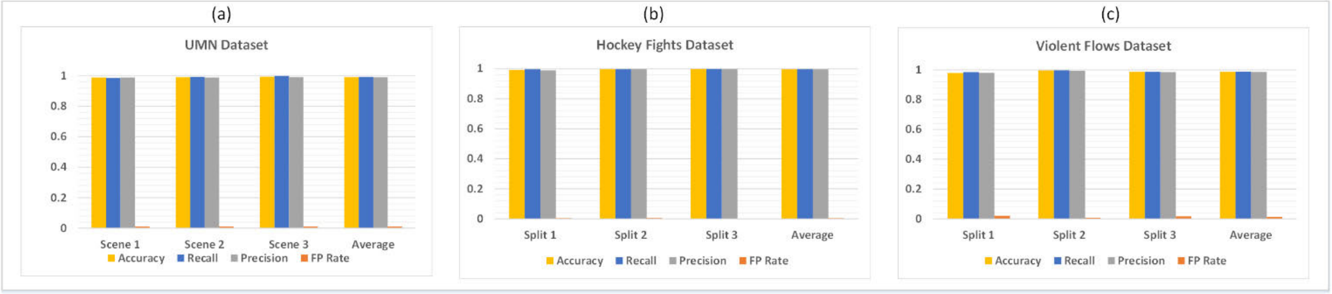

FIGURE 7. Comparison of classification results of the fused two-stream architecture with VOIs on different splits of the datasets. (a) UMN (b) Hockey fights (c) Violent flows.

TABLE 5. Classification results of the proposed approach on Violent flows dataset.

	Spatial	Temporal	Fused Two Stream (Full Frame)	Fused Two Stream (VOIs)
Recall	0.9567	0.9211	0.9703	0.9901
FP Rate	0.0488	0.0757	0.0285	0.0139
Precision	0.9515	0.9241	0.9715	0.9862
Accuracy	0.9540	0.9227	0.9709	0.9881

in crowded scenes with high accuracy while reducing the computational complexity involved in the optical flow computations required for the temporal stream. The approach presented here achieves this goal by replacing the optical flow frames in 3D CNNs with the SG3Is in the pre-trained 2D CNN. However, it is important to confirm the computational effectiveness of the use of SG3Is.

So, the next set of experiments was concerned with determining the execution times (in terms of frames per second) of the SG3Is method used in this study with the optical flow [11] and the dynamic image [12]. The results are shown in Table 6. It is important to mention here that we conducted this comparison to reinforce the results obtained by the SG3I authors [13] on the datasets encompassing a different context used in the current study. Note that we used our GPU configuration mentioned previously to get the optical flow, whereas the CPU-only configuration was used to measure the fps for the other two methods. As shown in the table, SG3I method results in significantly higher performance in generating input frames as compared with the other two methods, even though GPU was used to execute the optical flows.

C. COMPARISON WITH THE STATE-OF-THE-ART AND OTHER NETWORKS

To further evaluate the proposed approach for detection of abnormal behaviors in crowded environments, a comparison was conducted with the state-of-the-art works for the same purpose presented in the literature. For this comparison, some representative methods that yielded the highest accuracy in both categories addressed in this study (i.e., escape panic and violent interactions) were selected. Furthermore, since our dataset of SG3I images can be used with any pre-trained CNN, we also intended to determine the classification results by replacing the proposed modified Xception model with two other publicly available pre-trained models, i.e., Inception-v1 and DenseNet. It is important to note here that, since researchers have used various metrics to present their results, the comparative results will be presented using the metrics employed in the original paper. So, the methods in the escape panic category have used a mix of AUC values and accuracy, whereas the methods in violent flows category rely on accuracy as the main metric to report performance However, to conduct a more inclusive comparison of the performance of the current approach with others, the results are presented using both criteria.

TABLE 6. Execution times (frames per second) taken for input frames generation for the temporal stream.

Dataset	Optical Flow	Dynamic Image	SG3I
UMN	16.21	171.49	712.89
Hockey Fights	16.77	177.85	745.70
Violent Flows	15.90	180.12	793.36

The results of comparison with approaches in the escape panic category are tabulated in Table 7. The proposed method using the modified Xception model in two-stream architecture outperforms the existing methods in UMN dataset while providing AUC and accuracy values of 99.68% and 99.12%, respectively. In this way, the AUC value of the proposed method is better than the recently published approaches in this category including Fan *et al.* (99.6%), Hu *et al.* (98.9%), and Li *et al.* (99.60%). Similarly, the method provides higher accuracy as compared to the existing methods, such as Farooq *et al.* (98.75%) and Direkoglu (99.08%). Table 8 presents the results of comparison with approaches in the violent flows category. The current work provides better results in both Hockey Fights and Violent Flows datasets. Specifically, it provides an accuracy of 99.71% on hockey fights dataset as compared with the existing approaches in the category such as Li *et al.* (99.50%), Asad *et al.* (98.80%),

TABLE 7. Comparison of classification accuracy with the state-of-the-art methods in escape panic category.

Method	AUC% (UMN)	Accuracy% (UMN)
Fan et al. [35]	99.6	-
Farooq et al. [26]	-	98.75
Direkoglu [28]	-	99.08
Hu et al. [30]	98.90	-
Li et al. [32]	99.60	-
Ours (Inception-v1)	**98.99**	**98.33**
Ours (DenseNet)	**98.03**	**97.85**
Ours (Xception)	**99.68**	**99.12**

TABLE 8. Comparison of classification accuracy with the state-of-the-art methods in violent interactions category.

Method	AUC% (Hockey Fights)	AUC% (Violent Flows)	Accuracy% (Hockey Fights)	Accuracy% (Violent Flows)
Li et al. [36]	-	-	99.50	-
Asad et al. [37]	-	-	98.80	97.10
Ullah et al. [38]	-	-	96.00	98.00
Song et al. [39]	-	-	99.62	94.30
Ours (Inception-v1)	**99.05**	**97.63**	**98.54**	**97.12**
Ours (DenseNet)	**97.93**	**97.18**	**97.10**	**96.44**
Ours (Xception)	**99.76**	**98.91**	**99.71**	**98.81**

Ullah *et al.* (96%), and Song *et al.* (99.62%). Also, this approach shows an improvement in accuracy over the mentioned methods in violent flows dataset by providing an accuracy of 98.81% as compared with 97.10%, 98%, and 94.30% of Asad *et al.*, Ullah *et al.*, and Song *et al.*, respectively. The results in both categories of abnormal behaviors show that the approach works best with the Xception model as compared with Inception-v1 and DenseNet.

V. CONCLUSION

In this paper, an approach based on two-stream CNNs for the detection of abnormal behavior in crowded scenes was presented. The approach adopts a modified form of pre-trained 2D CNN for both the spatial and temporal streams. For this purpose, RGB image frames and the SG3Is (stacked grayscale 3-channel images) are used to fine-tune the spatial and temporal streams, respectively. Using SG3Is to represent motion features allows to avoid the use of computationally expensive alternatives for the temporal stream, such as optical flow or dynamic images. Consequently, the proposed approach achieves relatively better recognition accuracy as compared to the existing methods, while providing performance improvements comparative to the use of alternatives for the temporal stream. The experiments on UMN, hockey fights, and violent flows datasets show that the approach can efficiently detect different abnormalities included in these datasets with an accuracy of 99.12%, 99.71%, and 98.81%, respectively.

REFERENCES

[1] H. Ullah, A. B. Altamimi, M. Uzair, and M. Ullah, "Anomalous entities detection and localization in pedestrian flows," *Neurocomputing*, vol. 290, pp. 74–86, May 2018.

[2] R. T. Ionescu, F. S. Khan, M.-I. Georgescu, and L. Shao, "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7834–7843.

[3] S. Biswas and V. Gupta, "Abnormality detection in crowd videos by tracking sparse components," *Mach. Vis. Appl.*, vol. 28, nos. 1–2, pp. 35–48, Feb. 2017.

[4] R. Chaker, Z. Al Aghbari, and I. N. Junejo, "Social network model for crowd anomaly detection and localization," *Pattern Recognit.*, vol. 61, pp. 266–281, Jan. 2017.

[5] M. U. K. Khan, H.-S. Park, and C.-M. Kyung, "Rejecting motion outliers for efficient crowd anomaly detection," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 2, pp. 541–556, Feb. 2019.

[6] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, "Anomalynet: An anomaly detection network for video surveillance," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 10, pp. 2537–2550, Oct. 2019.

[7] K. Singh, S. Rajora, D. K. Vishwakarma, G. Tripathi, S. Kumar, and G. S. Walia, "Crowd anomaly detection using aggregation of ensembles of fine-tuned ConvNets," *Neurocomputing*, vol. 371, pp. 188–198, Jan. 2020.

[8] Y. H. Shao, W. F. Li, H. Y. Chu, Z. Y. Chang, X. Q. Zhang, and H. Y. Zhan, "A multitask cascading CNN with MultiScale infrared optical flow feature fusion-based abnormal crowd behavior monitoring UAV dagger," (in English), *Sensors*, vol. 20, no. 19, Oct. 2020, Art. no. 5550.

[9] W. Lin, J. Gao, Q. Wang, and X. Li, "Learning to detect anomaly events in crowd scenes from synthetic data," *Neurocomputing*, vol. 436, pp. 248–259, May 2021.

[10] X. Zhang, D. Ma, H. Yu, Y. Huang, P. Howell, and B. Stevens, "Scene perception guided crowd anomaly detection," *Neurocomputing*, vol. 414, pp. 291–302, Nov. 2020.

[11] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Computer Vision—ECCV 2004*. Berlin, Germany: Springer, 2004, pp. 25–36.

[12] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3034–3042.

[13] J.-H. Kim and C. S. Won, "Action recognition in videos using pre-trained 2D convolutional neural networks," *IEEE Access*, vol. 8, pp. 60179–60188, 2020.

[14] G. Singh, R. Kapoor, and A. Khosla, "Optical flow-based weighted magnitude and direction histograms for the detection of abnormal visual events using combined classifier," *Int. J. Cognit. Informat. Natural Intell.*, vol. 15, no. 3, pp. 12–30, Jul. 2021.

[15] Y. Feng, Y. Yuan, and X. Lu, "Learning deep event models for crowd anomaly detection," *Neurocomputing*, vol. 219, pp. 548–556, Jan. 2017.

[16] H. Purohit, R. Tanabe, T. Endo, K. Suefusa, Y. Nikaido, and Y. Kawaguchi, "Deep autoencoding Gaussian mixture model for unsupervised anomaly detection," presented at the ICLR, Feb. 2018.

[17] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1577–1581.

[18] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes," *Comput. Vis. Image Understand.*, vol. 172, pp. 88–97, Jul. 2018.

[19] Y. Hao, Z.-J. Xu, Y. Liu, J. Wang, and J.-L. Fan, "Effective crowd anomaly detection through spatio-temporal texture analysis," *Int. J. Autom. Comput.*, vol. 16, no. 1, pp. 27–39, Feb. 2019.

[20] E. Hatirnaz, M. Sah, and C. Direkoglu, "A novel framework and concept-based semantic search interface for abnormal crowd behaviour analysis in surveillance videos," *Multimedia Tools Appl.*, vol. 79, nos. 25–26, pp. 17579–17617, Jul. 2020.

[21] N. Li, X. Wu, D. Xu, H. Guo, and W. Feng, "Spatio-temporal context analysis within video volumes for anomalous-event detection and localization," *Neurocomputing*, vol. 155, pp. 309–319, May 2015.

[22] X. Hu, Y. Huang, Q. Duan, W. Ci, J. Dai, and H. Yang, "Abnormal event detection in crowded scenes using histogram of oriented contextual gradient descriptor," *EURASIP J. Adv. Signal Process.*, vol. 2018, no. 1, pp. 1–15, Aug. 2018.

[23] S. D. Bansod and A. V. Nandedkar, "Crowd anomaly detection and localization using histogram of magnitude and momentum," *Vis. Comput.*, vol. 36, no. 3, pp. 609–620, Mar. 2020.

[24] Y. Yuan, J. W. Fang, and Q. Wang, "Online anomaly detection in crowd scenes via structure analysis," (in English), *IEEE Trans. Cybern.*, vol. 45, no. 3, pp. 548–561, Mar. 2015.

[25] A. Sikdar and A. S. Chowdhury, "An adaptive training-less framework for anomaly detection in crowd scenes," *Neurocomputing*, vol. 415, pp. 317–331, Nov. 2020.

[26] M. U. Farooq, M. N. M. Saad, and S. D. Khan, "Motion-shape-based deep learning approach for divergence behavior detection in high-density crowd," *Vis. Comput.*, p. 25, Feb. 2021.

[27] Y. Xu, L. Lu, Z. Xu, J. He, J. Zhou, and C. Zhang, "Dual-channel CNN for efficient abnormal behavior identification through crowd feature engineering," *Mach. Vis. Appl.*, vol. 30, no. 5, pp. 945–958, Jul. 2019.

[28] C. Direkoglu, "Abnormal crowd behavior detection using motion information images and convolutional neural networks," *IEEE Access*, vol. 8, pp. 80408–80416, 2020.

[29] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 733–742.

[30] X. Hu, J. Dai, Y. Huang, H. Yang, L. Zhang, W. Chen, G. Yang, and D. Zhang, "A weakly supervised framework for abnormal behavior detection and localization in crowded scenes," *Neurocomputing*, vol. 383, pp. 270–281, Mar. 2020.

[31] A. Ramchandran and A. K. Sangaiah, "Unsupervised deep learning system for local anomaly event detection in crowded scenes," *Multimedia Tools Appl.*, vol. 79, nos. 47–48, pp. 35275–35295, Dec. 2020.

[32] N. Li, F. Chang, and C. Liu, "Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes," *IEEE Trans. Multimedia*, vol. 32, pp. 203–215, 2021.

[33] T. Wang, M. Qiao, Z. Lin, C. Li, H. Snoussi, Z. Liu, and C. Choi, "Generative neural networks for anomaly detection in crowded scenes," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 5, pp. 1390–1399, May 2019.

[34] R. Nawaratne, D. Alahakoon, D. D. Silva, and X. Yu, "Spatiotemporal anomaly detection using deep learning for real-time video surveillance," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 393–402 Jan. 2020.

[35] Z. Fan, J. Yin, Y. Song, and Z. Liu, "Real-time and accurate abnormal behavior detection in videos," *Mach. Vis. Appl.*, vol. 31, nos. 7–8, p. 13, Sep. 2020.

[36] H. Li, J. Wang, J. Han, J. Zhang, Y. Yang, and Y. Zhao, "A novel multi-stream method for violent interaction detection using deep learning," *Meas. Control*, vol. 53, nos. 5–6, pp. 796–806, May 2020.

[37] M. Asad, J. Yang, J. He, P. Shamsolmoali, and X. J. He, "Multi-frame feature-fusion-based model for violence detection," (in English), *Vis. Comput.*, vol. 37, pp. 1415–1431, Jun. 2020.

[38] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Violence detection using spatiotemporal features with 3D convolutional neural network," (in English), *Sensors*, vol. 19, no. 11, p. 2472, Jun. 2019.

[39] W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, and A. Wang, "A novel violent video detection scheme based on modified 3D convolutional neural networks," *IEEE Access*, vol. 7, pp. 39172–39179, 2019.

[40] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 1, 2014, pp. 568–576.

[41] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1446–1453.

[42] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[43] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[45] Z. Liu, Y. Chen, B. Chen, L. Zhu, D. Wu, and G. Shen, "Crowd counting method based on convolutional neural network with global density feature," *IEEE Access*, vol. 7, pp. 88789–88798, 2019.

[46] Y. Ding, F. Chen, Y. Zhao, Z. Wu, C. Zhang, and D. Wu, "A stacked multi-connection simple reducing net for brain tumor segmentation," *IEEE Access*, vol. 7, pp. 104011–104024, 2019.

[47] C. Shi, R. Xia, and L. Wang, "A novel multi-branch channel expansion network for garbage image classification," *IEEE Access*, vol. 8, pp. 154436–154452, 2020.

[48] *UMN Dataset, University of Minnesota*. Accessed: Jun. 30, 2021. [Online]. Available: http://mha.cs.umn.edu/movies/crowdactivityall

[49] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Computer Analysis of Images and Patterns*. Berlin, Germany: Springer, 2011, pp. 332–339.

[50] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 1–6.

ABID MEHMOOD received the M.Sc. degree in computer science from Quaid-i-Azam University, Islamabad, Pakistan, in 2001, and the Ph.D. degree in computer science from Universiti Teknologi Malaysia, Johor Bahru, Malaysia, in 2014. Prior to entering academia, from 2001 to 2009, he worked at the software development industry in different roles and contributed to the design and development of various high-performance enterprise applications. He is currently an Assistant Professor with the Department of Management Information Systems, King Faisal University, Saudi Arabia. His research interests include neural networks and deep learning, the Internet of Things, dynamic and self-adaptive systems, model-driven engineering, and aspect-orientation.