Video-based Domain Generalization for Abnormal Event and Behavior Detection

Salma Kammoun Jarraya, Alaa Atallah Almazroey
Computer Science Department, Faculty of Computing and Information Technology
King Abdulaziz University, KSA

***Abstract*—Surveillance cameras have been widely deployed in public and private areas in recent years to enhance security and ensure public safety, necessitating the monitoring of unforeseen incidents and behaviors. An intelligent automated system is essential for detecting anomalies in video scenes to save the time and cost associated with manual detection by laborers monitoring displays. This study introduces a deep learning method to identify abnormal events and behaviors in surveillance footage of crowded areas, utilizing a scene-based domain generalization strategy. By utilizing the keyframe selection approach, keyframes containing relevant information are extracted from video frames. The chosen keyframes are utilized to create a spatio-temporal entropy template that reflects the motion area. The acquired template is then fed into the pre-trained AlexNet network to extract high-level features. The study utilizes the Relieff feature selection approach to choose suitable features, which are then served as input to Support Vector Machine (SVM) classifier. The model is assessed using six available datasets and two datasets built in this research, containing videos of normal and abnormal events and behaviors. The study found that the proposed method, utilizing domain generalization, surpassed state-of-arts methods in terms of detection accuracy, achieving a range from 87.5% to 100%. It also demonstrated the model's effectiveness in detecting anomalies from various domains with an accuracy rate of 97.13%.**

***Keywords*—Domain generalization; abnormal event; abnormal behavior**

I. Introduction

Currently, improvements in technology and decreased costs for surveillance cameras have led to a rise in their utilization in public and private settings. Moreover, the need for an automated monitoring system is increasing because of heightened safety and security issues. An initial method for identifying irregularities from a surveillance camera was a non-intelligent approach where numerous displays were constantly monitored and checked, mainly by human operators. This activity was deemed urgent, demanding a high level of attention, as anomalies in video scenes are few compared to regular operations.

Developing an intelligent system is in high demands to detect anomalies and achieve the necessary outcomes automatically. The automatic system helps human operators detect abnormal events and behaviors and respond accordingly. Recent works focus on identifying anomalies in videos without using explicit models. Anomalies in video settings are usually infrequent and unpredictable, posing a challenge in training a model to encompass all possible domains of abnormal events and behaviors. Many limitations are associated with current anomaly detection systems, often developed using a manual methodology tailored to a given dataset to identify a particular anomaly. These methods encounter challenges when applied to new context and conditions due to the unique biases included in each scene.

Generating a detection model to identify abnormal events and behaviors in crowd scenarios is important for saving time, minimizing operator involvement, enhancing public safety, and verifying the model's ability to find abnormalities not previously recognized in various domains. Luo et al. [1] introduced a Future Frame Prediction Network for Video Anomaly Detection using deep learning methods to anticipate unusual video occurrences. Their approach showed strong performance in accurately identifying anomalous events, indicating potential research paths to improve generalization in new environments. Bhuiyan et al. [2] reviewed video analytics utilizing deep learning for crowd analysis, emphasizing the growing need for thorough techniques in video surveillance to identify abnormal events. This is the foundation for motivating the application of domain generalization in deep learning. This also involves utilizing transfer learning, which extends beyond individual activities and domains.

This work introduces a supervised deep-learning method with domain generalization to identify aberrant events and behaviors in crowd scenes. The research provides a thorough evaluation that focus on domain generalization and employs cross-domain transfer learning from the source domain to the target domain.

The remaining sections of the paper are structured as follows: In Section II, we provide a brief background on domain generalization and anomalies in video scenes, along with a literature review of previous works in these areas. Section III outlines our proposed offline method for generating anomaly detection models. Section IV, Experiments and Results, presents and discusses various experiments conducted to validate the techniques utilized in our proposed method and assess its contributions. Finally, in Section V, we conclude with a summary of our findings and recommendations for future research in detecting anomalies in video scenes.

II. Background and Literature Reviews

This section is divided into two subsections. In subsection A, we introduce the principle of domain generalization and review related works that apply domain generalization techniques. Subsection B describes anomaly detection and its various techniques, including an overview of existing works in this area.

A. Domain Generalization

Domain generalization (DG) is a recent study area within computer vision. Domain generalization can transmit information from the source domain to the target domain, referred to 'unseen domain'. The source domain pertains to the dataset used for training, while the target domain pertains to the dataset used for testing. However, in many visual applications, there are situations when there is labeled training data in one domain and unlabeled data in another. An optimal learning system should capture the broad concept of the visual world from limited accessible samples to prevent bias towards a specific domain. A model's performance is negatively affected when evaluated on a different domain due to domain discrepancy, viewpoint alteration, and changes in illumination. Furthermore, DG uses inexpensive data sources because of the unavailability and challenges in obtaining target domain data. These datasets reflect closely related but distinct tasks. The system attempts to learn by combining data from different source domains to create less sensitive visual classifiers for the target data. Domain generalization can be better comprehended with the provided example in Fig. 1.

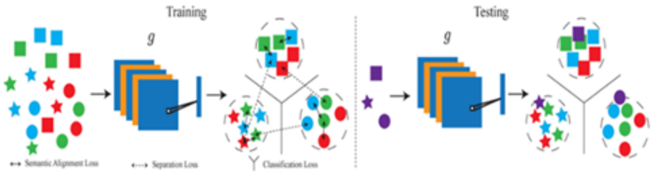

Fig. 1. Different datasets aggregated into single dataset and classifier is trained to classify the unseen test data [3].

Fig. 1 illustrates the integration of labeled data from many domains into a single dataset for training, resulting in the creation of less sensitive visual classifiers for the target data. Various domain distributions are depicted using distinct colors, representing each class by a unique shape. Following training, the model is evaluated using target data from a distinct domain that is not part of the training process.

Blanchard et al. [4] first defined the issue of DG. The authors introduced a kernel-based classifier inspired by multi-task learning, which theoretically ensured the performance across many related domains. Their proposed method efficiently deals with automatic flow cytometry gating. The Domain-Invariant Component Analysis (DICA) algorithm was introduced as a feature-learning algorithm utilizing the kernel in [8]. DICA is an extension of Kernel PCA that reduces the discrepancy between several source domains while preserving the functional connections with the feature label. This allows it to acquire a consistent transformation that applies across many domains.

DG has garnered interest in visual applications, including image-based analysis, object recognition [5], face spoofing [6], and activity recognition [7]. Dataset bias or domain shift poses a challenging problem in object recognition and must be resolved quickly. Any restricted collection of photographs is likely to only capture some facets of the subject because of the intricate nature of the visual realm. Thus, [5] introduced a Denoising Multi-Task Auto-encoder (D-MTAE) to extract domain-invariant features from pre-trained deep learning networks for object identification. This is achieved by learning feature representation across different domains using labels to establish connections. The classification accuracies were evaluated using multi-class SVM with the linear kernel (L-SVM). This method was applied in object identification and achieved an average classification accuracy rate of 68.60%.

The approach suggested in [7] utilizes Adversarial Auto-Encoders (AAE) to learn a feature representation through the joint optimization of a multi-domain auto-encoder, which is regularized by the Maximum Mean Discrepancy (MMD) distance. Employing the Adversarial Autoencoder (AAE) for feature learning decreases the likelihood of the model becoming overfitted to the source domains. It enhances the generalization of acquired features to unfamiliar target domains. Furthermore, a new classifier layer is appended to the acquired features to facilitate categorization. AAE-MMD is utilized in various visual applications such as handwritten recognition, action recognition, and object recognition, achieving average accuracy of 89.8%, 91.9% , and 72.3% for each application, respectively.

Moreover, DG is utilized to enhance the efficiency of biometric identification, specifically in face spoofing scenarios. In [6], a 3D CNN network extracts essential spatial and temporal characteristics. The model has utilized a generalization technique by reducing the MMD distance between several domains to ensure its ability to detect any abnormal event in an unobserved target domain. In addition, an open cross-domain visual search was created by [9] and implemented in a free-hand sketch program. This refers to searching for pairs of target and source domains. Carlucci et al. [10] created an unsupervised method for solving jigsaw puzzles. The method involves reconstructing the original image from its scrambled pieces and understanding spatial similarity concepts for classification purposes. Starting with photos from several domains, each image was divided into nine patches; an index labeled each patch and then randomly rearranged. Subsequently, the curated and randomized images were fed into a convolutional network. Two classifiers are employed: a jigsaw classifier based on a patch index and an object classifier based on an object label.

A domain flow generation model (DLOW) [11] proposed a method to aggregate two distinct domains by producing a continuous sequence of intermediary domains flowing from the source domain to the target domain. The primary advantage of the DLOW model is its ability to handle two scenarios. Initially, source images in intermediate domains are transformed into distinct styles. The gap between the source domain and the target domain is reduced by transferring photographs. Additionally, the DLOW model can produce novel image styles by training on numerous target domains not present in the training data—implementation of the DLOW model using Cycle GAN for unpaired image-to-image conversion.

Domain generalization has been used exclusively for image-based analytical tasks such as action identification, object recognition, and handwritten digit recognition, as shown in Table I. Thus, due to their video-based nature, domain generalization has yet to be utilized for identifying anomalous occurrences or behaviors. Anomaly is synonymous with abnormality, deviation from the ordinary, or something that appears strange and unexpected. The anomaly in the video scene refers to an action or activity that deviates from the film's context. It can be categorized into two forms, as seen in Fig. 2.

TABLE I. SUMMARY OF THE EXISTING METHODS APPLIED THE DOMAIN GENERALIZATION

Ref. year	Approach	Dataset	AUC	EER	Application
[8]	Invariant Feature Representation	GvHD	94.16	-	marrow transplantation
[5]	Feature learning approach	VLCS Office + Caltech dataset MINIST	68.60 86.29 89.8	- - -	Object recognition Handwritten digit recognition
[7]	Generative Adversal Network (GAN)	IXMAS VLCS Idiap	91.9 72.3 -	- - 0.3	Action recognition Object recognition
[6]	Spaito-Temporal approach	CASIA MSU Rose-Youtu	- - -	1.4 0.0 7.0	Face spoofing detection
[9] 2019	ConvNet	-	-	-	free-hand sketch
[10]	CNN	PACS VLCS Office-Home	80.51 73.19 61.20	- - -	Jigsaw puzzle
[11]	GAN	Van Gogh Van Gogh + Ukiyo-e	- -	- -	Image translation

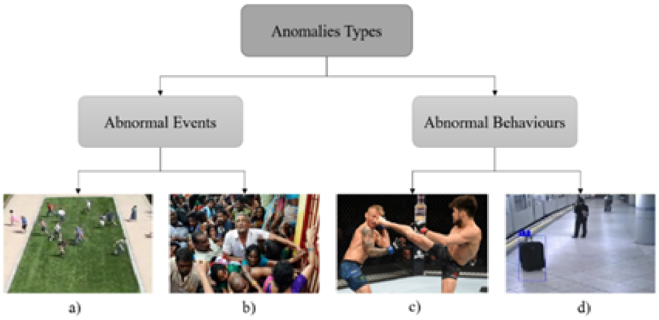

Fig. 2. Different types of anomalies: Abnormal Events (a) Escape and b) Stampedes), and Abnormal Behaviors (c) Fighting and d) Abandoned baggage)

B. Anomaly Detection based on Deep Learning Approaches

Abnormal event is an occurrence influenced by external factors, such as escape due to natural calamities like earthquakes or floods or induced by abnormal conduct like fighting [12]. Abnormal conduct refers to actions or attitudes displayed by an individual or group that deviate from the usual, such as throwing objects [13], walking, or driving in the incorrect direction. Abnormal events and behavior detection involve identifying and reacting to unusual video alterations. Researchers have been investigating methods to create an effective model for correctly detecting anomalies in video scenes.

Anomaly detection is a method used to identify uncommon objects or unexpected motion in video footage. There are two methods for detecting anomalies in videos: a hand-crafted approach and a deep-learning approach. Hand-crafted representation is an initial method to identify video scene anomalies. This method involves extracting information from the input video, necessitating an expert to create a model tailored to these qualities. Deep learning (DL) is a technique that utilizes the hierarchical structure of Artificial Neural Networks (ANNs) to perform machine learning. Its design is influenced by the human brain's operations known as artificial neural networks. The hand-crafted approach could be more satisfactory because it relies on extracted features tailored to detect a particular abnormality in a specific context. Therefore, this study emphasizes the utilization of a deep learning approach.

CNN has been utilized as a potent method for detecting anomalies in crowded scenes due to its effectiveness with high-dimensional data. A novel foreground object localization method is introduced [14]. This method extracts motion features using a Spatially Localized Multiscale Histogram of Optical Flow (SL-MHOF) and appearance features using a CNN-based model, eliminating the need to divide the video into multiple patches for fusion. Next, include the merged characteristics into a Gaussian Mixture Model (GMM) classifier for anomaly detection. Zhou and et al. [15] utilized a FightNet model to identify visual interactions using Temporal Segment Networks (TSN). Thus, FightNet was trained using three distinct input types: RGB, optical flow, and acceleration images for spatial and temporal networks. Subsequently, merge the outcomes acquired from various inputs to categorize the video. Song et al. [16] improved the methodology presented in [15] by incorporating 3D convolution and 3D pooling with a keyframe extraction approach to enhance the extracted features. Video frames are segmented into clips using keyframes to eliminate redundant frames and emphasize the movement between frames. CNN necessitates a substantial quantity of training films to prevent overfitting. Sabokrou et al. [17] were the first to employ fully convolutional neural networks (FCN) to address the limitations of CNN. Using a pre-trained CNN model decreases computational expenses by utilizing original frames as input rather than dividing the frame into smaller patches. Furthermore, a pre-trained Convolutional Neural Network (CNN) and optical flow are inputted into a Fully Convolutional Network (FCN). This method resulted in aberrant events being detected three times faster than merely a regular CNN.

The novel transfer learning strategy suggested in [18] detects violence by calculating the optical flows of the input video through the Lucas-Kanade method mentioned in [19]. Next, utilize the (OF) values to create many templates, which will serve as input for a pre-trained CNN to extract profound characteristics. A two-stream FCN network was proposed in [20]. The initial FCN stream processes the original frame input to extract appearance features, while the second stream utilizes optical flow to obtain motion features from the video frames. The combination of these features results in convolutional features. Binarize the convolution features using binary coding to aid in calculating the anomalous coefficient. The study referenced in [21] utilized a weakly supervised learning method to categorize videos as either 'normal' or 'abnormal' without pinpointing the exact frame where anomalies arise in abnormal videos. A pre-trained model that utilizes C3D to learn features for each segment. A model is trained to rate anomalies, predicting high scores for aberrant video portions. The study in [22] introduced fine-tuned CNN architectures using Aggregation of Ensembles (AOE), incorporating pre-trained CNNs such as AlexNet, VGGNet, and GoogleNet, each specializing in learning distinct features. Subsequently, different classifiers are employed to achieve the most favorable outcome for classification.

Subsequently, researchers integrated CNN with a long short-term memory (LSTM) network to extract spatial and temporal characteristics. Morales et al. [23] introduced a model for identifying violent robberies in Closed-Circuit Television Videos (CCTV) by utilizing a pre-trained VGG-16 network to extract characteristics, which were subsequently inputted into two convolutional long-short-term memory (convLSTM) layers. Finally, provide geographical and temporal characteristics to a fully connected layer group to obtain the categorization outcome. The technique mentioned in [24] involves preprocessing input frames by eliminating adjacent frames. The resulting data is fed into a pre-trained Alexnet model to extract spatial information. The study in [12] improved upon the technique introduced in [24] by introducing a Bidirectional Convolutional LSTM (BiConvLSTM) network. By utilizing a pre-trained network to extract appearance features and feeding them into the BiConvLSTM to capture temporal information bidirectionally for long-range context access, a more comprehensive understanding of the entire video is achieved, resulting in improved classification.

The study reviewed prior works in Table II and found that utilizing a deep learning approach for anomaly identification in a single dataset yields high detection accuracy. However, the approaches mentioned are specifically created to identify abnormal events or behavior in a given setting, but not simultaneously.

Various successful approaches in anomaly detection have been proposed, as summarized. Limitations are present in the methodologies outlined in this section. Domain generalization techniques have mainly been used in image-based analysis and have not been applied in video analysis models. Current anomaly detection approaches usually concentrate on identifying unusual occurrences or behavior in a video scene rather than both simultaneously, even though there may be numerous abnormalities in the video data.

This research seeks to overcome these limitation by using a supervised deep-learning method with domain generalization. We propose a comprehensive model to identify abnormal events and behavior in various domains. Furthermore, transfer learning will be used, its proven efficacy when combined with current methods.

III. Proposed Offline Method to Generate Anomaly Detection Model

This section elaborates on the proposed method, which utilizes a supervised deep learning methodology with domain generalization to identify aberrant events and behaviors in crowd video situations. The suggested method consists of two steps, as seen in Fig. 3. The initial phase involves pre-processing, commencing with transforming input films into a series of frames. Next, the keyframe selection method is applied to video frames using the Cosine Similarity (CS) algorithm [25] by measuring the similarity between two frames (current frame and prior keyframe). Next, compare the acquired result with the similarity threshold value to ascertain if the frame qualifies as a keyframe. Only the chosen keyframes are forwarded to the subsequent stage to create a spatio-temporal entropy template that emphasizes temporal and spatial variations among keyframes. The result from the initial stage, a spatio-temporal entropy template, is utilized as input for the subsequent step to derive deep features through the Convolutional Neural Network (CNN). The Relieff features selection method [18] is utilized to obtain impactful characteristics for accurately detecting anomalies. Various classifiers were tested for video classification, and the study chose the one that yielded superior classification outcomes.

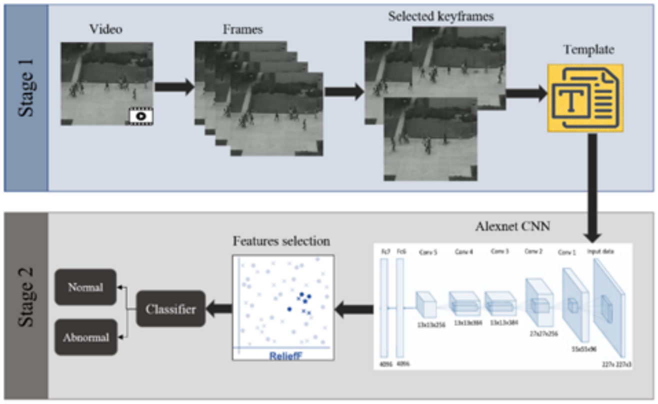

Fig. 3. The two stages of the proposed method.

The rest of this section is organized as follows: first, we introduce the preprocessing stage for the keyframe selection method and the process of generating a spatiotemporal entropy template. Then, we represent the feature extraction, feature selection method, and model generating, respectively. To make this section easy to read, some details and justifications related to each step of the proposed method are well described and validated in section IV.

The remainder of this section is structured as follows: We first provide the pre-processing stage for the keyframe selection approach and the procedure for creating a spatio-temporal entropy template. Next, we will present the feature extraction, feature selection approach, and model generation. Comprehensive explanations and validations for each step of the proposed technique are provided in section IV to enhance readability.

A. Pre-processing Stage

Pre-processing is the initial phase of the proposed approach. They first turned all input videos into individual frames. Subsequently, the keyframe selection technique can choose only frames with novel data. The chosen keyframes are utilized to create a spatio-temporal entropy template, which

TABLE II. Description of the Existing Supervised Deep Learning Methods for Detecting Anomalies from Video Scenes

Ref.	Deep Architecture	Features	Input data	Dataset	Anomaly Measurement		Abnormal Type
					AUC	EER	
[17]	Fully convolutional neural networks	Shape and motion features	Frame	UCSD ped2 Subway Entrance Subway Exit Hockey	- 90.4% 90.2% 94.4%	11% 17% 16% -	Behavior
[18]	CNN	Optical Flow	Frame	Movies ViF	96.5% 90.8%	- -	Behavior
[20]	Two-stream FCN	Spatial and Temporal	Frame	UMN UCSD ped1	97.6% 90.8%	- 15.9%	Event
[14]	(SL-MHOF) + CNN	Appearance and motion	Frame	UCSD ped2 Avenue	97.8% 87.2%	5.9% -	Behavior
[22]	Aggregation of Ensembles (AOE)	Appearance, motion feature	Frame	UCSD ped1 UCSD ped2 Avenue	94.6% 95.9% 89.3%	- - -	Behavior
[12]	Bidirectional Convolutional LSTM	Spatial and Temporal	Frame	Hockey Movies ViF	98.1% 100% 93.9%	- - -	Behavior
[15]	Deep ConvNets	Spatial and Temporal	Video	Hockey Movies	97.0% 100%	- -	Behavior
[16]	3D convolution	Spatial and Temporal	Frame	Hockey Movies ViF	98.96% 99.97% 93.5%	- - -	Behavior
[24]	CNN	Spatial and Temporal	Frame	Hockey Movies ViF	97.1% 100% 94.6%	0.55% 0% 2.34%	Behavior

is subsequently provided as an input to the second phase of the proposed technique. Keyframes are frames in a video that provide a comprehensive summary of the entire video and can be extracted to remove nearby repetitive frames effectively. Keyframe selection is the process of choosing frames that include new information [25]. The keyframe selection process aims to summarize the video by eliminating redundant adjacent frames to decrease the amount of information to be processed and reduce computational complexity [26]. Cosine resemblance (CS) quantifies the resemblance of video frames based on the cosine value of the frames. This study estimated the CS value for all input frames using equation (1) to establish the suitable similarity threshold [27].

$$Cosine\ Similarity\ (CS) = \frac{\sum_{i=1}^{n} A_i\ B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\ \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

Where A refers to the current frame and B refers to the next frame, and n states number of frames. Closer CS value to 1 means lower differences between the two frames [26].

Fig. 4 and Fig. 5 display a selection of 60 movies, comprising 30 regular videos and 30 abnormal videos from each dataset utilized in this research. The contrast score between each pair of successive frames is determined, and then the mean contrast score for each video is calculated. The line chart in Fig. 4) displays the average CS values for each standard sample video, ranging from 0.94 to 0.99. In contrast, Fig. 5 shows the average CS values for abnormal movies, ranging between 0.91 and 0.99. Most of the CS values fall within the range of 0.90 to 1.

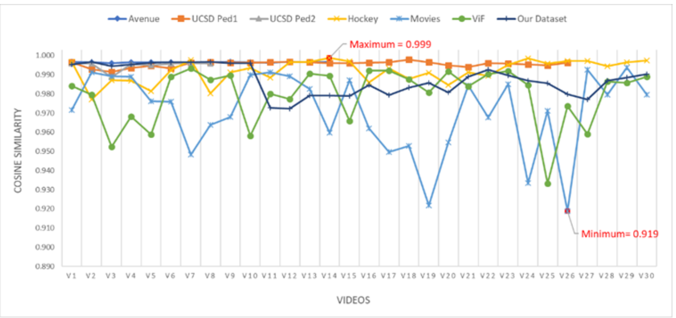

Fig. 4. Average cosine similarity values for normal videos.

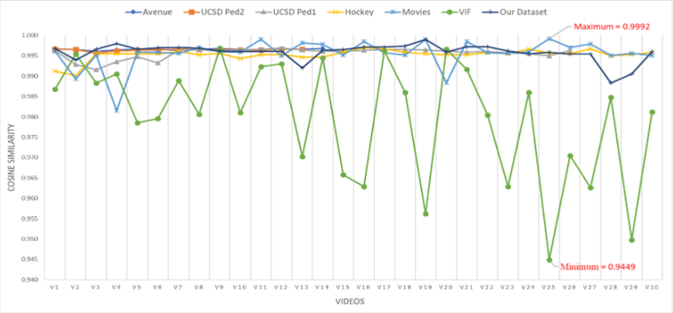

Fig. 5. Average cosine similarity values for abnormal videos.

The keyframe extraction method utilizes the CS algorithm [26] to identify keyframes from video frames by assessing the similarity between two frames. The process for extracting keyframes is illustrated in a flowchart in Fig. 6. This algorithm takes video frames as input and begins by verifying if the current frame is the first in the sequence. If the frame is the first keyframe, it is saved in a buffer. If it is not the first frame, the CS algorithm calculates the differences between the current frame and the previously extracted keyframe. If the CS value obtained does not surpass the similarity threshold value, it indicates that the two frames are different, and the current frame is then considered the new keyframe. The algorithm stored the keyframe in the buffer and utilized it to retrieve the subsequent keyframe. A higher cosine value suggests a similarity between the two frames. A lower cosine value implies a variation between the two frames. CS represents

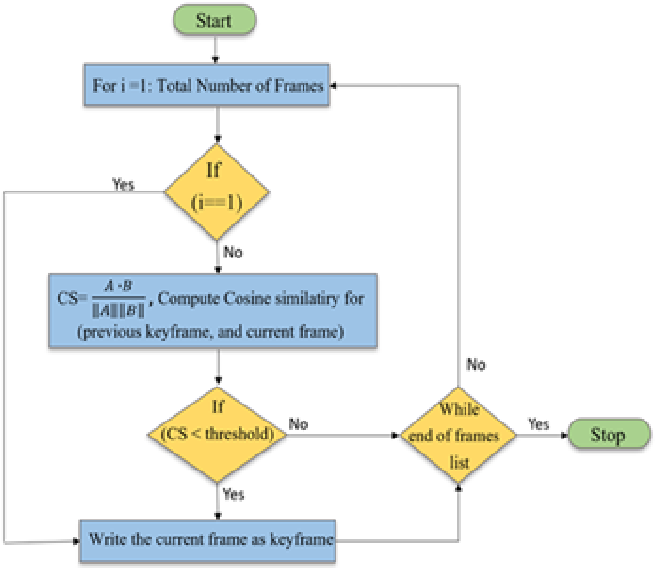
Start
For i =1: Total Number of Frames
If
(i==1)
Yes
No
CS= $\frac{A \cdot B}{\|A\|\|B\|}$, Compute Cosine similatiry for (previous keyframe, and current frame)
If
(CS < threshold)
No
Yes
While
end of frames
list
No
Yes
Stop
Write the current frame as keyframe

Fig. 6. Keyframe selection method.

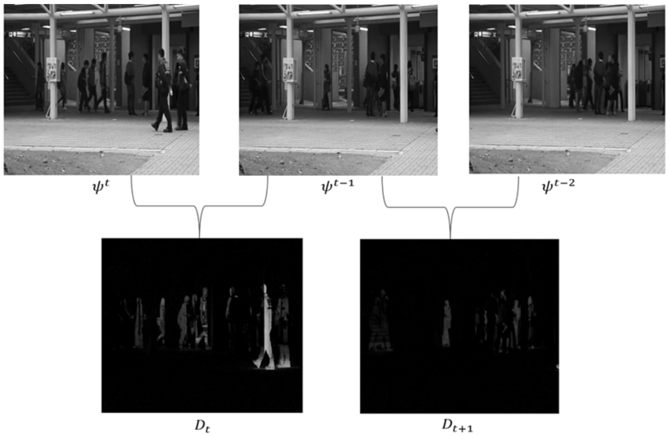

Fig. 7. Represents the frame differences generated by using the three-frame differences method. The top row Shows the original keyframes, and the bottom row shows the created frames differences.

the Cosine Similarity value, whereas I denotes the current frame index. The variable A represents the previously selected keyframe, whereas B represents the current frame. The average CS value obtained in the previous part falls between 0.90 and 1. Therefore, the threshold for comparing it with the CS value to extract keyframes should be within this range. The study established a similarity criterion of 0.995 after conducting multiple trials. Lower values were tested. However, no keyframe was recovered in specific videos. Only the keyframes from this section are forwarded to the next stage for generating the spatio-temporal entropy template.

Shannon (1948) developed entropy as a measure of 'disorder.' Entropy in a picture is a statistical metric of randomness that can describe the texture of the input image. Entropy is a measure used to assess visual information, where the entropy value rises as the unpredictability level increases. The aim of creating a spatio-temporal entropy template in this study is to concentrate the feature extraction process on motion regions rather than all spatio-temporal data. A spatio-temporal entropy template is created in this stage utilizing the selected keyframes from the previous step.

The process of creating a spatio-temporal entropy template involves four steps. Detect the motion region by utilizing the three-frame differences approach to calculate frame differences. I am applying the automatic dynamic threshold value to those difference frames. Create a pixel state card utilizing the state labels approach to determine if the pixel is part of a moving region. Compute the spatio-temporal entropy value for each pixel in the video keyframes. The initial and secondary processes detect motion regions, whereas the final two steps are utilized for modeling the background.

Motion region detection is the capability to recognize the pixels that show the movement of objects between video frames. This study initially utilized the three-frame differences method [28] to identify the motion region from keyframes and detect temporal changes in video keyframes. By choosing three consecutive keyframes (the current keyframe and the two preceding keyframes), the absolute variances between them are computed, resulting in two frame differences, as illustrated in Fig. 7.

The procedure started by converting the colored (RGB) keyframes to greyscale keyframes, then selecting the third keyframe $\left(\psi^t\right)$ from keyframes list and subtract it from the second keyframe $\left(\psi^{t-1}\right)$, and subtract the second keyframe $\left(\psi^{t-1}\right)$ from the first keyframe (ψ^{t-2}), as given by equations (2) and (3) [70]. Where D_t and D_{t+1} represent the frames differences using the three-frame differences method. The top row shows the original keyframes, and the bottom row shows the created frames differences.

$$D_t = \left|\Delta_{Gray}^{\psi^t, \quad \psi^{t-1}}\right| = \left|\psi^t_{Gray} - \ \psi^{t-1}_{Gray}\right| \tag{2}$$

$$D_{t+1} = \left|\Delta_{gray}^{\psi^{t-1}, \quad \psi^{t-2}}\right| = \left|\psi^{t-1}_{Gray} - \ \psi^{t-2}_{gray}\right| \tag{3}$$

Automatic threshold is a method that extracts essential information represented by pixels from the difference frames $(D_t, \ D_{t+1})$ by utilizing a feedback loop to optimize the threshold value. This process is cited in [29]. Automatic threshold effectively reduces background noise. The automatic threshold calculation procedure is depicted in a flowchart (Fig. 8). First, calculate the current threshold value to identify the mid-range pixels in the frame difference (D). Secondly, the binary value of D is determined by comparing its pixel value with the current threshold. The study categorizes pixels with values lower than the current threshold as background pixels and assigns them a value of 0. Pixels with values equal to or greater than the threshold are deemed foreground pixels and allocated 1 (where T represents the current threshold value) [30].

These processes create two images one for the background and another for the foreground. Thus, the mean for each of the two images is determined and used to determine the current threshold by taking the average of those mean values. Lastly, check whether the last threshold value is equal to the current threshold if it is then the loop will be stopped. Otherwise,

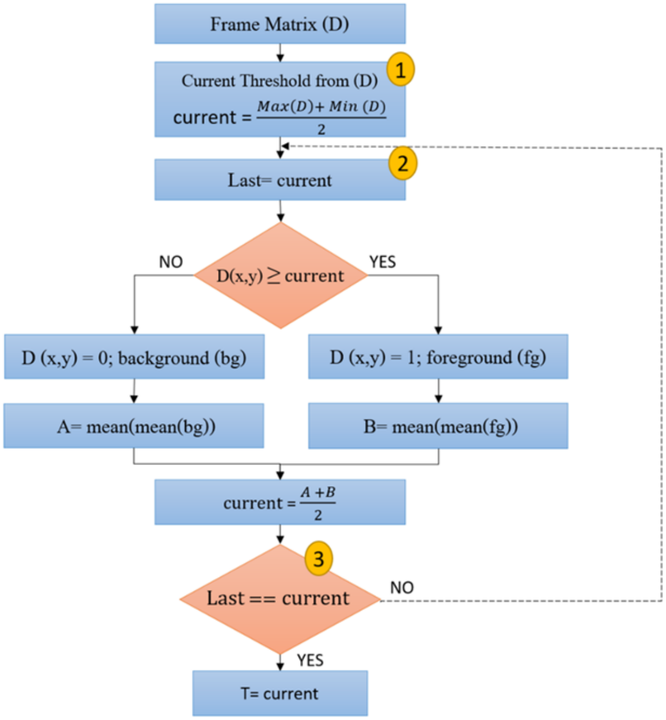
Frame Matrix (D)
Current Threshold from (D)
current = $\frac{Max(D) + Min(D)}{2}$
1
Last= current
2
D(x,y) ≥ current
NO
YES
D (x,y) = 0; background (bg)
D (x,y) = 1; foreground (fg)
A= mean(mean(bg))
B= mean(mean(fg))
current = $\frac{A+B}{2}$
3
Last == current
NO
YES
T= current

Fig. 8. Automatic threshold calculation.

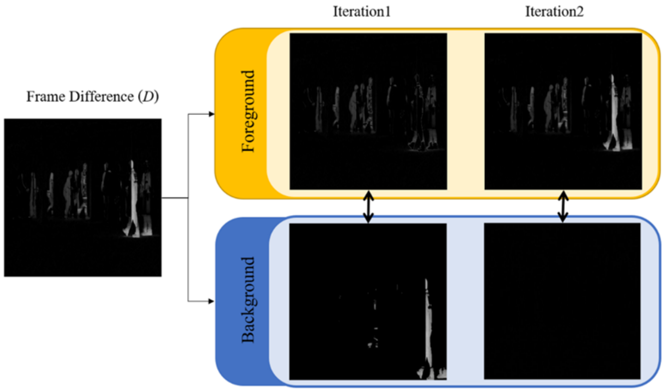

Fig. 9. Automatic threshold computation for each iteration.

then the whole process repeated starting from the second step using the original frame difference (D) and assigning the last threshold as the current threshold. All the classification decision in this procedure is associated with a pixel level, without considering its neighbors.

Fig. 9 displays the obtained images for background and foreground from (D) with different threshold iterations. In Fig. 9 the loop stopped after the second iteration when the new threshold was equal to the initial threshold, where the second iteration shown the perfect separation of background pixels from the foreground pixels.

The obtained threshold value used in the next section, as the following section clarifies building of the pixel state cards required to update the dynamic matrix. Building a pixel state card is used to update a dynamic matrix via a sliding window technique. As the sliding window is a rectangular area of fixed width and height that moves across a keyframe. Furthermore, the use of the sliding window improves decision-making by examining the pixel's neighbors in the sliding window to obtain a spatio-temporal entropy value of each pixel. The state labelling technique used to label the sliding window to determine the spatio-temporal entropy value for each pixel. Pixel labeling technique of frame ψ^t is based on differences of ($\Delta_{Gray}^{t-1,t-2}$ and $\Delta_{Gray}^{t,t-1}$), according to equation (4).

$$\begin{cases} \Delta_{Gray}^{(t-1,t-2)} = \psi_{Gray}^{(t-1)} - \psi_{Gray}^{(t-2)} \\ \Delta_{Gray}^{(t,t-1)} = \psi_{Gray}^{t} - \psi_{Gray}^{(t-1)} \end{cases} \tag{4}$$

The spatio-temporal sliding window (S) for each pixel is defined by Eq. (5) [29].

$$S = \{(i,j)_k \mid |i-x| \prec [w/2], |j-y| \prec [w/2], 0 \preceq t-1 \prec L\} \tag{5}$$

Where w and L are parameters that control the size of the sliding window (S). As $w * w$ refer to the height and width, and L refers to depth of S where $w = L = 3$.

A state-of-the-art technique is used to derive the label of S based on $\Delta_{Gray}^{(t-1,t-2)}$ and $\Delta_{Gray}^{(t,t-1)}$. The state of labels is defined as 0,1,2, with 0 representing no motion, 1 representing little motion, and 2 representing motion [29]. They initially assigned the state label 2 to all pixels in sliding windows L_1. The pixels in sliding windows L_2 and L_3 are allocated labels 0,1,2 based on comparison results with the thresholds.

The state labels within the Spatio-temporal sliding window are utilized to compute the probability density function for each pixel Π_{xy} by assessing the pixel's variation about its neighboring pixels using Eq. (6).

$$P_{(x,y,e)} = H_{(x,y,e)}/N \tag{6}$$

Where:

- N refers to the total number of labels in sliding window (S).
- $H_{(x,y,e)}$ refers to the number of label e in S as $e =$ 0,1,2.

The spatio-temporal entropy of pixel Π_{xy} now can be obtained by the following Eq. (7). Where E refers to spatio-temporal entropy value.

$$E_{(x,y)} = -\sum_{(i=0:2)} P_{(x,y,i)} \tag{7}$$

Calculating the spatio-temporal entropy value has been repeated for every pixel in the keyframe. Each video in the collection was eventually shown using a spatiotemporal entropy template. The created templates are utilized as input for the subsequent stage of the suggested method to extract profound features and create a model, as detailed in the next section.

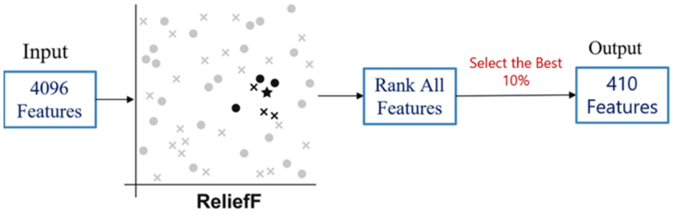

Fig. 10. Represents the process of feature selection.

B. Features Extraction and Model Generation Stage

The second part of the proposed strategy involves feature extraction and model creation. Feature extraction initiates the initial phase of the second stage in the suggested technique. Feature extraction reduces the resources needed to describe a vast dataset. The template created in the previous step is utilized as an input for the pre-trained Convolutional Neural Network (CNN) 'AlexNet' [31] to extract profound characteristics. The advantage of utilizing CNN for feature extraction lies in its simplicity and ease of implementation, making it easily applicable across many monitoring situations. Furthermore, this study utilized a pre-trained Convolutional Neural Network (CNN) due to its ability to perform well with limited training data. The Alexnet network necessitates an input size of 227×227×3, with 3 representing the number of color channels. The Alexnet network design comprises five convolutional layers and three fully connected layers (FC). Dropout regularization at a rate of 50% is implemented between the fully connected layers to mitigate overfitting [32]. The study retrieved features from the 'fc7' layer, resulting in 4096 features.

Feature selection is the process of optimizing retrieved features by choosing those that offer pertinent information for constructing a model efficiently. The study utilized the Relieff feature selection approach with k-nearest neighbors [18], where the input consists of the extracted features and labels vector. The output consists of the index of features ranked by the distinctiveness of their weight. The weight values of the features vary from -1 to 1, with significant positive weights indicating the feature's relevance. Feature selection benefits include reducing dimensionality, enhancing classification speed efficiency, and improving prediction performance. Only the top 10% of the total features, which amounts to 410 out of 4096, are chosen for creating the anomaly detection model, as shown in Fig. 10.

Model creation involves developing a model using the retrieved features from the training dataset. The study used a cross-validation technique to assess the model's effectiveness. The training films are randomly divided into five folds using five-fold cross-validation on the training dataset. Four folds are used for training the model, while the remaining fold is used to assess the model's effectiveness. Cross-validation prevents the creation of an overfitting model tailored to a specific dataset. Additionally, cross-validation is beneficial when used with a small dataset. Various classifiers have been utilized in the training folds to create a model. The experiment chose a linear Support Vector Machine (SVM) classifier with an 'auto' kernel scale and a Sequential Minimal Optimization 'SMO' solver. The linear SVM achieved superior accuracy results compared to other classifiers.

In the following section, we will examine and discuss the outcomes of the suggested model using various datasets. The study includes multiple experiments to assess the efficiency and performance of the proposed method.

IV. Experiments and Results

This section represents the experimentation and validation conditions and presents a discussion of the experimental results in order to evaluate all the used techniques and to evaluate the contributions to this research. Three different datasets are used in this study categorized according to the intent of use. The first dataset is the public datasets for anomaly detection, in this study six different public datasets were selected: UCSD Ped1 and UCSD Ped2 datasets [33], Avenue dataset [34], Hockey Fight dataset [35], Movies dataset [35], and Violent-Flows Crowd dataset (ViF) [36]. Sample of these video frames shown in Fig. 11.

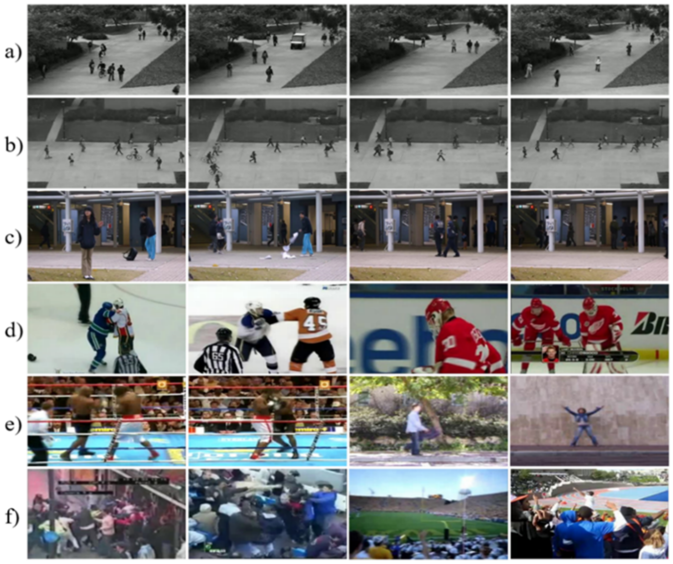

Fig. 11. Samples of public datasets: a) UCSD Ped1, b) UCSD Ped2, c) Avenue, d) Hockey, e) Movies, and f) ViF. The first two columns present abnormal frames, and the last two columns present normal frames.

Previous public datasets are limited since they only include abnormal behaviors. We have created a new dataset named the 'Collected Dataset',comprising 1654 movies categorized as normal and abnormal, as illustrated in Fig. 12. The movies compiled comprise atypical events such as panic induced by natural disasters like earthquakes and fires in vehicles and motorcycles, as outlined in Table III. This study specifically chose films from YouTube that were recorded by closed-circuit television (CCTV) cameras. They combine the public dataset with the gathered videos to create a comprehensive dataset that includes abnormal events and behaviors.

The Validation Dataset (unseen dataset) is the second constructed dataset in this study. It includes a collection of normal and abnormal events and behaviors videos that have been collected from YouTube. The Validation Dataset contains 89 videos, of which 44 videos of normal and abnormal events (fire and panic) that are a mixture of 26 fire videos and 18 panic

TABLE III. PUBLIC DATASETS DESCRIPTION

Dataset Name	Anomalies Type	The Scene	Level of Density	Challenges
UCSD Ped1 UCSD Ped2	Walking	Outdoor	Ranging from Sparse to Crowd	Complex occlusions and Crowd density.
Avenue	Walking, running, throwing an object.	Outdoor	Crowd	Camera shakes.
Hockey	Fighting	Indoor	Non-crowed	Adjacent frames contain overlap information.
Movies	Fighting	Indoor and Outdoor	Ranging from Sparse to Crowd	The resolution of videos frames is different.
ViF	Fighting	Outdoor	Crowd	Extreme crowd

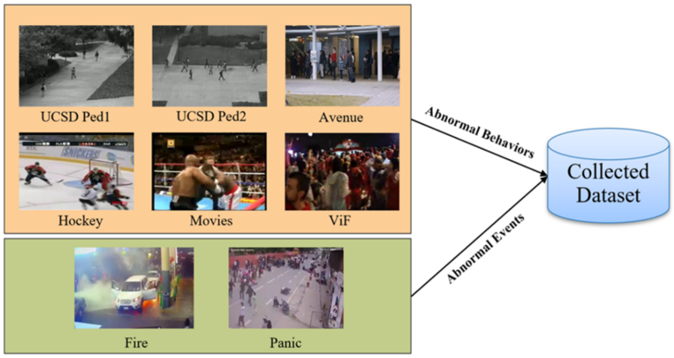

Fig. 12. The collected dataset.

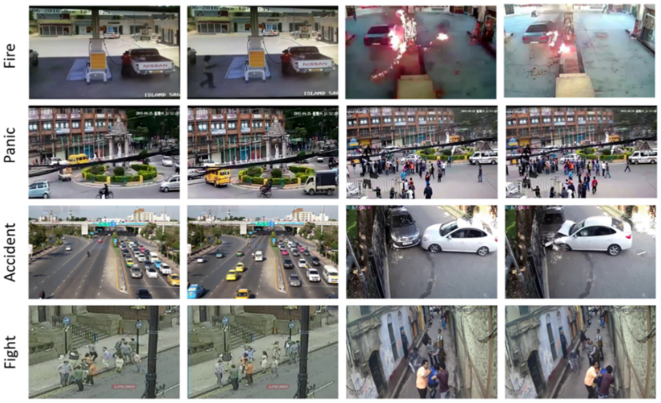

Fig. 13. The validation dataset sample frames: The first two columns show normal video frames, and the last two columns show abnormal video frames

videos. Whereas there are 45 normal and abnormal behavior videos (accident and fighting), with 18 accident videos and 27 fighting videos. In the selected video scenes, the density level varies from sparse to crowd, and their resolution is different. The purpose of creating the Validation Dataset (unseen dataset) is to evaluate the generality of the proposed model for the detection anomalies from unseen domains. A sample of video frames is shown for each abnormal event and behavior (Fig. 13).

In this research, the preparation of the dataset is the primary step of the proposed method by applying the keyframe selection method to all datasets. Selecting only the essential frames containing information from each video and discarding redundant frames to reduce computational complexity. (Table IV and Table V) show the average number of keyframes selected for each video from the Collected Dataset and the Validation Dataset, respectively.

As shown in (Table IV), the Collected Dataset combined six public datasets containing 1581 normal and abnormal behavior videos with 73 videos collected in this study. Furthermore, the study found that the number of frames extracted was significantly reduced in all datasets. As in the Collected Dataset, the average number of frames decreased by approximately two and a half times when the keyframe selection method was applied, thus minimizing the required computational complexity.

The study has also applied the keyframe selection method to all videos in the Validation Dataset. Table V presents the average number of frames and the average of the extracted keyframes for each normal/abnormal event and behavior videos. The average number of frames in the Validation Dataset is 202 frames. Where on average 74 of these frames have been extracted as keyframes that means the keyframe selection method reduced the required time for a process by about two and a half times.

It should be noted that after this preparation, each video in the public datasets, the Collected Dataset, and the Validation Dataset contains different number of keyframes. In this work, the performance of experiments results compared with previous works using well-known evaluation metrics as follows: Accuracy (ACC), Equal Error Rate (EER), Recall, Precision, F1-score, and Area Under the ROC Curve (AUC). Several experiments provided in this section to examine the research choices of the techniques used in the proposed method and to assess the contribution of the proposed method. These experiments will be structured as follows in this research:

- Experiment 1: Validate the research choices for the techniques used in the proposed method.
  - Experiment 1.1: Evaluate the keyframe selection method vs. all video frames.
  - Experiment 1.2: Evaluate the efficiency of using a spatio-temporal entropy template vs. an optical flow template.
  - Experiment 1.3: Evaluate the efficiency of the extracted features using different pre-trained networks.
  - Experiment 1.4: Evaluate the Relieff feature selection method with different sets of features.
  - Experiment 1.5: Evaluate the efficiency of the selected classifier.
- Experiment 2: Validate the contribution of the proposed method.
  - Experiment 2.1: Comparison with state-of-the-art methods.
  - Experiment 2.2: Validate the performance of the domain generalization in video based.

TABLE IV. Datasets Preparation: Public Datasets, Collected Videos, and the Collected Dataset

Dataset Name	No. of videos	Average Frames	Average Keyframes	Type of Anomaly
UCSD Ped1	70	200	93	Behavior
UCSD Ped2	28	163	57	Behavior
Avenue	37	180	76	Behavior
Hockey	1000	41	33	Behavior
Movies	200	50	16	Behavior
ViF	246	89	54	Behavior
Collected videos (panic and fire)	73	395	156	Events
Collected Dataset	**1654**	**195**	**83**	**Events and Behaviors**

TABLE V. Preparations of the Validation Dataset

Videos	Number of Videos	Average Frames	Average Keyframes	Type of Anomaly
Fire	26	182	53	Events
Panic	18	121	76	Events
Accident	18	240	106	Behavior
Fighting	27	245	65	Behavior
Validation Dataset	**89**	**202**	**74**	**Events and Behaviors**

    - Experiment 2.2.1: Evaluate cross-dataset performance without domain generalization.
    - Experiment 2.2.2: Evaluate the Performance of the domain generalization by applying cross-domains.
- Experiment 2.3: Validate the proposed model with domain generalization for detecting abnormal events and behaviors from crowd video scenes.
    - Experiment 2.3.1: Evaluate the efficiency of the proposed model using the Validation Dataset.
    - Experiment 2.3.2: Evaluate the proposed model with domain generalization compared to state-of-the-art methods.

A. Experiment 1: Validation of the Techniques used in the Proposed Method

The proposed method used five different techniques, which is the keyframe selection, generating a spatio-temporal entropy template, feature extraction using a pre-trained model, feature selection, and finally generating model using a classifier.

1) Experiment 1.1: Evaluate the Keyframe Selection Method Vs. all Video Frames: In this experiment, a comparison was conducted using the proposed method with and without the keyframe selection method. The aim of this experiment is to validate the use of the keyframe extraction method in terms of the required time of classification for each video and the accuracy of detection. This experiment was done on the Validation Dataset which consists of 89 normal and abnormal videos. The column charts in (Fig. 14 and Fig. 15) present the number of frames for each normal and abnormal video with and without using the keyframe selection method, respectively.

From these charts, it had been realized that using the keyframe selection method significantly reduces the number of frames that need to be processed. As the average number of frames in the Validation Dataset for normal and abnormal videos is about 202 frames, while the average number of keyframes extracted in the Validation Dataset is about 74 keyframes as illustrated. A comparison had been implemented on the model using the selected keyframes and all video frames based on two criteria: (1) The execution time for classification and (2) The accuracy of detecting anomaly.

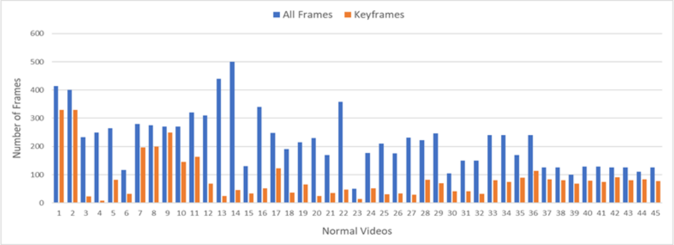

Fig. 14. Number of frames for each normal video with and without keyframe selection method.

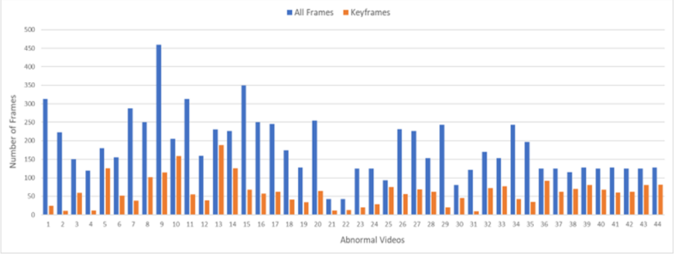

Fig. 15. Number of frames for each abnormal video with and without keyframe selection method.

Execution Time for Classification with and without using the Keyframe Selection Method.

Execution time is also known as the processing time that starts from receiving video keyframes until the video is classified as normal or abnormal. The execution time had been computed for each video in the Validation Dataset with and without keyframe selection method to estimate the required time for classifying a video. The process for calculating the execution time for each video consists of two stages: 1) The duration of generating a template and 2) The duration of extracting features and model classification. Then these durations had been accumulated to get the execution time for each video. The line chart in (Fig. 16 and Fig. 17) present the

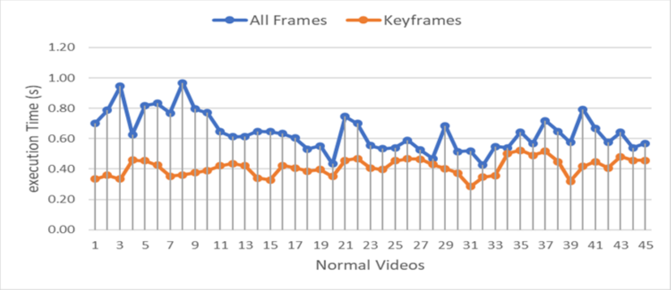

Fig. 16. The execution time for each normal video with and without keyframe selection method.

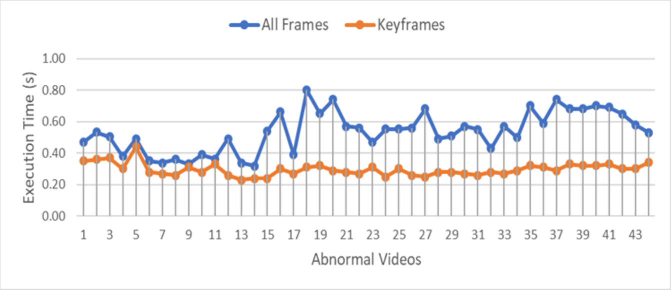

Fig. 17. The execution time for each abnormal video with and without keyframes selection method.

execution time with and without keyframe selection method for each normal and abnormal video in the Validation Dataset, respectively. This experiment showed that the average number of frames in the Validation Dataset using all video frames is 202 frames with an average duration of 10 seconds, which required on average (0.59 milliseconds) for classification. While using the selected keyframes, the number of frames reduced to an average of 74 keyframes, which required on average (0.35 milliseconds) for classification. Reducing the classification time with the use of keyframes is due to a decrease in the number of frames to be processed. As the number of frames has decreased by about three times compared to all frames, which has led to a decrease in the time required to generate the template. As the average time needed to generate a template using all frames is (0.35 milliseconds) while generating a template using the selected keyframes took only (0.15 milliseconds).

Generally, this experiment showed that the time required to classify a video using the keyframe selection method is approximately two times faster than using all video frames. Since the keyframe selection method discards redundant frames that need to be processed.

The Accuracy for Detecting Anomaly from Video with and without the Keyframe Selection Method.

After the study has shown that using the keyframe selection method to classify video is faster than using all video frames. In this section, the objective is to demonstrate the efficiency of the use of selected keyframes to detect anomaly perfectly. This experiment tested with all the datasets used in this research using the keyframe selection method and without it, as shown in (Table VI). It found that some of the public datasets provided the same accuracy with and without using the keyframe selection method, such as the Avenue dataset and the UCSD Ped2 dataset, where 87.5% and 100% accuracy obtained, respectively. Whereas, the rest of the datasets gave better detection when using the keyframe selection method.

TABLE VI. Accuracy for All Datasets Using the Selected Keyframes and All Video Frames

Dataset Name	Accuracy (%)	
Frames	**Keyframes**	**All Frames**
Avenue	**87.5**	**87.5**
UCSD Ped1	**95.24**	87.5
UCSD Ped2	**100**	**100**
Hockey	**98.67**	96
Movies	**100**	97
ViF	**97.3**	83.6
Collected Dataset	**97.13**	88.1

TABLE VII. Comparing the Accuracy of Anomaly Detection by Using Spatio-Temporal Entropy and Optical Flow Templates

Dataset Name	Accuracy (%)		
Technique	**Entropy Template**		**OF Templates [18]**
Used Frames	**Keyframes**	**All Frames**	**All Frames**
Hockey	**98.67**	96	94.4
Movies	**100**	97	96.5
ViF	**97.3**	83.6	80.9

To conclude, the keyframe selection method has demonstrated the efficiency of reducing computational complexity by minimizing the amount of redundant data and increasing detection accuracy since the model focuses only on keyframes containing new information.

2) Experiment 1.2: Evaluate the Efficiency of Using a Spatio-temporal Entropy Template Vs. an Optical Flow Template: This experiment aims to compare and represent the efficiency of a spatio-temporal entropy template that the study has implemented in the proposed method against the optical flow (OF) templates that are applied by Keçeli et al. in [18]. In the proposed method, an entropy template applied to detect the motion region between the keyframes. This template has been created by applying the three-frames differences method and calculating an automatic threshold to detect moving objects. Then the moving region detected by comparing the entropy value with the threshold. While in [18] all video frames are used to generate four 2D templates, by calculating the (OF) of vertical and horizontal velocity, magnitude and orientation for adjacent frames via the Lucas–Kanade method [19].

Table VII demonstrates a comparison between the proposed method using a spatio-temporal entropy template generated by the keyframes and all frames against the method applied in [18] that used (OF) templates with all video frames. Both methods used the AlexNet network to extract features, and both of them used the Relieff feature selection method [37]. The comparison made between some of the public datasets used in this study, i.e. Hockey dataset, Movies dataset, and ViF dataset.

By analyzing the (Table VII), the spatio-temporal entropy template with selected keyframes and all video frames has resulted in a more accurate detection result than the OF templates used. Whereas the use of the spatio-temporal entropy template with keyframes provided optimal accuracy results when compared to using the spatio-temporal entropy template

TABLE VIII. COMPARING THE EXECUTION TIME TO CLASSIFY ONE VIDEO OF HOCKEY DATASET WITH THE METHOD APPLIED IN [18] VERSUS THE PROPOSED METHOD

Method	**Execution Time (s)**
Keçeli et al. [18]	2.2
The Proposed Method	0.59

TABLE IX. EVALUATES THE ACCURACY RESULTS WITH DIFFERENT PRE-TRAINED MODEL

Dataset Name	**Accuracy (%)**		
Model	**AlexNet**	**ResNet18**	**SqueezeNet**
Avenue	**87.5**	**87.5**	**87.5**
UCSD Ped1	**95.24**	**95.24**	57.14
UCSD Ped2	**100**	**100**	75
Hockey	**98.67**	93.33	80
Movies	**100**	**100**	94.6
ViF	**97.3**	93	74.1
Collected Dataset	**97.13**	85.5	88.2

with all frames. The improved outcome of the detection in the proposed method is due to the use of the three-frame difference method, which reduced the drawback of the approach proposed in [18]. As [18] used the difference between two frames to determine the optical flow values, which cannot accurately detect moving objects unless the acceleration of the object is constant. In addition, the proposed method used an automatic threshold calculation as it is more efficient and precise than using a static threshold. The explanation is that if the static threshold is too large, then it may not be able to detect moving objects. On the contrary, if the static threshold is small, then there could be a lot of noise. Consequently, the use of an automatic threshold eliminates noise and precisely detects the motion region.

Furthermore, (Table VIII) represents the required classification time by using the proposed method against the method implemented in [18] to classify one video from the Hockey dataset with a duration of 1s for resolution of (360 × 288). The measurement includes the generation of templates, features extraction, and prediction.

The study found that the proposed method classifies the input video approximately four times faster than the method used in [18]. Since the [18] approach generates four templates and each time the features are extracted from each template separately, then combining all the extracted features, which increases the processing time required.

3) Experiment 1.3: Evaluate the Efficiency of the Extracted Features Using Different Pre-Trained Networks: Notably, the previous two sections talked about using the keyframes, and spatio-temporal entropy template, which gave a high detection result. This experiment compared the efficiency of the anomaly detection model using different pre-trained networks (AlexNet [31], ResNet18 [38], and Squeezenet [39]) that used to extract deep features from a spatio-temporal entropy template as shown in (Table IX). This experiment aims to demonstrate the efficacy of the selected pre-trained network ‘AlexNet’ in the proposed method.

The experiment has proved that the pre-trained ‘AlexNet’ achieved better detection with all datasets than ResNet18 and SqueezeNet networks. As a result, the proposed method selected the AlexNet network to extract its deep features.

TABLE X. DETECTION ACCURACY USING DIFFERENT PERCENTAGES OF FEATURES SETS

Dataset Name	**Accuracy (%)**		
Percentage	**10% of Features**	**50% of Features**	**100% of Features**
Avenue	**87.5**	50	50
UCSD Ped1	**95.24**	95.24	95.24
UCSD Ped2	**100**	100	100
Hockey	98.67	**99.67**	**99.67**
Movies	**100**	98.33	98.33
ViF	**97.3**	90.54	90.54
Collected Dataset	**97.13**	94.25	71

TABLE XI. TESTING RESULTS USING DIFFERENT CLASSIFIERS

Dataset Name	**Classifiers**		
Classifier	**SVM**	**KNN**	**Decision Tree**
Avenue	**87.5**	71.43	52.38
UCSD Ped1	**95.24**	61.9	52.38
UCSD Ped2	**100**	**100**	50
Hockey	**98.67**	96.66	94
Movies	**100**	98.33	98.33
ViF	**97.3**	94.59	75.68
Collected Dataset	**97.13**	87.47	41.9

4) Experiment 1.4: Evaluate the Efficiency of Feature Selection Method: The Relieff feature selection method has been applied for those features extracted by the AlexNet. Since the use of all the extracted features or large sets of features may in some cases, degrade the detection results, even if all the features are related to the input variable. Because of that, this study tested the detection models with the best 10%, 50%, and 100% of the features, as shown in (Table X), which ranked the features by their weights to find the best set of features for anomaly detection.

As stated in (Table X) the selection of the best 10% of features from the extracted features provided better results in most datasets than the selection of a higher percentage of features set, except that the Hockey dataset obtained a lower result by only 1% than the result received when using the best 50% of features or 100% of features. Despite this, as the difference is not too significant, this study decided to select the best 10% of the features to generate a model.

5) Experiment 1.5: Evaluate the Efficiency of the Selected Classifier: The study used the selected features from the previous experiment to provide a classifier with those features to generate a model. In this experiment, several classifiers (SVM, KNN, and Decision Tree) examined to demonstrate their results, as illustrated in (Table XI). The SVM classifier applied with the ’linear’ kernel function while the classifier type for the KNN and the Decision Tree classifiers are ‘Fine’.

From (Table XI) the study analyzed that the use of the linear SVM classifier provided optimum results for detecting anomaly over other classifiers. In addition, a further examination applied to the linear SVM classifier to deal with large datasets. By assigning the optional ’solver’ parameter of the linear SVM with different variables (Sequential Minimal Optimization ’SMO’ and Interative Single Data Algorithm ’ISDA’), as shown in (Table XII). Consequences, the linear SVM classifier is affective in detecting an anomaly using the ’Automatic’ kernel scale and the ’SMO’ solver as parameters.

TABLE XII. THE OUTCOMES OF THE TESTING DATASETS USING LINEAR SVM CLASSIFIER WITH DIFFERENT VARIABLES OF 'SOLVER' PARAMETER

Dataset Name	**Accuracy (%)**	
Solver	**SMO**	**ISDA**
Avenue	**87.5**	50
UCSD Ped1	**95.24**	85.71
UCSD Ped2	**100**	**100**
Hockey	**98.67**	97.33
Movies	**100**	98.33
ViF	**97.3**	95.95
Collected Dataset	**97.13**	87.5

B. Experiment 2: Validate the Contributions of the Proposed Method

This experiment evaluates the contributions of the research, which detects both abnormal events and behaviors from crowd video scenes and generalizes the proposed model by applying a domain generalization technique for the detection of anomalies from different domains.

Several sub experiments have been applied, where the first sub experiment compares the obtained results of the anomaly detection with state-of-the-art approaches. Whereas, the second and third sub experiments are applied to validate the performance of the proposed model with domain generalization, and to demonstrate the efficacy of the proposed model with domain generalization for the detection of both abnormal events and behaviors from different domains.

1) Experiment 2.1: Comparison with State-of-the-art Methods: The study compared the proposed method with several state-of-the-art methods for detecting anomalies. A combination of hand-crafted approaches [40], [26] and deep learning approaches [12], [45], [46], [49]-[16], [18], [22], [24], and [52] were presented in (Table XIII and Table XIV). The quantitative performance of the proposed method was evaluated based on frame-level Accuracy and EER evaluation matrices, and comparing the results obtained with several methods. The higher Accuracy value refers to better classification. On the contrary, the lower EER represents the better performance of detection.

As shown in (Table XIII), the AUC of the proposed method outperforms the state-of-the-art methods in UCSD Ped1 and UCSD Ped2 datasets by 95.24% and 100%, respectively. While the accuracy of detection for the Avenue dataset is slightly inferior to [49] by 2.8% AUC.

The study also compared the frame-level EER with some state-of-the-art anomaly detection approaches as presented in (Table XIII). The study analyzed that the proposed method achieved a better EER performance result for all the three datasets. As both UCSD scenes provided the lowest EER by 0.05% and 0%, respectively. While the highest EER are recorded via Li et al. [40] by 21% and 20% for UCSD Ped1 and Ped2, as that approach is a hand-crafted approach based on a dictionary-learning algorithm. Additionally, the method proposed in the Avenue dataset achieved 12.5%, which generated the best EER compared to [49] by 3%. As shown in (Table XIV), in the Hockey dataset, the proposed model reached 98.67%, which surpassed all state-of-the-art methods except [16] as the proposed model was slightly lower than [16] by 0.29%. In the Movies dataset, the proposed model accurately classified all the videos by 100%, as similar to the results obtained by [12], [15], and [24]. Whereas, in the ViF dataset, the accuracy result of the proposed model is 97.30%, which exceeded all other methods.

2) Experiment 2.2: Validate the Performance of the Domain Generalization in Video-based: The approach required for most surveillance applications is to construct a generalized model for the detection of anomalies that is capable of detecting anomalies from different domains. In the following subsections, two experiments discussed to evaluate the generality of the model with and without domain generalization.

Experiment 2.2.1: Evaluate Cross-Dataset Performance without Domain Generalization

Several cross-dataset experiments in this experiment conducted using a transfer learning technique. Thus, selecting one of the six public datasets used in this study (Avenue, UCSD Ped1, UCSD Ped2, Hockey, Movies, and ViF), as the source domain for each examination and using the remaining datasets as a target domain. The findings of these examinations are shown in (Table XV). In specific, the model trained in the source domain is used to detect anomalies in the target domain.

In general, the study observed that training the model using a source domain and testing the model with different target domain suffers from poor anomaly detection performance. In addition, the anomaly detection accuracy is not consistent because the detection result is affected by the extent to which the source domain relates to the target domain, as shown in (Table XV). Where the source domain (e.g., UCSD Ped2) achieved 97.2% with the target domain (UCSD Ped1) while the other target domain (ViF dataset) had a poor detection result of 50%. Even though this is the case with most of the existing anomaly detection methods, which train and test the model with a specific dataset in a particular scene and provide high-precision results that exceed all benchmarks. Consequences, the cross-dataset experiment deduced that the generation of a model from a single source domain cannot be generalized to detect anomalies from various domains accurately. While higher performance achieved when the source domain and target domain derived from a similar domain.

Experiment 2.2.2: Evaluate the Performance of the Domain Generalization by Applying Cross-Domains

This experiment aims to demonstrate the effectiveness of applying cross-domains to create a generalized model that goes beyond specific tasks and domains. Through training a model with different domains to construct a less sensitive classifier capable of detecting anomalies from different domains. Since collecting datasets from each domain is considered as a difficult task, as well as unavailability of datasets for all possible domains. The study evaluated the generality of the proposed model for anomaly detection by applying the cross-domain technique as in (Table XVI), which is also referred as leave one-domain-out, i.e. taking one domain as the test set and combining the remaining domains as the training set.

In this experiment, six domains were set up, each containing five datasets presented as follows:

- Domain 1: The first domain is composed of Avenue dataset videos, UCSD Ped1and Ped2 datasets, Hockey

TABLE XIII. COMPARISON AREA UNDER ROC CURVE (AUC) AND EQUAL ERROR RATE (EER) FOR ANOMALY DETECTION WITH STATE-OF-THE-ART METHODS

Methods	**UCSD Ped1**		**UCSD Ped2**		**Avenue**	
	AUC	**EER**	**AUC**	**EER**	**AUC**	**EER**
Li et al. [40]	87.2%	21%	89.1%	20%	-	-
Liu et al. [42]	83.1%	-	95.4%	-	85.1%	-
Stack Denoising AE [45]	92.1%	16%	90.8%	17%	-	-
(MGFC-AAE) [46]	85%	20%	91.6%	16%	-	-
AE+ RNN [49]	90.5%	13.5%	88.9%	11.5%	**90.3%**	15.5%
Convolutional AE + LSTM [50]	89.9%	12.5%	87.4%	12%	80.3%	20.7%
Convolutional AE [51]	89.1%	8%	94.8%	12%	-	-
SL-MHOF+CNN [14]	90.8%	15.85%	97.8%	5.9%	87.2%	-
Aggregation of Ensembles [22]	94.6%	-	95.9%	-	89.3%	-
3D_GAN [52]	-	-	-	-	79.6%	24.1%
Proposed Model on Testing Datasets	**95.24%**	**0.05%**	**100%**	**0%**	87.5%	**12.5%**

TABLE XIV. COMPARISON AREA UNDER ROC CURVE (AUC) FOR VIOLENCE DETECTION WITH STATE-OF-THE-ART METHODS

Methods	**AUC (%)**		
	Hockey	**Movies**	**ViF**
CNN + BiConvLSTM [12]	96.54	100	92.18
Spatio-temporal [15]	97.0	100	-
3D convolution [16]	**98.96**	99.97	93.5
Optical flow + CNN [18]	94.40	96.50	80.90
CNN + LSTM [24]	97.1	100	94.57
Proposed Model on Testing Dataset	98.67	**100**	**97.30**

dataset, and Movies dataset. While the ViF dataset is the domain that has left over to use it for testing.

- Domain 2: The second domain is composed of Avenue dataset, UCSD Ped1and Ped2 datasets, Hockey dataset, and ViF dataset. While the Movies dataset is left over for testing.
- Domain 3: The third domain is composed of Avenue dataset, UCSD Ped1and Ped2 datasets, Movies dataset, and ViF dataset. While the Hockey dataset is left over for testing.
- Domain 4: The fourth domain is composed of Avenue dataset, UCSD Ped1 dataset, Hockey dataset, Movies dataset, and ViF dataset. While the UCSD Ped2 dataset is left over for testing.
- Domain 5: The fifth domain is composed of Avenue dataset, UCSD Ped2 dataset, Hockey dataset, Movies dataset, and ViF dataset. While the UCSD Ped1 dataset is left over for testing.
- Domain 6: The sixth domain is composed of UCSD Ped1 dataset, UCSD Ped2 dataset, Hockey dataset, Movies dataset, and ViF dataset. While the Avenue dataset is left over for testing.

The average accuracy of these six domains from (Table XVI) is 83.04%, which considered to be a good result of the detection of anomalies from an unseen domain. All models generated in this experiment that using domain generalization provided a high accuracy result, except 'Domain1' since the density level for the source domain and target domain is not equivalent, where the density for all datasets combined in Domain1 ranges from sparse to crowd. In contrast, the density level for the target domain (ViF dataset) is extremely crowded. Because of that, most of the target domain (ViF dataset) videos classified as abnormal videos.

In conclusion, this experiment showed the advantages of applying the domain generalization technique, as it provided a high accuracy results for the detection of anomalies across different domains. A further advantage is the elimination of the need to gather datasets from all possible domains.

3) Experiment 2.3: Validate the Proposed Model for Detecting Both Abnormal Events and Abnormal Behaviors from Video Scenes: This experiment presents the efficacy of the proposed model with domain generalization to detect abnormal events and behaviors from different unseen domains and compare its effectiveness with other state-of-the-art approaches, as discussed in the following subsections.

Experiment 2.3.1: Evaluate the Efficiency of the Proposed Model Using the Validation Dataset.

This section assesses the efficiency of the proposed model with domain generalization for detecting both abnormal events and behaviors from various unseen domains using the Validation Dataset and compares the performance of the proposed model against other models generated by train the model using only one of the public datasets to detect specific abnormal behavior, as illustrated in (Table XVII).

The study has proven that applying the domain generalization technique to the detection model improves the detection of anomalies from different domains. As illustrated in (Table XVII), the proposed model trained in the Collected Dataset detected anomalies from different unseen domains perfectly with an accuracy of 89.9%. As the precision metric recorded 0.97% accurate classification of abnormal videos, where the proposed model misclassified only one abnormal video and classified it as a normal video. Whereas the proposed model rightly classified normal videos by 0.80% as achieved by the recall metric.

Overall, this experiment showed that the proposed model with domain generalization outperforms all other models trained in a particular domain. As the proposed model is more generalized, which capable of detecting both anomalous events and behaviors from the Validation Dataset with high accuracy of 89.9%.

Experiment 2.3.2: Evaluate the Proposed Model with Domain Generalization Compared to state-of-the-art Methods

In this section, the study compared the efficiency of the proposed model with domain generalization for detecting abnormal events and behaviors from video scenes with several

TABLE XV. REPRESENTS CROSS-DATASET PERFORMANCE WITHOUT DOMAIN GENERALIZATION

Source Vs Target Dataset	Accuracy (%)						Range of Accuracy
	Avenue	**UCSD Ped1**	**UCSD Ped2**	**Hockey**	**Movies**	**ViF**	
Avenue	87.5	95.24	100	80	95	54	54%-100%
UCSD Ped1	100	95.24	83.3	26	48.33	50	26%-100%
UCSD Ped2	75	97.2	100	65	80	50	50%-100%
Hockey	50	57.14	83.3	98.67	96.67	59.46	50%-98.67%
Movies	50	23.8	33	92.67	100	52.7	23.8%-100%
ViF	75	57.14	50	54.33	60	97.3	50%-97.3%

TABLE XVI. THE ACCURACY RESULTS OF CLASSIFICATION USING THE CROSS-DOMAINS

Target Domain Source Domain	**ViF**	**Movies**	**Hockey**	**UCSD Ped2**	**UCSD Ped1**	**Avenue**
Domain 1	50%	x	x	x	x	x
Domain 2	x	91.76%	x	x	x	x
Domain 3	x	x	91%	x	x	x
Domain 4	x	x	x	100%	x	x
Domain 5	x	x	x	x	90.48%	x
Domain 6	x	x	x	x	x	75%

TABLE XVII. REPRESENTS THE AREA UNDER ROC CURVE (AUC) AND EQUAL ERROR RATE (EER), RECALL, PRECISION, AND F1-SCORE VALUES FOR DETECTING ANOMALIES FROM THE VALIDATION DATASET

Training Dataset Name	**Results on The Validation Dataset**				
Metric	**AUC**	**EER**	**Recall**	**Precision**	**F1-score**
Avenue	75.3%	0.25%	0.73%	0.77%	0.75%
UCSD Ped1	60%	0.40%	0.53%	0.62%	0.57%
UCSD Ped2	71.9%	0.28%	0.44%	**1%**	0.62%
Hockey	68.5%	0.31%	0.69%	0.69%	0.69%
Movies	50.6%	0.49%	0.02%	**1%**	0.04%
ViF	43.8%	0.56%	**0.82%**	0.47%	0.59%
Collected Dataset	**89.9%**	**0.11%**	0.80%	0.97%	**0.88%**

TABLE XVIII. COMPARISON OF THE PROPOSED MODEL WITH DOMAIN GENERALIZATION AGAINST STATE-OF-THE-ART APPROACHES IN THREE ANOMALY DATASETS

Accuracy (%)			
Ref.	**UCSD Ped1**	**UCSD Ped2**	**Avenue**
Li et al. [43]	87.2	89.1	-
Liu et al. [40]	83.1	95.4	85.1
Xu et al. [44]	92.1	90.8	-
Li and Chang [41]	85	91.6	-
Wang et al. [47]	90.5	88.9	**90.3**
Chong et al. [48]	89.9	87.4	80.3
Yang et al. [49]	89.1	94.8	-
Chen et al. [50]	90.8	97.8	87.2
Singh et al. [20]	94.6	95.9	89.3
Yen et al. [37]	-	-	79.6
Our Model with DG	**100**	**100**	87.5

state-of-the-art approaches, as illustrated in (Table XVIII and Table XIX).

In particular, the proposed model with domain generalization enhanced the accuracy for anomaly detection in the UCSD Ped1, UCSD Ped2, and ViF dataset compared to all state-of-the-art methods. While the accuracy of the Avenue dataset, the Hockey dataset, and the Movies dataset are slightly lower than the highest accuracy recorded by the state-of-the-art methods for each of these datasets by a maximum of 2.8%. Notably, the proposed model with domain generalization achieved, on average 96.52% accuracy as a result of the detection of different anomalies perfectly from different domains.

TABLE XIX. COMPARISON OF THE PROPOSED MODEL WITH DOMAIN GENERALIZATION AGAINST STATE-OF-THE-ART APPROACHES IN THREE VIOLENCE DATASETS

Ref.	**Accuracy (%)**		
	Hockey	**Movies**	**ViF**
Keçeli et al. [18]	94.4	96.5	80.9
Zohu et al. [15]	97	**100**	-
Song et al.[16]	**98.96**	99.97	93.5
Sudhakaran et al. [24]	97.1	**100**	94.57
Hanson et al. [12]	96.54	**100**	92.18
Our Model with DG	98.67	98.33	**94.59**

V. CONCLUSION

The work conducted in this research contributes to the field of the anomaly detection from crowd video scenes. Compared to other existing approaches, the novelty of this work lies in twofold. Firstly, applying a supervised deep learning approach to detect abnormal events and abnormal behaviors from crowd video scenes. Secondly, employ the domain generalization technique in a video-based model to improve the generality of the proposed model to detect anomalies from different domains.

The proposed model uses the keyframe selection method to select only the important frames and eliminate the nearby redundant frames. Also, it constructs a spatio-temporal entropy template for motion detection using the three-frame difference method and a dynamic threshold and using the pixel status cards technique to calculate the entropy value for each pixel. Furthermore, it employs the Relieff feature selection method to select the appropriate features, which extracted by a pre-trained network. We built two new datasets. Each of these datasets contains normal and abnormal events and behaviors videos. In particular, the Collected Dataset designed to evaluate the effectiveness of the proposed model in detecting abnormal events and abnormal behaviors from video scenes. Whereas the Validation Dataset created to evaluate the proposed model for the detection of anomalies from unseen domains. The comprehensive experimental study shows that the proposed method detects both abnormal events and behaviors in the Collected and Validation Dataset at a high accuracy rate of 97.13% and 89.9%, respectively. It also outperforms state-of-the-art methods with accuracy rates ranging between (87.5% to 100%). As future work, the proposed method can be extended to apply the domain generalization based on a semi-supervised approach for adaptability.

REFERENCES

[1] Luo, Weixin., Liu, Wen., Lian, Dongze., & Gao, Shenghua. (2021). Future Frame Prediction Network for Video Anomaly Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence , 44 , 7505-7520 .

[2] Bhuiyan, Md Roman., Abdullah, J.., Hashim, N.., & Farid, Fahmid Al. (2022). Video analytics using deep learning for crowd analysis: a review. Multimedia Tools and Applications , 81 , 27895 - 27922 .

[3] Motiian, S., et al. Unified deep supervised domain adaptation and generalization. in Proceedings of the IEEE International Conference on Computer Vision. 2017.

[4] Blanchard, G., G. Lee, and C. Scott. Generalizing from several related classification tasks to a new unlabeled sample. in Advances in neural information processing systems. 2011.

[5] Ghifary, M., et al. Domain generalization for object recognition with multi-task autoencoders. in Proceedings of the IEEE international conference on computer vision. 2015.

[6] Li, H., et al., Learning Generalized Deep Feature Representation for Face Anti-Spoofing. 2018. 13(10): p. 2639-2652.

[7] Li, H., et al. Domain generalization with adversarial feature learning. in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR). 2018.

[8] Muandet, K., D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. in International Conference on Machine Learning. 2013.

[9] Thong, W., P. Mettes, and C.G. Snoek, Open cross-domain visual search. arXiv preprint arXiv:1911.08621, 2019.

[10] Carlucci, F.M., et al. Domain generalization by solving jigsaw puzzles. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[11] Gong, R., et al. DLOW: Domain flow for adaptation and generalization. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[12] Hanson, A., et al. Bidirectional Convolutional LSTM for the Detection of Violence in Videos. in Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[13] AL-DHAMARI, A., R. SUDIRMAN, and N.H. MAHMOOD, ABNORMAL BEHAVIOR DETECTION IN AUTOMATED SURVEILLANCE VIDEOS: A REVIEW. Journal of Theoretical & Applied Information Technology, 2017. 95(19).

[14] Chen, Z., et al. Robust Anomaly Detection via Fusion of Appearance and Motion Features. in 2018 IEEE Visual Communications and Image Processing (VCIP). 2018.

[15] Zhou, P., et al. Violent interaction detection in video based on deep learning. in Journal of Physics: Conference Series. 2017. IOP Publishing.

[16] Song, W., et al., A Novel Violent Video Detection Scheme Based on Modified 3D Convolutional Neural Networks. IEEE Access, 2019. 7: p. 39172-39179.

[17] Sabokrou, M., et al., Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. Computer Vision and Image Understanding, 2018. 172: p. 88-97.

[18] Keçeli, A. and A. Kaya, Violent activity detection with transfer learning method. Electronics Letters, 2017. 53(15): p. 1047-1048.

[19] Barron, J.L., D.J. Fleet, and S.S. Beauchemin, Performance of optical flow techniques. International journal of computer vision, 1994. 12(1): p. 43-77.

[20] Wei, H., et al. Crowd abnormal detection using two-stream Fully Convolutional Neural Networks. in 2018 10th International Conference on Measuring Te

[21] Sultani, W., C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[22] Singh, K., et al., Crowd anomaly detection using Aggregation of Ensembles of fine-tuned ConvNets. Neurocomputing, 2020. 371: p. 188-198.

[23] Morales, G., et al. Detecting Violent Robberies in CCTV Videos Using Deep Learning. in IFIP International Conference on Artificial Intelligence Applications and Innovations. 2019. Springer.

[24] Sudhakaran, S. and O. Lanz. Learning to detect violent videos using convolutional long short-term memory. in 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). 2017. IEEE.

[25] Nasreen, A. and G. Shobha, Key frame extraction from videos-A survey. International Journal of Computer Science & Communication Networks, 2013. 3(3): p. 194.

[26] Li, Y., et al., Key Frames Extraction of Human Motion Capture Data Based on Cosine Similarity. vectors, 2017. 11(12): p. 1.

[27] Yu, H., et al., Translation domain segmentation model based on improved cosine similarity for crowd motion segmentation. Journal of Electronic Imaging, 2019. 28(2): p. 023011.

[28] Sehairi, K., F. Chouireb, and J. Meunier, Comparative study of motion detection methods for video surveillance systems. Journal of Electronic Imaging, 2017. 26(2): p. 023025.

[29] Hammami, M., S.K. Jarraya, and H. Ben-Abdallah, On line background modeling for moving object segmentation in dynamic scenes. Multimedia tools and applications, 2013. 63(3): p. 899-926.

[30] Zhang, Y., X. Wang, and B. Qu, Three-frame difference algorithm research based on mathematical morphology. Procedia Engineering, 2012. 29: p. 2705-2709.

[31] Krizhevsky, A., I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. in Advances in neural information processing systems. 2012.

[32] Srivastava, N., et al., Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 2014. 15(1): p. 1929-1958.

[33] Mahadevan, V., et al. Anomaly detection in crowded scenes. in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2010. IEEE.

[34] Lu, C., J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. in Proceedings of the IEEE international conference on computer vision. 2013.

[35] Nievas, E.B., et al. Violence detection in video using computer vision techniques. in International conference on Computer analysis of images and patterns. 2011. Springer.

[36] Hassner, T., Y. Itcher, and O. Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. in 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. 2012. IEEE.

[37] Robnik-Šikonja, M. and I. Kononenko, Theoretical and empirical analysis of ReliefF and RReliefF. Machine learning, 2003. 53(1-2): p. 23-69.

[38] Stock, P., et al., And the bit goes down: Revisiting the quantization of neural networks. arXiv preprint arXiv:1907.05686, 2019.

[39] Iandola, F.N., et al., SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and¡ 0.5 MB model size. arXiv preprint arXiv:1602.07360, 2016.

[40] Li, N., et al., Spatio-temporal context analysis within video volumes for anomalous-event detection and localization. Neurocomputing, 2015. 155: p. 309-319.

[41] Feng, Y., Y. Yuan, and X. Lu. Deep representation for abnormal event detection in crowded scenes. in Proceedings of the 24th ACM international conference on Multimedia. 2016. ACM.

[42] Liu, W., et al. Future frame prediction for anomaly detection–a new baseline. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[43] Ghrab, N.B., E. Fendri, and M. Hammami. Abnormal events detection based on trajectory clustering. in 2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV). 2016. IEEE.

[44] Zhou, X.-G. and L.-Q. Zhang. Abnormal event detection using recurrent neural network. in 2015 International Conference on Computer Science and Applications (CSA). 2015. IEEE.

[45] Xu, D., et al., Detecting anomalous events in videos by learning deep representations of appearance and motion. Computer Vision and Image Understanding, 2017. 156: p. 117-127.

[46] Li, N. and F. Chang, Video anomaly detection and localization via multivariate gaussian fully convolution adversarial autoencoder. Neurocomputing, 2019. 369: p. 92-105.

[47] Ravanbakhsh, M., et al. Abnormal event detection in videos using generative adversarial nets. in 2017 IEEE International Conference on Image Processing (ICIP). 2017. IEEE.

[48] Sabokrou, M., et al., Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Transactions on Image Processing, 2017. 26(4): p. 1992-2004.

[49] Wang, L., et al. Abnormal Event Detection in Videos Using Hybrid Spatio-Temporal Autoencoder. in 2018 25th IEEE International Conference on Image Processing (ICIP). 2018. IEEE.

[50] Chong, Y.S. and Y.H. Tay. Abnormal event detection in videos using spatiotemporal autoencoder. in International Symposium on Neural Networks. 2017. Springer.

[51] Yang, B., et al., Anomalous behaviors detection in moving crowds based on a weighted convolutional autoencoder-long short-term memory network. IEEE Transactions on Cognitive and Developmental Systems, 2018.

[52] Yan, M., X. Jiang, and J. Yuan. 3D Convolutional Generative Adversarial Networks for Detecting Temporal Irregularities in Videos. in 2018 24th International Conference on Pattern Recognition (ICPR). 2018. IEEE.