IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 4, 2021, accepted January 28, 2021, date of publication February 15, 2021, date of current version February 26, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3059519

Multi-Modal Anomaly Detection by Using Audio and Visual Cues

ATA-UR REHMAN[1], HAFIZ SAMI ULLAH[2], HAROON FAROOQ[2], MUHAMMAD SALMAN KHAN[3], TAYYEB MAHMOOD[2], AND HAFIZ OWAIS AHMED KHAN[4]

[1]Department of Electrical Engineering, University of Engineering and Technology at Lahore, Lahore 54890, Pakistan
[2]Department of Electronic Systems Engineering, University of Regina, Regina, SK S4S 0A2, Canada
[3]Department of Electrical Engineering, University of Engineering and Technology at Peshawar, Peshawar 25120, Pakistan
[4]Department of Electrical Engineering, Lahore University of Management Sciences, Lahore 54792, Pakistan

Corresponding author: Ata-Ur Rehman (a.ur.rehman@uet.edu.pk)

This work was supported by the Higher Education Commission of Pakistan under Grant 9827.

ABSTRACT This paper considers the problem of anomaly detection in an outdoor environment where surveillance cameras are usually installed to monitor activities of general public. A novel solution is proposed which combines audio and visual data to automatically detect abnormal activities. The proposed anomaly detection algorithm makes use of both visual and audio features to automatically detect anomalous activities in scenes. Visual features such as optical flow technique combined with particle swam optimization and social force model are used, whereas, acoustic features such as, energy, zero crossing rate, volume, spectral-centroid, spectral spread, spectral roll-off, spectral flux, cross correlation and the mel-frequency cepstral coefficients (MFCCs) are used. An anomaly inference is developed which is based on both visual and audio features. The performance of the proposed algorithm is evaluated by testing it on the publicly available UMN datasets combined with the audio recordings. The proposed algorithm is compared with state-of-the-art techniques and is shown to achieve improved performance in terms of accuracy.

INDEX TERMS Anomaly detection, SVM, particle swam optimization, social force model, mel-frequency cepstral coefficients.

I. INTRODUCTION

Over the past few decades, security has become a very challenging task. To monitor maximum urban areas, installation of surveillance cameras has rapidly increased around the world. With increase in the number of cameras, it has become almost impossible to manually monitor activities of general public through camera recordings. Therefore, automatic crowd behaviour analysis has become an emerging field in computer vision applications. Automatic video analysis to detect anomalous activities in surveillance videos is also known as anomaly detection [1]. Anomaly detection can be a pre-alarm or may signal a handler for reviewing more carefully the current scene. It also enhances the efficiency of surveillance by reducing man power and increasing safety precautions.

Anomaly detection is a very challenging task because a complete list of anomalies is usually not known. Most of the anomaly detection algorithms in the literature are based on data acquired through different types of sensors, for instance, video cameras or audio recordings. However, most of the existing anomaly detection algorithms for surveillance, consider only one sensor; either audio or visual sensors.

Several audio based anomaly detection algorithms have been proposed in the literature during the last one decade. For instance, algorithms for analysis of the individual events including speech recognition by machines are presented in [2], [3]. Auditory scene categorization are proposed in [4], [5], while, speaker identification systems are presented in [6], [7]. Moreover, there has been a growing interest in audio behaviour analysis for event detection in surveillance applications pertaining to public transport security [8]–[10]. There is an extensive research work available in the literature for audio classification by using hand engineered features. These hand engineered features are mainly utilized to train networks like Support Vector Machine (SVM) for classification purpose. Hand designed features to train SVM model to segment the audio signals are presented

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

in [10]. They propose the Expectation-Maximisation (EM) algorithm to estimate the parameters for the representation of feature space in Gaussian distribution [10]. Also different hand-crafted features are utilized in [11] for screaming and gunshot detection. Altogether they use 49 distinctive features based on spectral, perceptual, and temporal distribution. They perform feature selection based on hybrid wrapper method. The classifier is made up of two Gaussian Mixture Model (GMM) which are trained using Figueiredo and Jain algorithm [12]. Bag of words approach is used in [8] to train the classifier. At first, features like energy, volume, spectral flux, and spectrum dispersion are computed. Using K-Means clustering, features are converted to clusters which are then used to train SVM. They report average accuracy of 96%. Audio anomaly detection is wide in terms of scope and there can be multiple applications where it can be helpful. For instance, anomalous sounds in road accidents are detected in [13]. In another research presented in [14], different combination of features are utilized to train SVM classifier for voice clip and music clip. They conclude that using MFCC and frequency energy to train SVM can lead improvement in classification.

With the boom of Convolutional Neural Network (CNN) there is increasing trend towards solving problem of different complexity without the need of hand designed features. CNN provides end-to-end solution for different image, audio, text related tasks. In audio classification, [15] propose a model for end to end audio sample classification. They first extract the low-level feature using 1-D convolutional layer and use the traditional image classification network hierarchy to perform the audio classification task. Artificial data augmentation is applied by adding noise and normalizing the audio packet. At a sampling rate of $44.1k$ they achieve 85.65% accuracy [15]. J. Pons and X. Serra provide a brief comparison of randomly weighted CNN for audio classification. The CNN is fed with spectrograms or wave-forms of signals of size $29sec$. Baseline model with MFCC feature and SVM is used for comparison. Given the configuration utilized in their work, Visual Geometry Group (VGG) spectrogram based and temporal feature based front ends achieve better performance [16]. In [17] a method to initialize the weights for Deep Neural Network (DNN) given the different data distribution under road surveillance is proposed. They classify background sound signals in hazardous and non-hazardous class. They use hand engineered features based on spectral and temporal changes in the signal and train a recurrent neural network to objectify the class of given audio sample. They achieve more than 90% accuracy and F1 score [17]. In [18] the anomaly detection problem is treated as a regression-based classification problem within a deep learning framework. It proposes a solution to detect abnormal operation sounds in complex machines. Their solution is based on an auto-encoder, and it uses the residual error, which stands for its reconstruction quality, to identify the anomaly in machine. However, when the target machine sound is non-stationary, a reconstruction error tends to be largely independent of an anomaly, and its variations increase because of the difficulty of predicting the edge frames [19]. This problem is overcome by deep learning based autoregressive networks presented in [20] and interpolation deep neural network presented in [19]. These deep learning based methods achieve high accuracy which is evident from their Area under the Curve (AUC) scores, however, computational complexity of such algorithms is greater than the conventional methods. Furthermore, deep neural network based audio classifiers presented in [20] and [19] have not been tested in surveillance problems. In general, the audio detection is best achieved using SVM with a cost of feature extraction and not having end to end solution. We use extensive study to help us select prominent features, best representing the detailed information about the signal, to train the model.

Several Video based crowd analysis and anomaly detection algorithms have also been proposed. For instance, visual anomaly detection using different visual features is presented in [21]–[23], Social Force Model (SFM) based method is proposed in [24] and particle advection based techniques in [25], [26]. This particle advection based method detects anomalies and track them as well by placing a rectangular grid of particles on each video frame, which synchronises with the flow of objects within the frame. However, in crowded scenes it is very difficult to do such segmentation for anomaly detection and tracking. In such cases particle advection eliminates the use of tracking part [25]. Here the flow of advected particles is computed using fourth order RungeKutta-Fehlberg algorithm [27] with further support of bilinear interpolation of optical flow field. This method is improved by including coherent and incoherent scene by use of chaotic invariants [26]. In [21], spatial and temporal changes for given scene are computed with high accuracy by using flow of particles computed with streak lines in parallel with particle advection scheme. After fluid (crowd) flow estimation, Latent Dirichlet Allocation (LDA) [28] provides role of classifier and gives a score for force magnitude to classify anomaly detection. There are also some trajectory-based approaches to locate objects by tracking or frame-difference [29], [30]. The research work presented in [31] uses Hidden Markov Model (HMM) for scene categorisation to decide anomaly behaviour. Social force model aided with particle advection can also be used to translate flow of particles with the help of Particle Swarm Optimization (PSO) [32], [33].

Video based anomaly detection algorithms perform well when the video data is free of noise and anomalies are only visual anomalies. Therefore, accuracy of video based anomaly detection algorithms mentioned above may degrade when the video data is noisy, video data is not available temporarily for some reason, or anomalies are not visual anomalies such as gun shot. In such cases multiple modalities such as audio plus video can give better anomaly detection accuracy. Algorithms related to audio visual behaviour analysis are presented in [9], [10], [34]–[36], however, these algorithms analyze social behaviours of human beings

(e.g., friendly/aggressive) and some specific applications related to traffic. In best of our knowledge, there is no algorithm in the literature which uses both audio and video modalities to detect abnormal activities in scenes which are usually monitored by cameras only.

This research work proposes a novel framework for anomaly detection by using multiple modalities. A multi-modal anomaly detection framework is proposed which uses audio and video data for surveillance purposes. Advantages of the proposed method is its robustness to detect anomalies even when one of the modalities is temporarily not available or includes noise. The proposed method makes use of audio and visual classifiers and incorporates them into a novel anomaly detection inference model to detect anomalies in complex surveillance scenes.

The video classifier in the proposed technique calculates interaction force in each frame by using a social force model. These forces are aligned with particles flow with respect to a reference point. Forces are further compared with empirically chosen threshold, video frames with values of social forces lower than the threshold indicate ''Normal'' video frame, whereas, frames with forces higher than threshold are classified as ''Abnormal''. These interaction forces are computed by using particle advection and SFM [37] optimised by using PSO [33] as a robust algorithm.

The audio classifier is a SVM [38] which is trained with positive and negative data. Audio event detection is somewhat a pattern recognition problem as stated in [39]. The usual technique to evaluate the data is to evaluate the data on the basis of feature vectors containing one or more features for a single packet where, vectors for positive examples are classified as training set and are used for training a classifier. There are other methods which can also be used to formulate and evaluate feature vector for training purpose e.g. defining them on clusters with the help of means of set features as done in [13] but it takes extra efforts even though it minimises the problem of time span of testing signal.

To evaluate the performance of the proposed audio-visual anomaly detection framework, we have also implemented an interpolation based deep neural network for audio classification, incorporated it in the proposed audio-visual anomaly detection framework and evaluated its impact on the accuracy of the overall framework. Performance of the proposed audio-visual anomaly detection framework is evaluated with two different audio classifiers (SVM based classifier and the deep neural network based audio classifier) and comparison results are presented in Section III

Contributions of the proposed research are summarized below

1. A novel anomaly detection framework is proposed which utilizes audio and visual features to robustly detect anomalies in scenarios where video data can possibly be missing or noisy.
2. For detecting video anomalies, a SFM based approach is developed in which particle swam optimization algorithm is incorporated to optimize location and velocity of particles.
3. For detection of audio anomalies, we have implemented two different audio classifiers, incorporated them separately in the proposed anomaly detection framework and compared the performance of the overall audio-visual anomaly detection framework with two different audio classifiers.

The remainder of the paper is organized as follows: Section II describes methodology of the proposed framework, extensive experimental validation is given in Section III and conclusions are presented in Section IV.

II. METHODOLOGY

An anomaly detection algorithm is proposed which can recognise normal and abnormal events in surveillance area by using audio and visual cues. The proposed architecture is mainly based on two sections: visual classification and audio classification. In Fig. 1 bock diagram of visual classification based on particle advection and event classification using SFM and PSO is shown, whereas, Fig. 3 shows the block diagram of the proposed audio based classification by using Support Vector Machine (SVM) with the help of audio features extracted from the audio data.

A. VISUAL CLASSIFICATION

In the proposed framework for anomaly detection, we have developed a SFM [37] based method to detect abnormal crowd behavior. Social force model gives a mechanism for estimating interaction forces among individuals in a crowd. Therefore, it describes behavior of the crowd on the basis of interactions of individuals. Hence, normal activities in a crowd can be represented by normal social forces among the individuals whereas, abnormal social forces in the crowd portray abnormal behaviors.

Most commonly used method to compute social forces among the different objects is to first track them and then compute social forces [40], [41]. However, tracking individual targets in crowd is a very challenging task. Therefore, to estimate interaction forces we treat crowd as a collection of particles. Similar to [24], we place a set of particles on the video frame and move these particles according to the proposed particle advection technique and instead of computing interaction forces among individual targets in a crowd we compute interaction forces among these particles by using the social force model.

1) PARTICLE ADVECTION

Particle advection is useful to understand the flow of particles which represent objects moving in crowed scene. We propose a particle advection method which is based on optical flow combined with particle swarm optimization. Previously, in [24], [25] particle advection is employed by introducing a rectangular particles grid over each frame and then velocity for each particle is computed using fourth-order Runge-Kutta-Fehlberg algorithm [27] with interpolation of the

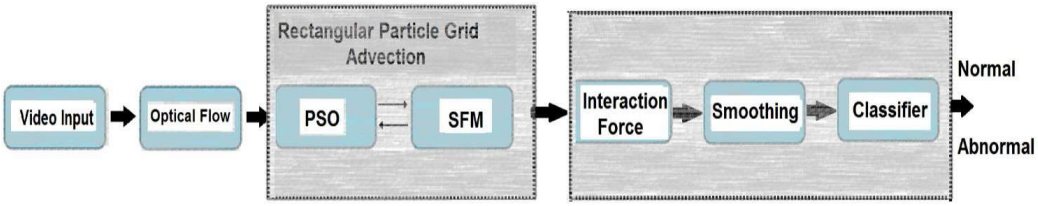

FIGURE 1. Visual event detection.

optical flow field. In crowded scene particles/objects may move with random trajectories making this approach inappropriate as it considers that the particles follow the fluid dynamical model. To overcome this problem we propose an improved particle advection technique which replaces fluid dynamical model with a robust modelling for understanding flow by using the proposed optical flow and particle swarm optimization method.

2) OPTICAL FLOW

Optical flow or optic flow is the sequence of visible motion of object in a displayed scene which is the result of relative motion between observer and the scene. Optical flow [42] reflects the changes occurred in an image when it moves. It is a representation of motion of an image or specifically the object of the image in terms of motion vector. To draw the motion vector, we analyse video frames. Suppose we are analyzing a video at time t_1 and t_2, we can draw the estimation for the next sequence of motion in term of motion vectors as shown in Fig. 2.

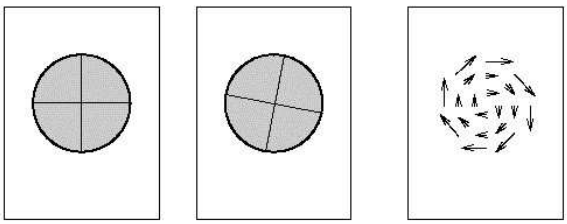

FIGURE 2. Optical flow estimation.

There are a number of methods to estimate the optical flow based on partial derivatives of the images i.e. for lower and higher-order partial derivatives, such as Lucas–Kanade method, Horn–Schunck method, Buxton–Buxton method and Black–Jepson method. For the proposed algorithm we have used Lucas-Kanade algorithm [43] to obtain horizontal and vertical optical flow components of particles. Once we have optical flow information of all the particles, we calculate the average optical flow field O_{avg} of particles over a fixed window of N video frames. Average optical flow field of particle with pixel location x_i and velocity v_i is represented as $O_{avg}(x_i)$, whereas, optical flow field in the current video frame is represented as $O(x_i)$.

3) PARTICLE SWARM OPTIMIZATION

Particle Swarm is a robust optimization technique which is normally used to control a swarm of data in an iterative way resulting in optimized solution to a problem [33]. It optimizes a criterion function called fitness function taking initially a swarm of randomly organized particles with X- dimensions over a finding area or search space. The optimization function tries to manage the flow of particles in accordance with the fitness function iteratively. There are two positions namely the p_{best} and the g_{best}. They depend on i-th particle and independently valid for whole swarm respectively. P_{best} is the best position related to either maximum or minimum fitness function where g_{best} is the best position among all. The position change between two consecutive p_{best} can be termed as velocity as it is the change in position with respect to time and can be represented as v_i (i.e. the velocity of the i-th particle). Particle swarm optimization is an iterative process which updates velocity of every particle according to following equations [33] to find the optimum velocity

$$v_i^{new} = Iw \times v_i^{old} + C_1 \times rand_1 \times (pbest_i - X_i^{old}) + C_2 \times rand_2 \times (gbest - X_i^{old}) \quad (1)$$

$$x_i^{new} = x_i^{old} + v_i^{new} \quad (2)$$

where I_w is inertial weight, it is a tuning parameter tuned to balance the global and local explorations, well-tuned value may reduce running time cost. Constants C_1 and C_2 are the drifting parameters for target location, lower value of these constants allow the particles to move away from target region while higher values cause abrupt drift towards target location. The $rand_1$ and $rand_2$ are two random numbers with values between 0 and 1. Finally, x_i^{new} and v_i^{new} are the updated values of location and velocity of i^{th} particle respectively while x_i^{old} and v_i^{old} are the previous values of particle location and velocity respectively.

Once we have optimum location and velocity of all the particles we can use social force model to calculate the interaction forces for every particle.

4) SOCIAL FORCE MODEL (SFM)

With the help of following equations, we define social force model [37] for particle motion dynamics by considering personal and environmental interactions. It facilitates the formulation for the movement of each particle in the scene.

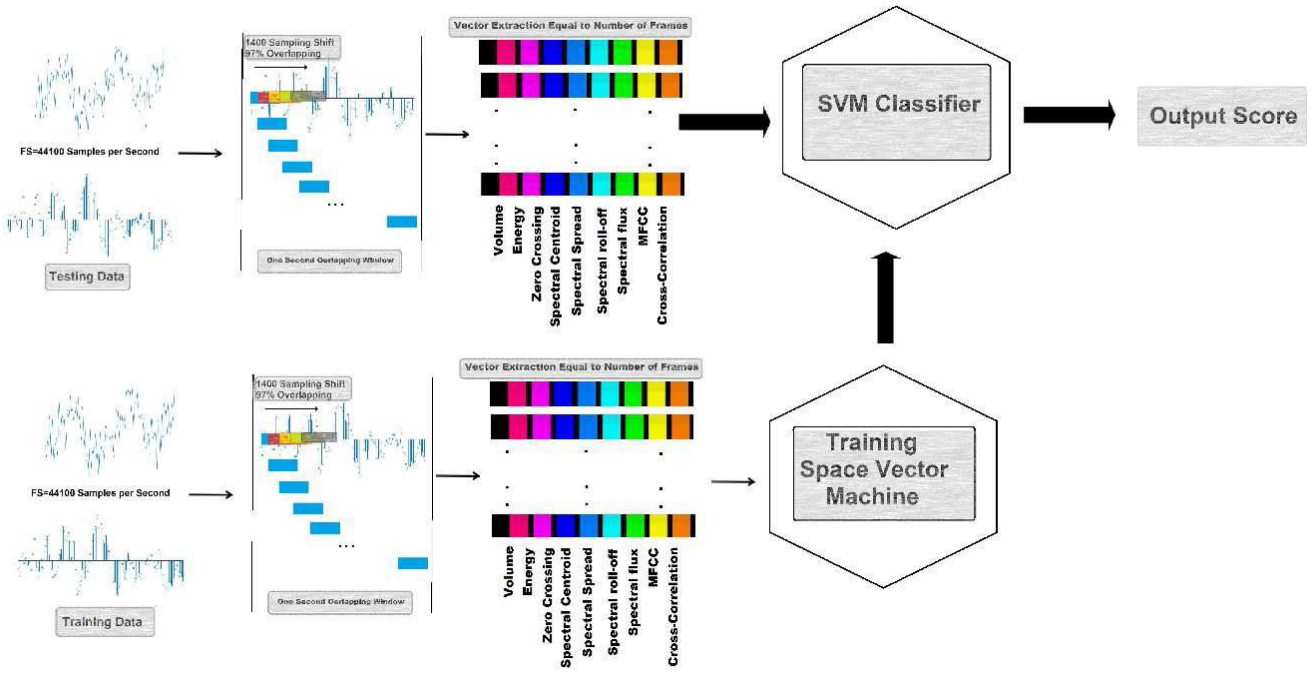

FIGURE 3. Audio classification.

Mathematically, SFM can be described as:

$$\begin{aligned} F_a &= m_i \times a_i \\ &= m_i \times \frac{dv_i}{dt} \\ &= F_p + F_{int} \end{aligned} \tag{3}$$

where F_p represent the personal force, F_{int} describes the interaction force, m_i and v_i being the mass and velocity of i-th particle respectively. We can write F_p as [37]

$$F_p = m_i \times \frac{v_i^p - v_i}{\tau_i} \tag{4}$$

where, v_i is the actual velocity of particle i and v_i^p is its desired velocity. In the proposed work, we compute these velocities as

$$v_i = O_{avg}(x_i^{new}) \tag{5}$$

and

$$v_i^p = (1 - q_i)O(x_i^{new}) + q_i O_{avg}(x_i^{new}) \tag{6}$$

where q_i is the parameter which describes individual behaviour of i^{th} particle. Particle exhibits individual behaviour if its values approaches zero and herding behaviour if it approaches unity.

The interaction force F_{int} is a combination of repulsive/attractive force F_{ped} and a psychological repulsive force between pedestrian and walls/buildings called environmental force F_w i.e.

$$F_{int} = F_{ped} + F_w \tag{7}$$

Interaction force for an i^{th} particle is calculated as

$$F_{int}(x_i^{new}) = m_i \times \frac{dw_i}{dt} - \frac{m_i}{\tau_i} \times (v_i^p - v_i) \tag{8}$$

where $\frac{dw_i}{dt}$ can be evaluated by calculating the difference between two consecutive optical flows that is $\frac{dw_i}{dt} = (O(x_i^{new})|_t - O(x_i^{new})|_{t-1})$. This equation states that the force allows the particles to direct from desired path to actual one. The particles are shifted towards area of the large motion and driven by optical flow.

Based on the interaction forces we define a global classification criteria to classify video frames into anomaly and normal classes.

5) VISUAL ANOMALY DETECTION

The proposed algorithm classify video frames into normal and abnormal on the basis of interaction forces. Lower magnitude of interaction forces describe normal behaviour, thus, our fitness function tries to drive the particles toward area of minimum force. Less interaction force means a regular movement of particle. For classification, we set an empirically defined threshold, interaction force less than the predefined threshold represents a normal video frame, whereas, force above that threshold represents an anomalous video frame. Using fitness function to move particles towards small interaction force, thereby allows the particles to simulate a ''normal'' situation of the crowd or vice versa. Main routine for the proposed audio visual anomaly detection algorithm is presented in Algorithm I.

Algorithm 1 Main Routine for Audio Visual Anomaly Detection

Input: Video, Audio signals
Output: Classification of video frames into normal and abnormal classes
Initialization:
1. Initialize particle velocity as $v_{i=1:k}$ and position $x_{i=1:k}$, where k is the total number of particles.
2. Initial p^i_{best} with local best position
3. Initialize g_{best} with global best position
4. Initialize O and O_{avg} with a $n \times m$ matrix whose all elements are zero. where $n \times m$ is the resolution of the video frame.

for Frame $i = 1$: Total no of Frames **do**
  if Remainder of Frame and Frame Rate = 0 **then**
    Get audio sample for the next second
    Calculate Score using Algorithm 2
    return detection score
  end if
  Resize frame and Making Usable by setting no of rows = 120 and no of columns = 160
  Calculate Average optical flow and optical flow for each video frame by using $Out = Oavg$ and O
  while $Itr \leq maxiter$ and $F_{int}(g_{best}) \leq BestFit$ **do**
    for for i = 1: k **do**
      Calculate F_{int} using Equation 8
      if $F_{int}(x^{new}_i) < F_{int}(p^i_{best})$ **then**
        $p^i_{best} = x^{new}_i$
      end if
    end for
    Set $g_{best} = argmin(p^i_{best})$
    Estimate new positions for particles using equation 1 and 2
  end while
  Output: Optimized interaction forces and
  New particles' positions.
  Store previous usable values
  if Frame $i <= 10$ **then**
    Calculate F_r using Equation 10 where r is 10
  end if
  if Frame $i > 10$ **then**
    Calculate F_t using Equation 11
    Calculate C_t based on difference mentioned in Equation 12
    C^s_t: Apply moving average filter to smooth C_t
    if detection score >th1 & C^s_t >th2 **then**
      Abnormal with Firing
    end if
    if detection score <=th1 & C^s_t <=th2 **then**
      Normal
    end if
  end if
end for

The fitness function can be described as:

$$FitnessFunction = min[F_{int}(x^{new}_i)] \tag{9}$$

where i indicates the i-th particle. Algorithm I includes *maxiter* term which indicates the number of iteration, in the proposed work its value is set to 100. For anomaly detection the reference force is taken from the first 10 frames. The average of force is calculated based on equation 10 given below:

$$F_r = \sum_{1=1}^{k} F_{int}(X^{new}_i)|_r \tag{10}$$

where $r = 10$ because we are taking 10 frames for training data and F_r represent the normal/trained/reference data. For testing data, we will also calculate the interaction sum using equation 11 as:

$$F_t = \sum_{1=1}^{k} F_{int}(X^{new}_i)|_t \tag{11}$$

The next step is to calculate absolute force difference between reference or training frames and current frame using equation 12 as:

$$C_t = |F_t - F_r| \tag{12}$$

The second last step is to use moving average filter to smooth output and at the end the output is categorized based on threshold chosen experimentally.

$$R_t = \begin{cases} Abnormal & \text{if} \quad C^s_t > th \\ Normal & \text{otherwise} \end{cases} \tag{13}$$

where, C^s_t is output from smoothing filter, th is defined threshold while the R_t holds the final decision for normal or abnormal classification.

B. ACOUSTIC CLASSIFICATION

This section describes in detail the steps involved in classifying the audio events into normal and abnormal events. The architecture for audio classification presented in this paper is based on bags of words approach. In the proposed work, the audio stream is divided into small audio frames known as packets. For each audio packed we extract number of audio features which are further used to evaluate the probability of the occurrence of these packets. Based on these probabilities we categorise audio events into normal and abnormal events. Features extracted from audio packets are explained in the next section

1) AUDIO FEATURE EXTRACTION

Unlike video frames, the audio changes very rapidly, therefore, even within a fraction of a second the audio events may change, which means, every event needs to be processed before the occurrence of the next event. Therefore, to detect every audio event and to classify it into normal and abnormal audio event, from audio sequences accurately, we chose the

window size small. The window size T_w is chosen carefully to cover all the frequencies. No matter what the audio frequency is, lower or higher, the frame size remains the same. For each frame, the feature vector is computed which can either be used for high level feature extraction or can be used directly to predict the score using trained classifier with positive and negative examples.

For every audio packet we compute two different sets of audio features. The first set of features includes 1) energy, 2) zero crossing rate, 3) volume, 4) spectral-centroid, 5) spectral flux, 6) spectral spread, 7) spectral roll-off and 8) cross correlation [8], [44], [45], whereas, the second set of features include Mel Frequency Cepstral Coefficients (MFCC) [46]. **Energy:** To calculate the weighted sum of samples energy of sound within the time window, We have used hamming window, a type of raised cosine defined as

$$\mathbf{w}[n] = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{M-1}), & 0 \leq n \leq M-1 \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

whereas, the energy is calculated as

$$E_n = \sum_{m=-\infty}^{m=+\infty} (x[m]w[n-m])^2 \quad (15)$$

where E_n is the short time energy, $x[m]$ is the input signal, and $w[n]$ represents the hamming window. Based on the properties of short time energy algorithm we can find; (a) Maximum Window Energy (b) Total Energy (c) Signal Envelope (d) Signal Extraction.

By comparing individual window energies, we can find out window energy or signal sample which contains maximum window energy. Abnormal audio events such as a gun shot, produce impulsive sound, we can differentiate them from other signals based on maximum window energy, for instance in our case most of the normal sound events have maximum window energy not more than hundred units, so we can say that energies more than 100 units is optimum condition for a gunshot. The total energy is sum of energies carried by all the windows.

Zero crossing rate: The Zero crossing rate is the sign change going from positive value to a negative value and vice versa, it can be defined as

$$zcr = 1/(2K) \sum_{m=-1}^{m=K} [sign(x[m+1]) - sign(x[m])] \quad (16)$$

where K is the length of the signal $x[n]$, and $sign(\cdot)$ is a function that evaluates the sign of extracted value.

Volume: The volume of signal is the RMS value of amplitude of audio samples which can be evaluated as

$$V = \sqrt{\frac{1}{K}\sum_{m=1}^{K} x[m]^2} \quad (17)$$

where V represent volume and K is the length of the signal $x[m]$.

Spectral Centroid: The spectral Centroid SC shows the weighted mean of frequencies available in a signal and can be evaluated as

$$SC = \frac{\sum_{i=b0}^{b1} f_i s_i}{\sum_{i=b0}^{b1} s_i} \quad (18)$$

where s_i is the spectral magnitude at bin i and f_i represents the centre frequency of the $i^t h$ bin, whereas, b_0 and b_1 are the lower and upper limit of the bin for which spectral centroid calculated.

Spectral Flux: The Spectral Flux SF is the measure of how quickly the power spectrum of a signal is changing, it can be computed as the square difference between the normalized magnitude of spectra of two consecutive windows as reported in equation given below

$$SF = \sum_{i=b_0}^{b_1} [s_i(t) - s_i(t-1)]^2 \quad (19)$$

where $s_i(k)$ represents the kth normalized Discrete Fourier transform (DFT) at i^{th} frame and b_0 and b_1 are the edges of the band.

Spectral Spread: The Spectral Spread is spread of the spectrum around its mean value. It can be computed as [45]

$$SS = \sqrt{\frac{\sum_{k=b0}^{b1}(f_k - \mu_1)^2)s_k}{\sum_{k=b0}^{b1} s_k}} \quad (20)$$

where f_k is the frequency of k^{th} bin in hertz, s_k is the spectral value for k^{th} bin, and b_0 and b_1 are the edges of the band over which spectral spread is calculated. Parameter μ_1 is the Spectral Centroid.

Spectral rolloff The Spectral rolloff point is the frequency under which the 85% percent of spectral energy lies. It can be computed as

$$\sum_{k=b0}^{R_t} s_k = \alpha \sum_{k=b0}^{b1} s_k \quad (21)$$

where α is controllable threshold and specifies how much percentage of total energy contained from b_0 to R_t, for our case its 0.85. $b0$ and $b1$ are the band limits over which we are evaluating spectral spread.

Cross-Correlation The Cross-Correlation is the measure of similarity of two signals. It is defined as

$$(f * g)[n] \triangleq \sum_{i=-\infty}^{\infty} \overline{f[i]} g[i+n] \quad (22)$$

$\overline{f[i]}$ defines the conjugate of $f[i]$ where f and g are two functions under consideration. Envelopes of all the gunshots are very similar to each other, so, we can distinguish gunshot from the noise by taking correlation of envelope of signal with the envelopes of the reference gunshots.

Mel Frequency Cepstral Coefficients (MFCC): At the end, the MFFCCs are extracted. The MFCCs are obtained

by applying Discrete Cosine Transform (DCT) to log transformed energy. Mathematically it can be understood as

$$c(l) = \sum_{m=1}^{M} X'(m) cos(l\frac{\pi}{M}(m - \frac{1}{2})), \quad wherel = 1, 2, \ldots M \tag{23}$$

where $X'(m)$ is scaled version of $X(k)$ using Mel filterbank [46]. Usually first 13 coefficents are considered.

2) TRAINING AND CROSS VALIDATION OF CLASSIFIER

After extracting low level feature vectors, the task of supervised learning starts. The feature vectors from positive example of event we are interested in detecting, can be used to train a classifier. We used linear support vector machine [38] for classification purpose. The training dataset should be roughly ten times the number of features, but it also works for the lower number of training sets. We are using 44 positive and negative examples to train classifier to recognise the gun voice. In the proposed algorithm binary SVM is used which can be extended to an N class classifier. This N class classifier will actually have $N + 1$ classes, the one extra class is for non-matching voices. For better results we used SVM with linear kernel as it changes the distribution of data on axes and act as dot product.

The SVM is a binary input machine as it can classify two examples i.e. positive or negative. Thus, a pool of SVM Figure 4 is realized in order to face the multi-class problem at hand. Algorithm for audio event detection by using SVM classifier is summarized in Algorithm II.

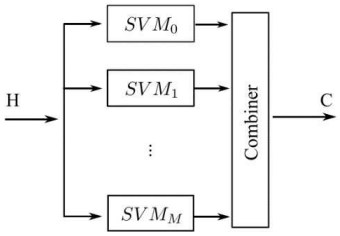

FIGURE 4. SVM Architecture.

The i-th classifier is trained by considering training examples from class C_i as positive examples while samples from all other classes are taken as negative examples. The SVM cross validation outputs a score which is taken as a matching probability. This output probability corresponds to level of matching between the testing voice with the trained voice.

$$C = \begin{cases} C_0 & S_i < \delta \\ \arg\max S_i & \text{otherwise} \end{cases} \tag{24}$$

hlIf all the classifiers give a confidence score $si < \lambda$ the time interval is classified as a background sound in class C_0. For our experiments the threshold is set as $\lambda = 10$. When SVM is used for background noise classification as well it becomes more reliable and enhance itself while reducing the false indication.

Algorithm 2 Audio Event Detection Using SVM Classifier

Input: Audio File
Output: Score
Initialization:
1. Initialize A_i and X_i, where i = No of features
2. L = compute length of signal
3. E = Apply window and calculate energy by adding the squares of magnitude of samples
4. Window = Move window forward by 500 and again calculate till the last sample is approached

1: Compute the Volume by using Eq. 17
2: Compute Maximum Window Energy, Total Energy and the envelope.
3: Compute Spectral roll off = $arg[(|X(f)|^2) * 0.85 < |X_{half}(f)|^2]$, where, Spectral Roll off is the fraction of bins in the power spectrum at which 85% of the power is at lower frequencies
4: Take FFT of input define frequency axis
5: Starting from 0 Hz combine energy of samples till it become 85% of half of the energy of total spectrum
6: Compute Spectral Centroid by using Equ. 18
7: Compute Spectral flux by using Equ. 19
8: Compute ZCR by using Equ. 17
9: Compute Cross Correlation between sample vectors, X and Y
10: Compute the mel frequency cepstral coefficients of a speech signal using the mfcc function of MATLAB
11: Design a linear predictor (FIR) filter for reference gunshot signal using MATLAB.
12: Pass the signal to be classified through that filter
13: compare estimated error for signal with that of reference
14: if value is within threshold than signal is gunshot else noise
15: Store feature Matrix into X, i.e. $X = FeatureMatrix$
16: **if** Second = 1 **then**
17: M = Load Training Data
18: SVMTRAINED = Training SVM Classifier
19: SAVE SVMTRAINED
20: **end if**
21: LOAD SVMTRAINED
22: Cross Validate SVMTRAINED with "X" Feature Vector
23: Return SCORE

III. RESULTS

The performance of the proposed algorithm is evaluated by measuring the accuracy of the proposed technique. The audio and video anomalies are different streams, but one can cause the other to happen. For example, if there is a gunshot, crowd can get panic, which is a visual anomaly clue. Combining both anomalies, the result for the anomaly detection can be improved. We used this hypothesis to perform our experiments.

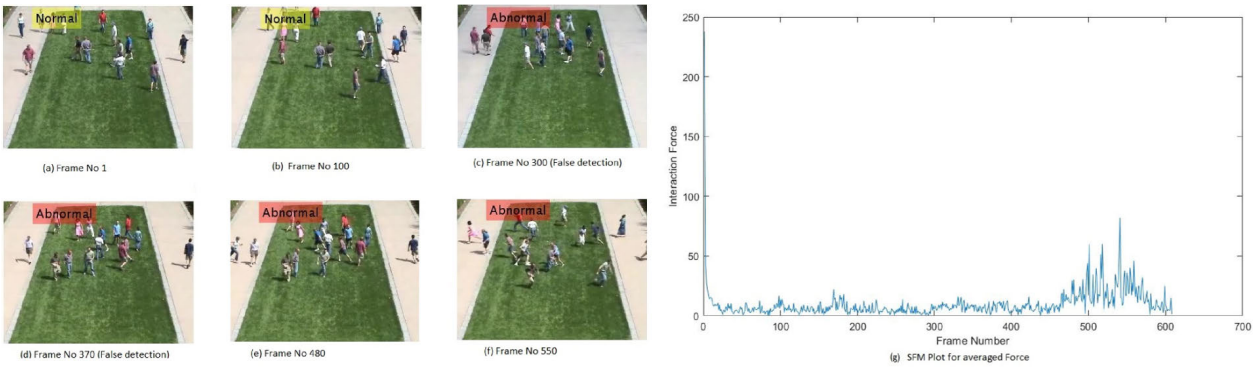

FIGURE 5. UMN data set visual Classification Sequence 1.

For faster and reliable approach towards the accurate anomaly detection, the size of video frame which we randomly choose to be 120×160, the particle adverted on each frame are kept constant i.e. 100. This ensures that output quality remains the same as the input. Windowing for moving averaging filter is kept at 15 samples per filter to capture the steady state response, so that filter can work smoothly.

To decide the decision boundary between anomaly and normal condition we define two thresholds. These thresholds are selected empirically to get a tradeoff between precision and recall. Threshold $th1$ is chosen to be 70, and threshold $th2$ is selected as 0.03. Videos used for evaluation of the proposed technique are recorded at 30 frames per second. To reduce false positives while restricting missed detections to rise, we empirically selected a wait time of 100ms. This ensures that wrongly detected anomalies which last for less than three video frames should not be detected.

Three different types of extensive evaluations are presented in this section.

1. Comparison of video based anomaly detection algorithms with the proposed anomaly detection framework without audio cues is presented. For comparison, proposed method is compared with the SFM based anomaly detection method [24].
2. Comparison of audio classification based on SVM based classifier and interpolation based deep learning audio classifier [19] is presented.
3. Comparison of the accuracy of the proposed audio-visual anomaly detection framework by using two different audio classifiers is presented. Furthermore, the proposed audio-visual anomaly detection algorithm is compared with the other state-of-the-art in the field of online visual anomaly detection. Methods chosen for the comparisons are online anomaly detection algorithms which are the state-of-the-art in the field of visual anomaly detection

In the best of our knowledge, there is no publicly available audio-video datasets, therefore, we have used separate publicly available datasets for evaluation of audio and video anomaly classifiers. Audio anomaly classifiers are tested and compared on the publicly available DCASE dataset [47]. DCASE is an audio only dataset and it does not include videos. To compare the performance of video classifiers we have used publicly available UMN datasets [48]. For audio-visual anomaly detection, we have combined audio and video data to create audio-visual data for evaluation of the proposed audio-visual anomaly detection framework. The process used for creation of audio-visual data is explained in Section III-C.

A. ANOMALY DETECTION WITH VIDEO CUES

Evaluation of the proposed anomaly detection algorithm without using audio cues is conducted on video datasets including UMN [48] dataset.

Results with video cues by using SFM and particle advection-based visual anomaly detection method are presented in Figs. 5 - 9. It can be seen in these figures that anomaly detection based on visual data only, can sometimes fail to correctly classify the events into normal and abnormal events. There are false negatives as well as false positives in detection. Another phenomena which is obvious in these Fig. 9 is that there is late identification of anomaly in the existing techniques. This delay in event detection is of approximately half of a second.

The results based on anomaly detection on video data only, shows miss-classifications as can be seen in Fig.6 (c and d) and in Fig. 7 (e and f). The problem of late detection and false detection has been solved in the proposed anomaly detection framework. It can be seen in Table 1 that improved detection accuracy is achieved even without using audio cues. However, it is later shown in the results that the inclusion of audio cue has further improved the accuracy. Furthermore, in the previous PSO based particle advection methods it was hard to define threshold, we eliminated this problem by introducing a reliable filter and considering anomaly if it happens in three consecutive video frames. Results presented in Table 1 show that the proposed method even without inclusion of audio cues shows better results compared to [24].

B. ANOMALY DETECTION BY USING AUDIO CUES ONLY

In this subsection we compare accuracy of SVM based audio classifier with the deep learning based audio classifier [19]. Datasets used for comparisons are taken from the DCASE

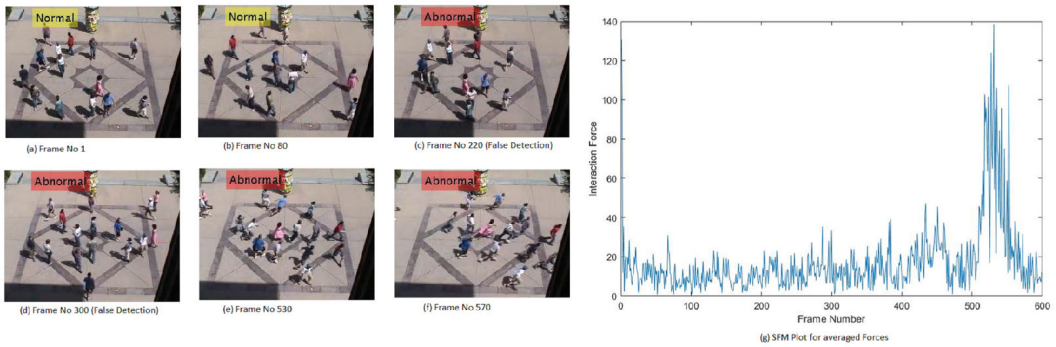

FIGURE 6. UMN data set visual classification Sequence 2.

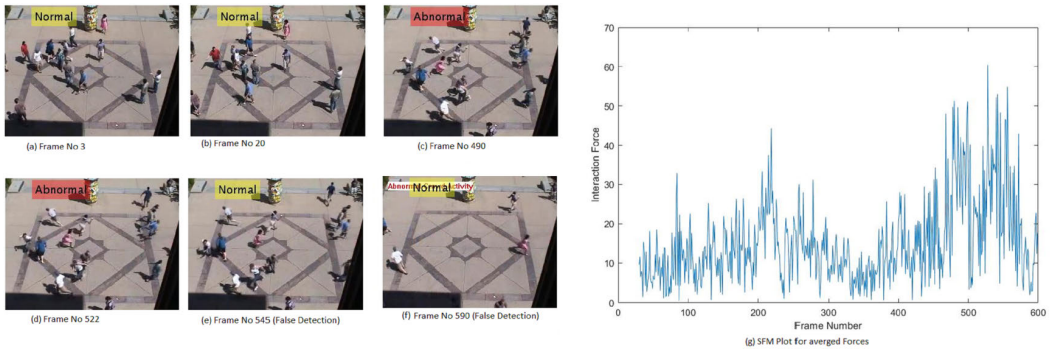

FIGURE 7. UMN data set visual classification Sequence 3.

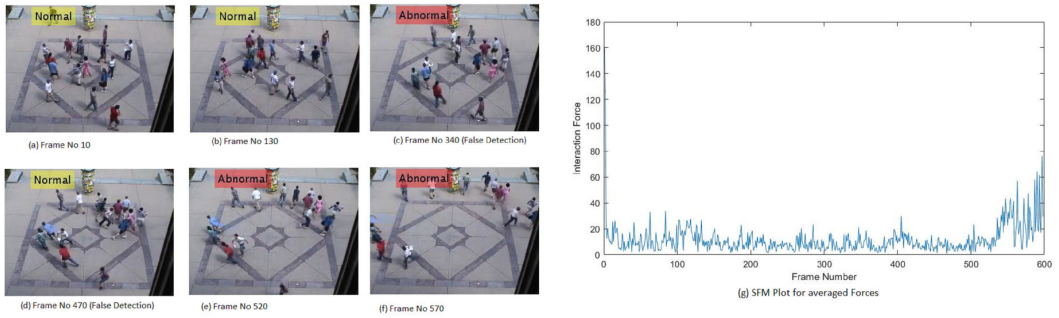

FIGURE 8. UMN data set visual classification Sequence 4.

TABLE 1. Comparison of video based anomaly detection.

Video Sequences	Video based Anomaly Detection [24]				Proposed Method without Audio Cues			
	TP	FP	FN	Accuracy	TP	FP	FN	Accuracy
UMN Dataset Seq 1	97	2	7	0.91	99	1	6	0.93
UMN Dataset Seq 2	73	3	6	0.89	75	2	5	0.91
UMN Dataset Seq 3	78	5	6	0.87	79	3	7	0.88
UMN Dataset Seq 4	53	6	8	0.79	59	5	3	0.88
UMN Dataset Seq 5	73	3	9	0.85	77	3	5	0.90
Total	374	19	36	0.87	389	14	26	0.90

Challenge Task 2 [47]. Audio datasets with three different anomalous events: baby cry, glass break and gunshot are taken from the data set and used for comparison of audio results. These 15 datasets are recorded in 15 different

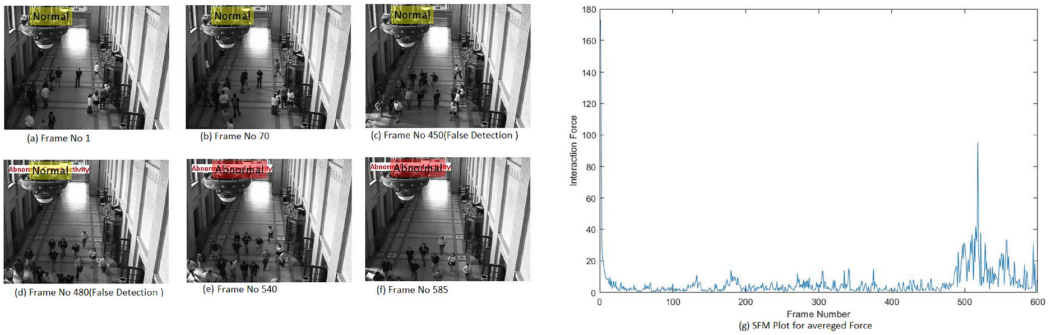

FIGURE 9. UMN data set visual classification Sequence 5.

environments which are listed in Table 2. Area under the Curve (AUC) scores of [19] and the SVM based technique are compared in Table 2. It can be seen that the performance of both audio classifier is somewhat similar.

TABLE 2. Comparison of AUC score for deep learning based Autoregressive networks audio classifier and SVM based audio classifier.

Scene	Deep learning based audio classifier [19]	SVM based audio classifier
beach	0.72	0.71
bus	0.83	0.81
cafe/restaurant	0.76	0.75
car	0.82	0.83
city center	0.82	0.80
forest path	0.72	0.71
grocery store	0.77	0.78
home	0.69	0.69
library	0.67	0.69
metro station	0.79	0.80
office	0.78	0.78
park	0.80	0.79
residential area	0.78	0.76
train	0.84	0.82
train	0.87	0.86

C. ANOMALY DETECTION BY USING AUDIO AND VIDEO CUES

The audio visual data sets which include examples of different type of gunshots are used for the evaluation purposes. There is no publicly available audio visual data set for anomaly detection, therefore, we have created audio-visual data set by combining publicly available audio dataset with the video dataset.

Our solution is relevant to unusual crowd activity with possible reason of anomalous activity. Considering this we have taken audio data from DCASE Challenge Task 2 [47] which describes situations where people are in some hallway, park, and train station, we also collect almost hundred examples for gunshots with variety of gun shot types. There are challenges in overlapping and aligning the anomaly in video with the audio sequence. Under the given circumstances we did not have the audio having all the characteristics for the proposed task. The first task is to find the binary decision pattern for the anomaly in video sequence. The output is the prediction only based on visual cues. This binary decision is further refined by manual input by observing the individual frame in the dataset. After extracting the anomalous decision boundary, we prepare multiple audio clips by mixing of sound clip of people in hallway, train station, park with gunshots. The mixing is done in such a way that the gunshot clip is only chosen if the visual decision boundary evaluates to true and there is manual input of allowing the gunshot to be added. We add this chosen gunshot to sound clips of crowd in public and normalize the range of $[1, -1]$.

For better understanding we explain the process with the help of a test example in Figs. 10-16. After taking a decision from the video sequence Fig. 10, we perform manual decision stum as shown in Fig. 11. Based on the visual result we perform final selection of stump as shown in Fig. 12. Next we take an example of audio sample representing a gunshot as shown in Fig. 13. A portion of this gunshot audio based on the decision stump is extracted which is shown in Fig. 14. As a final step we take a normal audio signal such as from a train station scenario and merge the gunshot audio which this normal audio sequence.

Figure 17 through 19 demonstrates that the inclusion of audio event and overall anomaly detection based on video and audio data helps to achieve faster and reliable anomaly detection.

There can be scenarios where the visual data can temperately be missing, however, a high score of audio event is available. In such cases, the proposed algorithm correctly identifies anomaly which would not be possible with anomaly detection by using video cues only.

As mentioned earlier, performance of the proposed audio-visual anomaly detection framework is tested with two different audio classifiers

SVM based classifier

autoregressive network based audio classifier

Detection results of the proposed audio-visual anomaly detection framework with two different audio classifiers are summarized in Table 3.

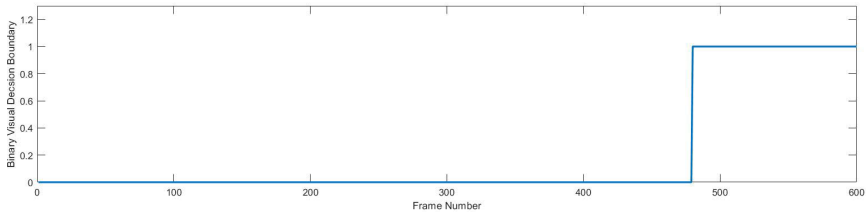

FIGURE 10. Decision from video signal.

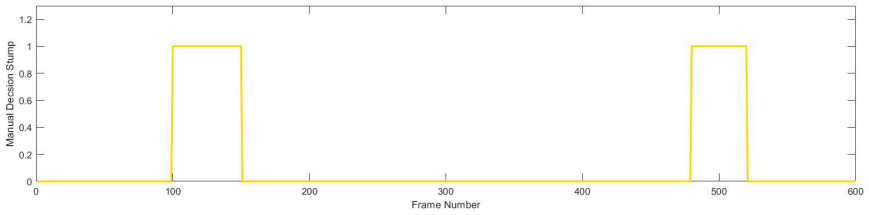

FIGURE 11. Manual stamping.

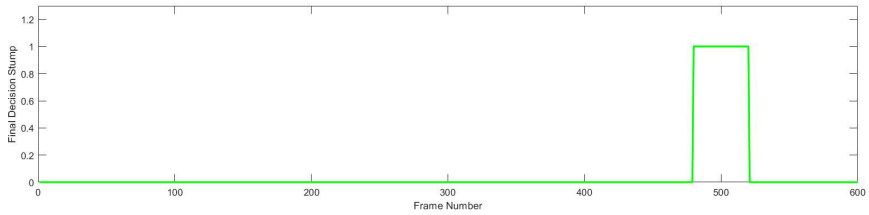

FIGURE 12. Final selection stump.

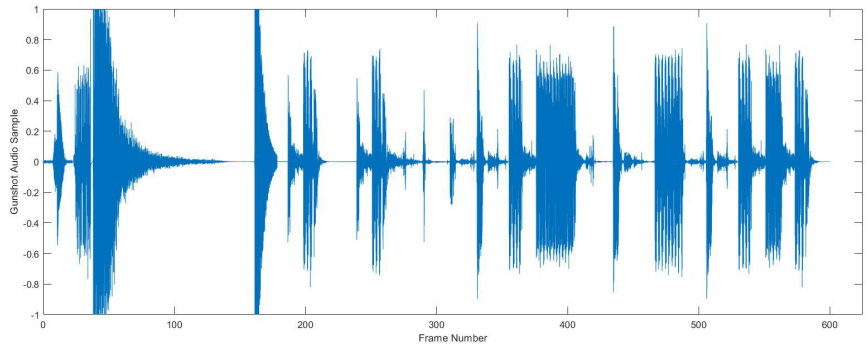

FIGURE 13. Gunshot audio sample.

If we compare Table 3 with the Table 1, we can see that that the proposed algorithm has out performed the algorithms based on video based anomaly detection only. There is a clear improvement in overall accuracy of the proposed algorithm. In total, false positive and false negative rate has been reduced while true positives has been increased and the overall accuracy has been improved.

It can also be seen in the Table 3 that the performance of the proposed anomaly detection framework with the SVM audio classifier remains almost the same even if we replace the audio classifier with the more complicated deep learning based autoregressive network audio classifier. The is because the audio classifier assists video classifier to improve the overall anomaly detection accuracy by detecting audio anomalies when video data is missing or noisy. Therefore, we can conclude that the proposed anomaly detection framework even with a simple SVM based classifier is enough to get highly accurate anomaly detection results.

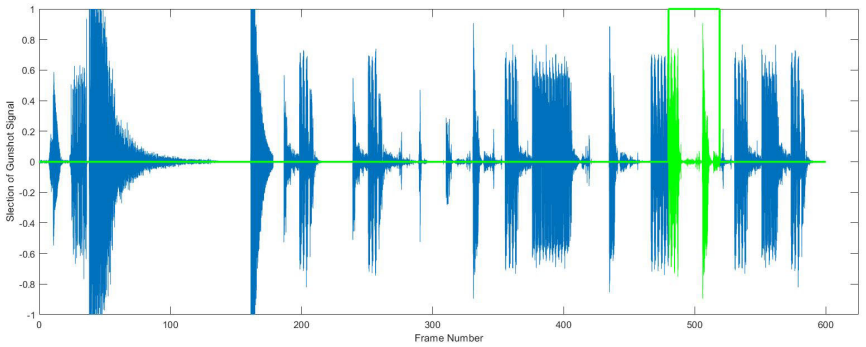

FIGURE 14. Selection of gunshot based on decision stump.

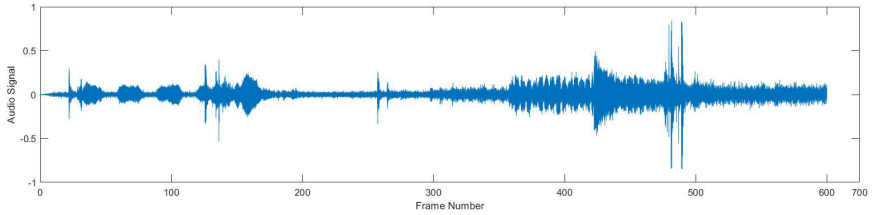

FIGURE 15. Audio signal (Crowd on a train station).

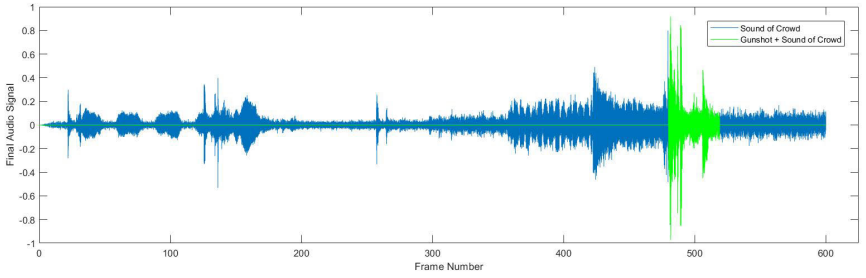

FIGURE 16. Final merger of gunshot with the normal sound clip.

TABLE 3. Comparison of audio-visual anomaly detection by using two different audio classifiers.

Video Sequences	Audio-Visual Anomaly Detection deep learning based audio classifier [19]				Audio-Visual Anomaly Detection with with SVM audio classifier			
	TP	FP	FN	Accuracy	TP	FP	FN	Accuracy
UMN Dataset Seq 1	100	2	4	0.94	102	1	3	0.96
UMN Dataset Seq 2	76	2	4	0.92	76	2	4	0.93
UMN Dataset Seq 3	80	4	5	0.90	81	2	6	0.91
UMN Dataset Seq 4	64	1	2	0.95	63	3	1	0.94
UMN Dataset Seq 5	81	1	3	0.95	80	2	3	0.94
Total	401	9	19	0.93	402	10	17	0.94

Plot for the true positive against the false positive is shown in the Figure. 20. This shows a very good true positive rate compared to the false positives.

Comparison of the overall accuracy of the proposed method with the existing techniques is presented in Table 4. The proposed method exploits audio and video cues, therefore, for comparison, five sequences of UMN dataset combined with the audio data are used to calculate the overall accuracy. Whereas, all other techniques are video based anomaly detection techniques, therefore, accuracy of video

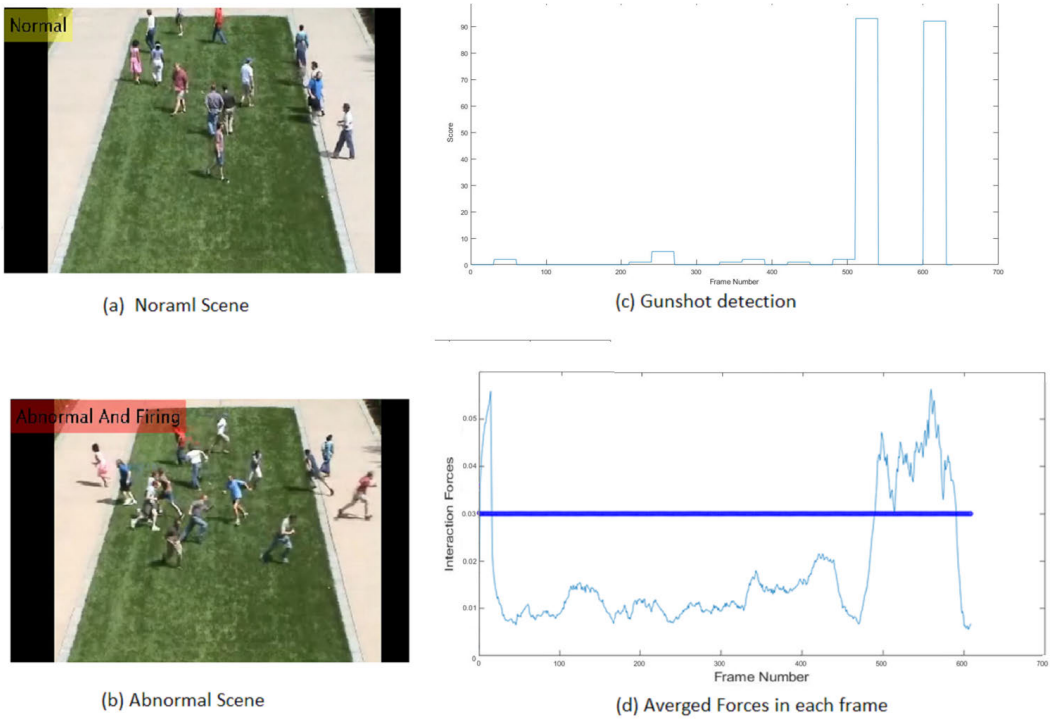

FIGURE 17. Audio video event classification sequence 1 (a) Normal scene detection (b) Abnormal scene focused with gunshot detection (c) Gunshot detection score two peaks shows two shots (d) Interaction force with threshold declaration.

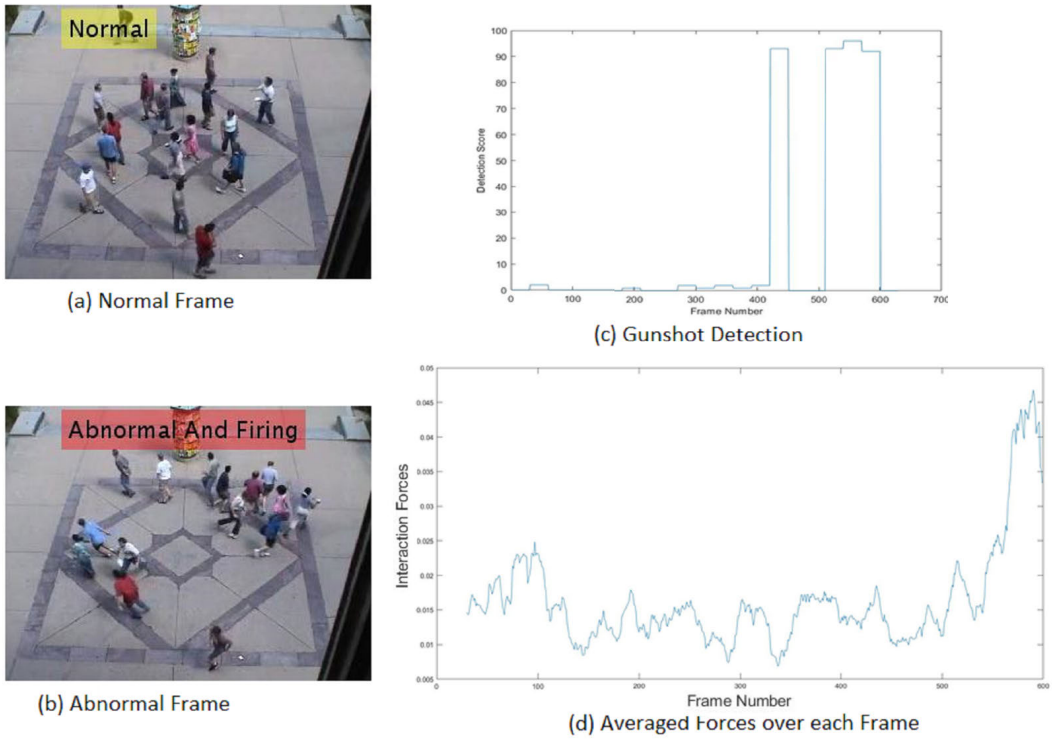

FIGURE 18. Audio video event classification sequence 4 (a) Normal scene detection (b) Abnormal scene focused with gunshot detection (c) Gunshot detection score four peaks shows three shots with in three seconds (d) Interaction forces plot for each frame seconds (d) Interaction forces plot for each frame.

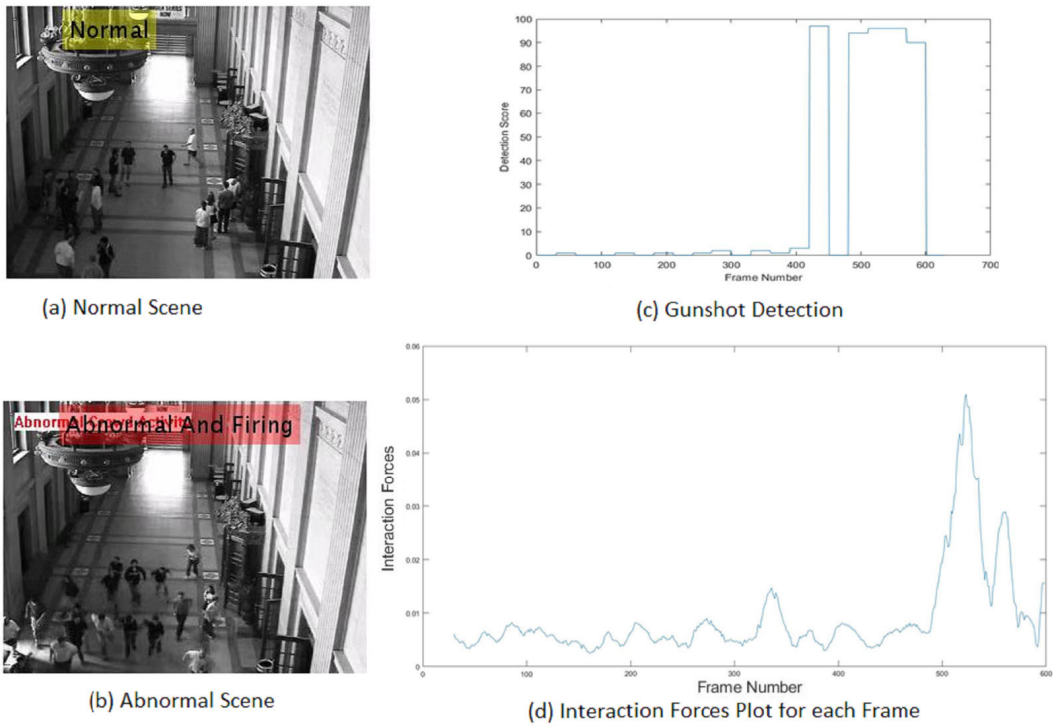

FIGURE 19. Audio video event classification sequence 5 (a) Normal scene detection (b) Abnormal scene focused with gunshot detection (c) Gunshot detection score for almost five peaks shows four continuous shots (d) Averaged forces plot against each frame.

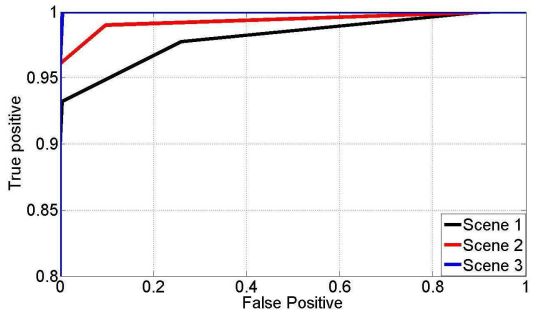

FIGURE 20. True positives vs false positives.

based methods shown in Table. 4 is evaluated by using video sequences of UMN dataset. This can again be seen in Table 4 that the proposed method has higher accuracy compared to the existing techniques for anomaly detection.

D. DISCUSSIONS

In the proposed framework we have provided a solution for anomaly detection in an outdoor environment where surveillance cameras are usually installed to monitor activities of general public. Existing techniques for anomaly detection rely on video data only to automatically detect anomalies in the surveillance area. However, most of the abnormal activities in surveillance area cause anomalies in the video as well as audio. Although the audio and video anomalies are different streams, but one can cause, the other to happen. We have proved with the proposed anomaly detection framework that combining both modalities result in better detection accuracy as can be seen in Table 4. To prove this hypothesis, we presented three different types of evaluation result

1. The proposed technique with video only modality is compared with other state-of-the-art in the area of anomaly detection for video surveillance and it is shown Table 1 that the proposed framework even without audio modality performs slightly better than the other methods.
2. The proposed technique with audio only modality is compared with the other state-of-the-art in the area of audio based anomaly detection and it is shown in Table 2 that the audio anomaly detection method used in the proposed framework performs equivalent to the other deep learning based audio anomaly detection methods.
3. As mentioned in the Introduction section of the paper, in best of our knowledge, there is no algorithm in the literature which uses both audio and video modalities to detect abnormal activities in scenes which are usually monitored by cameras only. The proposed framework is the first effort to prove that the audio modality combined with the video modality can perform better to accurately detect abnormal activities. Therefore, in the third and final comparison, we have shown and proved that the proposed audio-video based anomaly detection framework gives better results. These comparisons are presented in Table 3 and 4, whereas, the supremacy of the proposed framework can also be seen in Figs. 17-20.

TABLE 4. Accuracy comparison between proposed and existing methods.

Method	Accuracy
Social Force Model based [24]	0.84
Chaotic Invariants [26]	0.87
Anomaly Detection With Compact Feature Sets [49]	0.91
Sparse Reconstruction [50]	0.87
Deep model [51]	0.91
Spatiotemporal Anomaly Detection Using Deep Learning [52]	0.93
Purposed framework	0.95

IV. CONCLUSION

An efficient anomaly detection technique has been presented for the detection of abnormal behaviours among crowds in outdoor environment. The developed technique is based on audio and visual data. This has been shown that the video based event detection combined with the audio classifications has helped to improve the anomaly detection accuracy. The comparison of the proposed algorithm with the existing ones show that the proposed method has superior performance in terms of improved accuracy and reduced number of false detections.

REFERENCES

[1] H. Lin, J. D. Deng, B. J. Woodford, and A. Shahi, "Online weighted clustering for real-time abnormal event detection in video surveillance," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 536–540.
[2] D. R. Reddy, "Speech recognition by machine: A review," *Proc. IEEE*, vol. 64, no. 4, pp. 501–531, Apr. 1976.
[3] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Commun.*, vol. 56, pp. 85–100, Jan. 2014.
[4] R. Cai, L. Lu, and A. Hanjalic, "Co-clustering for auditory scene categorization," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 596–606, Jun. 2008.
[5] H. Malik, "Acoustic environment identification and its applications to audio forensics," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 11, pp. 1827–1837, Nov. 2013.
[6] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento, "A real-time text-independent speaker identification system," in *Proc. 12th Int. Conf. Image Anal. Process.*, Sep. 2003, pp. 632–637.
[7] Z. Saquib, N. Salam, R. P. Nair, N. Pandey, and A. Joshi, "A survey on automatic speaker recognition systems," in *Signal Processing and Multimedia*. Berlin, Germany: Springer, 2010, pp. 134–145.
[8] V. Carletti, P. Foggia, G. Percannella, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance using a bag of aural words classifier," in *Proc. 10th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, Aug. 2013, pp. 81–86.
[9] W. Zajdel, J. D. Krijnders, T. Andringa, and D. M. Gavrila, "CASSANDRA: Audio-video sensor fusion for aggression detection," in *Proc. IEEE Conf. Adv. Video Signal Based Surveill.*, Sep. 2007, pp. 200–205.
[10] J.-L. Rouas, J. Louradour, and S. Ambellouis, "Audio events detection in public transport vehicle," in *Proc. IEEE Intell. Transp. Syst. Conf.*, Sep. 2006, pp. 733–738.
[11] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Proc. IEEE Conf. Adv. Video Signal Based Surveill.*, Sep. 2007, pp. 21–26.
[12] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 381–396, Mar. 2002.
[13] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 1, pp. 279–288, Jan. 2016.
[14] W. Shuiping, T. Zhenming, and L. Shiqiang, "Design and implementation of an audio classification system based on SVM," *Procedia Eng.*, vol. 15, pp. 4031–4035, Jan. 2011.
[15] J. J Huang and J. J. A. Leanos, "AclNet: Efficient end-to-end audio classification CNN," 2018, *arXiv:1811.06669*. [Online]. Available: http://arxiv.org/abs/1811.06669
[16] J. Pons and X. Serra, "Randomly weighted CNNs for (Music) audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 336–340.
[17] Z. Mnasri, S. Rovetta, and F. Masulli, "Audio surveillance of roads using deep learning and autoencoder-based sample weight initialization," in *Proc. IEEE 20th Medit. Electrotech. Conf. (MELECON)*, Jun. 2020, pp. 99–103.
[18] D. Oh and I. Yun, "Residual error based anomaly detection using auto-encoder in SMD machine sound," *Sensors*, vol. 18, no. 5, p. 1308, Apr. 2018.
[19] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 271–275.
[20] E. Rushe and B. M. Namee, "Anomaly detection in raw audio using deep autoregressive networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3597–3601.
[21] R. Mehran, B. E. Moore, and M. Shah, "A streakline representation of flow in crowded scenes," in *Comput. Vision—ECCV*. Berlin, Germany: Springer, 2010, pp. 439–452.
[22] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *Int. J. Comput. Vis.*, vol. 74, no. 1, pp. 17–31, Apr. 2007.
[23] S. Tariq, H. Farooq, A. Jaleel, and S. M. Wasif, "Anomaly detection with particle filtering for online video surveillance," *IEEE Access*, early access, Jan. 25, 2021, doi: 10.1109/ACCESS.2021.3054040.
[24] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 935–942.
[25] S. Ali and M. Shah, "A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–6.
[26] S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2054–2060.
[27] F. Lekien and J. Marsden, "Tricubic interpolation in three dimensions," *Int. J. Numer. Methods Eng.*, vol. 63, no. 3, pp. 455–471, 2005.
[28] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
[29] X. Wang, X. Ma, and W. E. L. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 539–555, Mar. 2009.
[30] S. Ali and M. Shah, "Floor fields for tracking in high density crowd scenes," in *Computer Vision—ECCV*. Berlin, Germany: Springer, 2008, pp. 1–14.
[31] 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.
[32] R. Raghavendra, A. Del Bue, M. Cristani, and V. Murino, "Optimizing interaction force for global anomaly detection in crowded scenes," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 136–143.
[33] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE Int. Conf. Neural Netw.*, vol. 4, Nov. 1995, pp. 1942–1948.
[34] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "Identifying human behaviors using synchronized audio-visual cues," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 54–66, Jan. 2017.

[35] V.-T. Vu, F. Bremond, G. Davini, M. Thonnat, Q.-C. Pham, N. Allezard, P. Sayd, J.-L. Rouas, S. Ambellouis, and A. Flancquart, "Audio-video event recognition system for public transport security," in *Proc. IET Conf. Crime Secur.*, Jun. 2006, pp. 414–419.
[36] Q.-C. Pham, A. Lapeyronnie, C. Baudry, L. Lucat, P. Sayd, S. Ambellouis, D. Sodoyer, A. Flancquart, A.-C. Barcelo, F. Heer, F. Ganansia, and V. Delcourt, "Audio-video surveillance system for public transportation," in *Proc. 2nd Int. Conf. Image Process. Theory, Tools Appl.*, Jul. 2010, pp. 47–53.
[37] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 51, pp. 4282–4286, May 1995.
[38] L. Wang, "Support vector machines: Theory and applications," in *Studies in Fuzziness and Soft Computing*, vol. 302. Berlin, Germany: Springer, Jan. 2005.
[39] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," 2014, *arXiv:1409.7787*. [Online]. Available: http://arxiv.org/abs/1409.7787
[40] S. M. Naqvi, L. Mihaylova, and J. A. Chambers, "Multi-target tracking by using particle filtering and a social force model," in *Proc. 17th Int. Conf. Inf. Fusion (FUSION)*, Jul. 2014, pp. 1–6.
[41] A. Ur-Rehman, S. M. Naqvi, L. Mihaylova, and J. A. Chambers, "Multi-target tracking and occlusion handling with learned variational Bayesian clusters and a social force model," *IEEE Trans. Signal Process.*, vol. 64, no. 5, pp. 1320–1335, Mar. 2016.
[42] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Computer Vision—ECCV*. Berlin, Germany: Springer, 2004, pp. 25–36.
[43] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 221–255, Feb. 2004.
[44] Z. Liu, Y. Wang, and T. Chen, "Audio feature extraction and analysis for scene segmentation and classification," *J. VLSI signal Process. Syst. for signal, image video Technol.*, vol. 20, no. 1, pp. 61–79, Oct. 1998.
[45] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the Cuidado project," Institut de Recherche et Coordination Acoustique/Musique, Paris, France, CUIDADO Ist Project Rep., Jan. 2004.
[46] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *J. Comput. Sci. Technol.*, vol. 16, no. 6, pp. 582–589, Nov. 2001.
[47] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proc. DCASE Workshop Detection Classification Acoustic Scenes Events*, Munich, Germany, Nov. 2017, pp. 1–8. [Online]. Available: https://hal.inria.fr/hal-01627981
[48] R. Mehran, A. Oyama, and M. Shah. *Unusual Crowd Activity Dataset of University of Minnesota*. Accessed: 2009. [Online]. Available: http://mha.cs.umn.edu/movies/crowdactivity-all.avi
[49] R. Leyva, V. Sanchez, and C.-T. Li, "Video anomaly detection with compact feature sets for online performance," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3463–3478, Jul. 2017.
[50] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proc. CVPR*, Jun. 2011, pp. 3449–3456.
[51] Y. Feng, Y. Yuan, and X. Lu, "Learning deep event models for crowd anomaly detection," *Neurocomputing*, vol. 219, pp. 548–556, Jan. 2017.
[52] R. Nawaratne, D. Alahakoon, D. De 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.

ATA-UR REHMAN received the B.Eng. degree in electronic engineering from Air University, Islamabad, in 2006, and the M.Sc. (Hons.) and Ph.D. degrees from Loughborough University, Loughborough, U.K., in 2010 and 2014, respectively. He then joined the University of Engineering and Technology, Lahore, as a Lecturer, in 2007. He moved to Loughborough University, in 2009. He worked as a Postdoctoral Research Associate with the Department of Automatic Control and Systems Engineering, University of Sheffield, from December 2013 to January 2016. Since January 2016, he has been working as an Assistant Professor with the University of Engineering and Technology, Lahore. His main research interests include computer vision, machine learning, pattern recognition, and multi-target tracking.

HAFIZ SAMI ULLAH received the B.S. degree in electrical engineering from the University of Engineering and Technology, Lahore, Pakistan, in 2018. He is currently pursuing the M.Sc. degree with the Department of Electronic Systems Engineering, University of Regina, Canada.

HAROON FAROOQ received the Ph.D. degree in electrical engineering from Glasgow Caledonian University, U.K., in 2012. He is currently an Assistant Professor with the Department of Electrical Engineering, Rachna College of Engineering and Technology (RCET, Gujranwala), University of Engineering and Technology, Lahore, Pakistan. His research interests include power quality, renewable energy systems, electric vehicles, and demand side management.

MUHAMMAD SALMAN KHAN received the B.S. degree (Hons.) in electrical engineering from the University of Engineering and Technology (UET), Peshawar, Pakistan, in 2007, the M.S. degree in electrical engineering from The George Washington University, Washington, D.C., USA, in 2010, and the Ph.D. degree in electrical engineering from Loughborough University, Loughborough, U.K., in 2013. He was a Postdoctoral Fellow with the Department of Electrical Engineering, Universidad de Chile, Santiago, Chile, from 2013 to 2015. Since 2015, he has been an Assistant Professor with the Department of Electrical Engineering, Jalozai Campus, UET Peshawar. He is the Principal Investigator of the HEC-funded projects on Artificial Intelligence in Healthcare, the Intelligent Information Processing Laboratory, and the National Center of Artificial Intelligence, UET Peshawar. His research interests include information processing, pattern recognition, machine learning, big data, and artificial intelligence. In 2019, he was included as a member of the World Health Organization (WHO) Roster of Experts on Digital Health.

TAYYEB MAHMOOD received the B.S. degree in electrical engineering from the University of Engineering and Technology, Lahore, Pakistan, in 2006, and the Ph.D. degree in information and communication engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2013. He is currently working as an Assistant Professor with the Department of Electrical Engineering, University of Engineering and Technology, Lahore. His research interests include low-power computing, the Internet of Things, and Embedded systems.

HAFIZ OWAIS AHMED KHAN received the B.S. degree in electrical engineering from the University of Engineering and Technology, Lahore, Pakistan, in 2018. He is currently pursuing the M.Sc. degree with the Department of Electrical Engineering, Lahore University of Management Sciences, Lahore.