Detection of Abnormal Activity to Alert the Nearby Persons via M-DNN Based Surveillance System

SHANKARGOUD PATIL[1], KAPPARGAON S. PRABHUSHETTY[2]
[1]Department of Electronics and Communication Engineering,
S. G. Balekundri Institute of Technology, Belagavi
Shivabasavnagar, Belagavi, Karnataka, INDIA

[2]Department of Electronics and Communication Engineering,
Veerappa Nisty Engineering College, Hasanapur
Shorapur, Yadgir, Karnataka, INDIA

Abstract: - In today's environment, video surveillance is critical. When artificial intelligence, machine learning, and deep learning were introduced into the system, the technology had progressed much too far. Different methods are in place using the above combinations to help distinguish various wary activities from the live tracking of footages. Human behavior is the most unpredictable, and determining whether it is suspicious or normal is quite tough. In a theoretical setting, a deep learning approach is utilized to detect suspicious or normal behavior and sends an alarm to the nearby people if suspicious activity is predicted. In this paper, data fusion technique is used for feature extraction which gives an accurate outcome. Moreover, the classes are classified by the well effective machine learning approach of modified deep neural network (M-DNN), that predicts the classes very well. The proposed method gains 95% accuracy, as well the advanced system is contrast with previous methods like artificial neural network (ANN), random forest (RF) and support vector machine (SVM). This approach is well fitted for dynamic and static conditions.

Key-Words: - Data fusion, Feature extraction, Modified deep neural network (M-DNN), Video surveillance, Wary activities.

Received: May 25, 2021. Revised: November 20, 2021. Accepted: December 4, 2021. Published: December 20, 2021.

1 Introduction

To improve public safety, surveillance cameras are increasingly being deployed in public spaces such as roadways, crossroads, banks, shopping malls, and so on [1]. Law enforcement agencies' monitoring capabilities, on the other hand, have not kept up. As a result, there is a conspicuous deficit in the use of surveillance cameras, as well as an unworkable camera-to-human-monitor ratio. Detecting abnormal events such as traffic accidents, crimes, or unlawful activity is one of the most important tasks in video surveillance [2]. Anomaly events are uncommon in comparison to usual activities. As a result, developing sophisticated computer vision algorithms for automatic video anomaly detection is a vital requirement to save labor and time loss. The purpose of a realistic anomaly detection system is to detect and communicate activity that deviates from typical patterns in a timely manner, as well as to determine the time window in which the abnormality occurs [3]. Like a way, anomaly detection can be thought of as a high-level video knowledge that separates abnormalities from typical patterns. Once an anomaly has been identified, classification techniques can be used to categorize it into one of the specialized activities.

Research can be carried out at several levels, ranging from preparation stages utilizing mostly image processing techniques to analysis and interpretation. The extraction of prominent features to represent the moving objects in the scene, their classification, tracking of their movements, and behavior analysis [4]. In the advanced system, discover an advanced of data fusion technique for feature extraction. Data fusion is a rather old technique [5]: from the 1970s, when it exploded in popularity in the United States, through the 1990s and into the present, this research subject has remained attractive owing to its polymorphism and practical benefits. In our work, the image which is pre-processed is extracted by fifteen feature

extraction approaches. The one-by-one extracted outcomes are fusion to get a more accurate result as well as this fusion technique is termed as data fusion approach. The main advantage of using fusion technique is detection of aberrant actions in both local and global environments [6]. Moreover, this method delivers "high-resolution" and localized estimations because it directly computes abnormality levels at each local node. The final step is classifier, in surveillance systems, classification is a critical stage [7]. A classifier is typically thought of as a standalone entity that makes a judgement on an issue by producing a factual or binary result, or basically a label. A classifier's goal is to turn feature outputs into useful information so that higher-level decisions may be made [8]. Various intelligent approaches are available for classification. Among them modified-deep neural network (M-DNN) is preferred for the proposed approach since it gives a rapid operation to detect whether the situation is normal or abnormal. If the situation is normal, the system cannot generate any signal, when the condition of the place is changed to abnormal, the advanced system is analyzed to create the alert signal instantly to secure the persons. The advantage of the advanced system is, it works for both dynamic and static environment.

The key objectives of the paper are: (i) Video surveillance system of human activity recognition is created according to the condition of normal and abnormal activity dataset as well as the accuracy of the system is analyzed. (ii) Video clips of the surveillance is segmented and pre-processed to remove the unwanted things, and the outcome of the pre-processing is extracted. In the proposed system fifteen feature extraction are utilized to extract the video clip. With the usage of fifteen feature extraction approach, it is termed as fusion technique. (iii) Data fusion technique gives more accurate result, then the outcome is classified to find the condition of that place. The classification section is carried by modified DNN, which delivers a rapid operation and more accurate outcome. (iv) The proposed system accuracy is analyzed for three sections, (a) for five feature extraction method, (b) for eight feature extraction method, (c) for fifteen feature extraction method. Moreover, the performance of the system is compared to the previous techniques of ANN, SVM and RF.

The remaining part of the paper contains: part 2 presents current research in human activity recognition approaches and part 3 discusses the proposed technique and mathematical expressions. Part 4 proposes detailed results. Finally, the conclusion is provided in part 5.

2 Literature review

Various ideas and techniques are utilized for human activity recognition to find the abnormal event. A few methods are discussed below.

Azar Mahmoodzadeh [9] proposed a technique to recognize the human activity, the technique name was deep belief network classifier. A hybrid technique that extracted the feature from image with the usage of scale invariant feature transform (SIFT), global invariant feature transform (GIFT) and histogram of orientated gradient (HOG). The final step of classification is carried by deep belief network (DBN). Need for large data set owing to the fusion technique, had some problem, therefore bag of work (BoW) technique was preferred for individual feature set extract. PASCAL VOC challenge 2010 dataset was applied to simulate the result.

Nida Khalid and et al. [10] have suggested the stereoscopic action recognition according to the fusion of depth sensor and RGB. Activity of human were tracked by four features, such as 3D Cartesian-plane features, geodesic distance, way-points trajectory generation as well as joints Motion Capture (MOCAP) features. Particle swarm optimization (PSO) was used to optimize the above features, then classifier of neuro fuzzy technique was utilized to classify the features and recognize the activity. Three challenging datasets were used for simulation, the data set name was a UoL (University of Lincoln) 3D social activity dataset, a Nanyang Technological University (NTU) RGB+D dataset and a Collective Activity Dataset (CAD). With the NTU RGB+D dataset, 93.5% accuracy, 92.2% accuracy with the UoL dataset, and 89.6% accuracy with the Collective Activity dataset were attained.

Amna Shifa and et al. [11] have presented the privacy protected approach of muti-level video security system (MuLViS). A Smart Surveillance Security Ontology (SSSO) was linked in to MuLViS for autonomous choosing of privacy matching to recognize the activity. Data which is captured from the videos were protected with various encryption stages. The approach was suited for the protection of surveillance shots, and it may be made GDPR compliant, guaranteeing that legitimate data access respects individuals' privacy rights.

Rashmika Nawaratne and et al. [12] has presented the activity identification of human by the usage of Growing Self Organizing Map (GSOM). The method works on the process of accepting two proven ideas of old-style deep learning, hierarchical and multi-stream learning, enforced to the GSOM architecture of self-structuring to allow for learning from unprocessed video data with a variety of properties, as well as applying a transience

characteristic in the algorithm addresses overfitting and the effects of obsolete data on neural architecture. Three benchmark datasets were utilized to simulate and verified the result of human activity identification.

Ana-Cosmina Popescu and et al. [13] has suggested the approach of automated machine learning to determine the activity of human. Data from the channel of 3D video was linked with independently fleeting the data via 2D CNN. The network outcome of all channels was linked into a class scores array with the usage of fusion approach that gave an accurate outcome from the raw video. Three public datasets as well as a new data set of PRECIS HAR were used by them to test and validate the result. The automated machine learning approach based HAR system proved that the advanced system achieved a high accuracy of 98.43% and delivered a fast process to detect the activity.

Wen Heng and et al. [14] presented the recognition technique of human activity via SVQA methods. The method works on two various tasks such as distorted face identification (DFI) and distorted face verification (DFV), according to these tasks, the method contains two processes namely DFI-SVQA and DFV-SVQA as well as equivalent quality metrics. DFI-SVQA and DFV-SVQA are the core mechanisms used to extract the features and were classified by CNN. Moreover, the approach contains real world data set of video surveillance that was used to analyze face resolution, compression level, and how the characteristics of compressed video surveillance was affected. This approach was more effective at determining the quality of surveillance videos whereas keeping a reasonable time frame in mind.

Wei Liu and et al. [15] suggested the Motion-From-Memory (MFM) to enhance the video surveillance system. MFM maintains the dynamic input order and output features of individual frame. The cost of the system is little high, which was the only drawback of the approach. It was very useful for identification of moving object. In mAP, the functioning of a light-weight MobileNet-based is better than RCNN detector and was improved by 13.93%, making it analogous to that of a strong ResNet-50-based detector. This method was three times faster than the conventional method because it contains 540x960 surveillance footage operating at 33 frames per second on a reasonable commercial GPU (NVIDIA GTX 1080Ti.

Bo-Hao Chen and et al. [16] presented the detection of human beings in multiple surveillance cameras by combining low-rankness and sparsity with circumstantial regularization. The advanced method solves the nearest outlier detection issues via the usage of contextual regularization of low rank limitation. Moreover, the performance was verified by multiple states of background with the utilization of dictionary learning-based sparse. Qualitative and quantitative approaches shown that this method outperformed and also more suitable for robust operation.

Hongzhou Zhang and et al. [17] suggested an entropy framework for evaluating the ambiguity of video surveillance attributions for law enforcement. Within the model, public security risk was separated into three categories based on the source of the threat: fixed targets (or limited zones), video information quality and moving objects. The advanced method was very effective in detecting the objects without any issues.

Roshan Singh and et al. [18] presented the identification of a view-invariant human activity. The frame work of the advanced system presents three modules, they are background subtraction, function extraction and activity is recognized by means of a set of hidden Markov models (HMMs). Uniform rotation local binary patterns, contour-based distance signal feature and optical flow-based motion feature were the methods used for feature extraction. i3DPost multi-view dataset, KTH action recognition dataset, and MSR view-point action dataset were used for activity recognition.

In ref [9] the author explains a deep belief network classifier for recognizing human activities. Yet the method did not give an accurate result since the classifier did not work properly at large data set. In ref [10] the author describes the stereoscopic action identification using a depth sensor and RGB. Sensors are high in cost and the process of the method is slow in operation. In ref [11] the author explained a novel technique of multi-level video surveillance (MuLViS) systems privacy-protected approach. This method detects only the event is normal and abnormal but not to give any alert for saving the living beings. In ref [12] the author explained the approach of identification of human activity through the use of a Growing Self-Organizing Map (GSOM). This method is not suitable for crowded place, since the method provides poor performance for large dataset. In ref [13] the author explains the method of using automated machine learning to assess human activity. This method does not give a rapid operation and cannot detect the activity more accurately. In ref [14], the author elaborates the method of determining human activity using automated machine learning. This method gives an accurate result but it cannot deliver a rapid operation to detect the activity. In ref [15], the author explains the Motion-From-Memory

(MFM) to enhance the video surveillance system. This method provides a fast activity detection with well accuracy but it contains some impacts like not cost effective and error detection is low. In ref [16], combination of low-rankness and sparsity with circumstantial regularization to detect the human beings in multiple surveillance cameras is presented. This method gives the most accurate result still it is not suitable for crowded place activity detection. In ref [17], the author explains an entropy framework for assessing the ambiguity of video surveillance credits for law application. This method gives a well performance but the process of operation is slow. In ref [18], the author determines the recognition of a view-invariant human activity. This method gives an accurate outcome with low cost, but it was not well fitted for detection of crowd activity.

3 Proposed methodology

The majority of visual (or video) surveillance systems use numerous cameras to expand the observation area and provide different views of the situation. Objects are tracked and their activities are recognized using overlapping fields of vision in a multi-camera surveillance system, which are predetermined by a set of actions or scenarios, or even to learn new behavior patterns or information. In the proposed method, the abnormal activity is detected in the crowed place like bank, airport, etc. with the support of video surveillance. Initially the images are collected according to the activity of the place. After the images are collected, the images are pre-processed to get converted as computer language since the images which are obtained directly from the camera are not able to read by the computer. In the proposed system, sobel techniques of edge-based segmentation is preferred. The Sobel edge detection technique for image classification uses the derivatives of Sobel approximation to discover edges. Then the image is extracted at data fusion techniques, here fifteen various extraction techniques are utilized to extract the image. At last, the extracted values are classified in the DNN algorithm to find the normal and abnormal activities of that place. Figure 1 shows the schematic structure of advanced system.

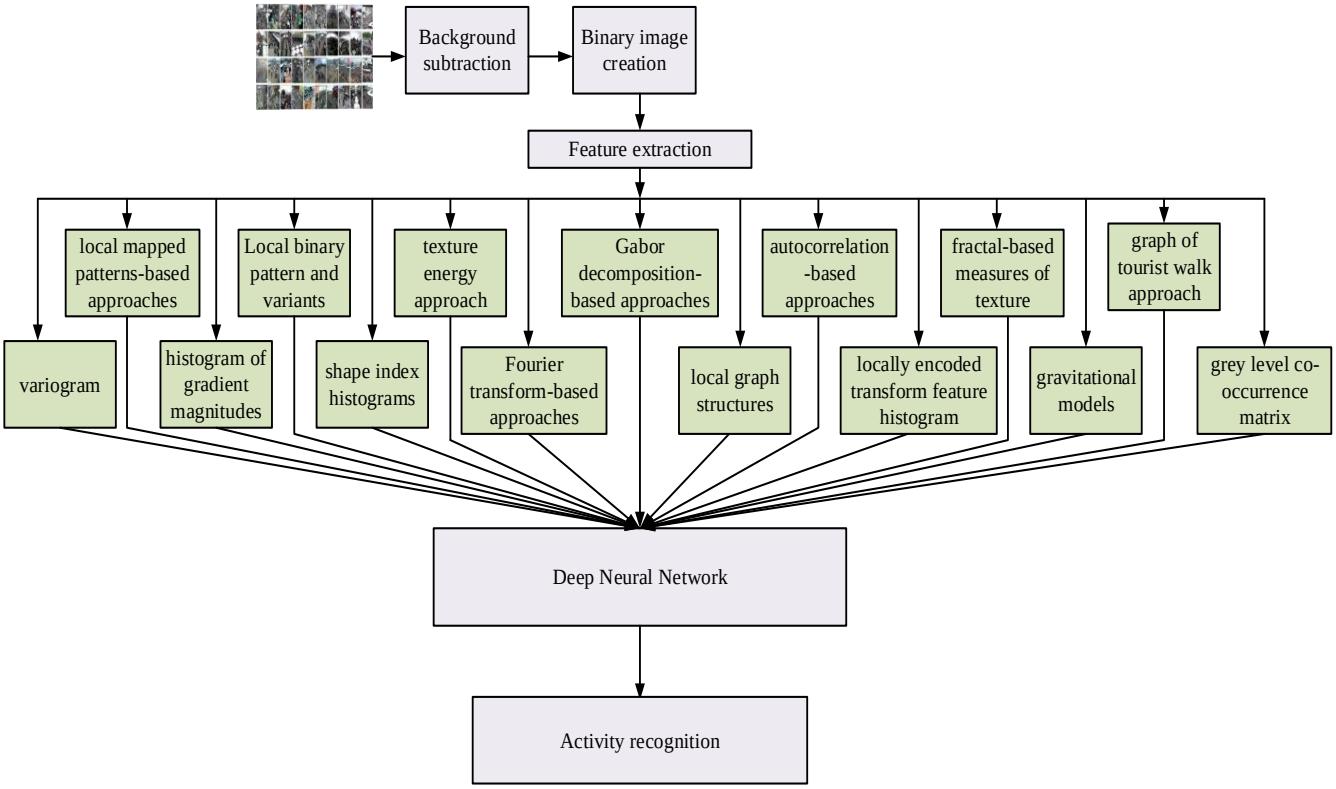


Figure 1 Architecture of proposed method

3.1 Data collection

The first stage in a video surveillance system is to install a CCTV camera and monitor the footage. The data acquisition device is the video camera, and in this case is a digital camera in video capture mode. The video

sequences captured by the video camera are transformed into static color image datasets. Various types of videos are taken from various cameras, which cover the entire surveillance region. Because our implementation uses frames for processing, the videos are transformed into frames. Background subtraction is used to identify the foreground objects in the collected video data.

3.2 Pre-processing

Pre-processing is critical due to noisy, inconsistent, and incomplete data. Preprocessing refers to procedures performed on a video frame at the weakest pace of data. The given video sequence, which is originally in RGB format, captures the input frame. The primary goal of preprocessing is to improve the frame data by augmenting some salient elements in preparation for better processing. Gabor Filter method of pre-processing is used to process the image [19]. Because of its optimal localization, in computer vision and image processing, the Gabor function has been acknowledged as a particularly valuable tool, particularly for analysis of texture.

Gabor filter:

The 2D Gabor filter function is at the heart of Gabor filter-based feature extraction.

$$\psi(x,y) = \frac{f^2}{\pi\gamma\eta} e^{-\left(\frac{f^2}{\gamma^2}x'^2 + \frac{f^2}{\eta^2}y'^2\right)} e^{j2\pi f x'} \quad (1)$$

$$x' = \mathrm{x}\cos\theta + y\sin\theta$$

$$y' = -\mathrm{x}\cos\theta + y\sin\theta$$

The Gabor filter is a frequency domain wave amplified by an origin-centered Gaussian in spatial domain. The central frequency of the filter is denoted by f, the rotation angle is denoted by θ, the sharpness (bandwidth) along the Gaussian major axis is represented by γ, and the sharpness (bandwidth) along the Gaussian minor axis are all represented by η. The aspect ratio of the Gaussian in the given form is η/γ. And in frequency domain, this variable has the following analytical value.

$$\psi(u,v) = e^{-\frac{\pi^2}{f^2}\left(\gamma^2(u'-f)^2 + \eta^2 v^2\right)} \quad (2)$$

$$u' = \mathrm{u}\cos\theta + v\sin\theta$$

$$v' = -\mathrm{u}\cos\theta + v\sin\theta$$

The variable ψ is a singular real-valued Gaussian centered at f in the frequency domain (Eq. (2)). The simple version imposes a series of filters that are self-similar, i.e., rotated and scaled analogues of one another (Gabor wavelets), irrespective of orientation θ and frequency f.

Gabor features, also known as Gabor banks, Gabor jets, or multi-resolution Gabor features, are created by combining the feedbacks of Gabor filters in (1) or (2) and applying several filters at different orientations θ_n and frequencies f_m. In this example, frequency correlates to scale information and is hence derived from it.

$$f_m = k^{-m} f_{max}, m = \{0, \ldots\ldots, M-1\} \quad (3)$$

where, $f_0 = f_{max}$ represents highest frequency desired, f_m is the m^{th} frequency and $k > 1$ represents the frequency scaling factor. The orientation is expressed as,

$$\theta_n = \frac{n2\pi}{N}, \qquad n = \{0, \ldots\ldots, N-1\} \quad (4)$$

where, N is number of orientations and θ_n is the n^{th} number of orientations. Exponential spacing is used to select the filter bank scale as well as linear spacing is selected for orientation. The improved fram image is then segmented by the edge based segmentation.

Edge based segmentation:

It is the technique of splitting an image into sections with different texture, color, brightness, grey level and contrast attributes. The edge of a portion of the image is calculated using discontinuity calculations. Segmentation contains two main operation that are edge detection and linking. Active contour models have been used to produce a number of ways for segmenting mammograms. Under edge detection, sobel edge detection [20] is available. It is one of the greatest edge detection approaches because it is a smaller amount sensitive to noise image. As a result, mathematical gradient estimates must be used. A pair of 3×3 convolution kernels make up the sobel operator. One kernel is simply 90° rotated from the other. The kernels are structured for running both horizontally and vertically that is corresponding to pixel network, moreover each kernel is perpendicular to two orientations. For the process, the kernels are applied discretely in the image which is given as input, to generate discrete observation of apiece orientation's gradient component. All of them are converted together to get the complete magnitude of each point gradient and its orientation. The magnitude of the gradient is expressed as

$$|G| = \sqrt{G_x^2 + G_y^2}$$

Normally, the magnitude of approximate image is measured as,

$$|G| = |G_x| + |G_y|$$

The position of the edge's orientation (relative to the pixel grid) that causes the spatial gradient is determined by:

$$\partial = \arctan(G_x/G_y) \quad (5)$$

At the end of the process, the images are smoothened. After pre-processing stage, feature extraction is carried out.

3.3 Feature extraction

Feature extraction is one of the amplitude reducing techniques. Raw input is split as well as reduced the amplitude to create several clusters, this helps to make the process very easier. The key objective of the input data set is, having a high number of variables. Due to high variables, several mathematical resources are needed to route them. Therefore, features of the images are extracted to get a finest data to combine and select the variable [21]. The features of images are simple to extract as well as to find the originality and accuracy of the original input data set. In the advanced work, texture extraction is utilized to extract the large raw data set.

Texture based feature extraction

Texture is an important part of human vision, and it's employed in a lot of video surveillance applications. Differentiating between textures is a simple operation for the eyes. Despite this, no specific definition of texture has yet been established. In the advanced system fifteen features extraction method is presented that is termed as data fusion techniques. Each feature extraction method is explained separately in the below section.

Local binary pattern and variants

Each image pixel q_c is represented by a binary sequence in the LBP approach. The latter is defined as the variation between grey - level rating of pixelq_c and the radius R of its circular neighborhood centered at q_c. The code of LBP is measured by,

$$LBP_{P,R}(q_c) = \sum_{P=0}^{P=1} S(x)\, 2^P \quad (6)$$

where, $x = q_P - q_c$ represents the level of intensity among neighboring pixels q_P and central pixelq_c, Inside the circular neighborhood of radius R and P nearby pixels. Moreover, $s(x)$ is

$$S(x) = \begin{cases} 1 & x \geq 0 \\ 0 & otherwise \end{cases} \quad (7)$$

The codes range from 0 to 2^P since each digit of an LBP code is either 0 or 1. The LBP code is stable against every monotonic modification of picture brightness according to the sign function $S(x)$. As a texture descriptor, the histogram of these distinct labels can be used. Thus, the distribution of LBP structures can be used to characterize a texture image, and a picture can be represented by an LBP histogram vector h.

$$h = \sum_{i=1}^{W} \sum_{j=1}^{H} \delta\big(LBP_{P,R}(i,j) - k\big) \quad (8)$$

where $0 \leq k < d = 2P$ represents LBP patterns number, W and H are image dimensions, and δ represents the Heaviside function.

Shape index histograms

Capturing images in our detection system are defined by curvedness and shape index values. The shape index S_l, a numerical measure of the surface's shape at a vertex p, is defined as,

$$S_l(p) = \frac{1}{2} - \frac{1}{\pi} \arctan\left(\frac{k_{max}(p) + k_{min}(p)}{k_{max}(p) - k_{min}(p)}\right) \quad (9)$$

where, k_{max} and k_{min} denote the principal curves of the vertex surface p, with $k_{max} > k_{min}$ is expressed by:

$$k_{max}(p) = H(p) + \sqrt{H^2(p) - K(p)}$$
$$k_{min}(p) = H(p) - \sqrt{H^2(p) - K(p)}$$

where, $k(p)\ and\ H(p)$ are the Gaussian curvatures and mean respectively. The mean curvature Hand Gaussian curvature K for a normal parametric surface mapping x from $\Re 2$ into $\Re 3, x: u\Re 3, u = (u, v)$ is given by,

$$K = (eg - f^2)/(EG - F^2) \quad (10)$$
$$H = (eG - 2fF + gE)/(2\{EG - F^2\}) \quad (11)$$

where, E, F and G are first fundamental coefficients and e, f and g are second fundamental coefficients.

$$E = \|X_u\|^2, F = X_u X_v, G = \|X_v\|^2$$

$$e = \frac{\det(X_{uu} X_u X_v)}{\sqrt{EG - F^2}}, f = \frac{\det(X_{uv} X_u X_v)}{\sqrt{EG - F^2}},$$

$$g = \frac{\det(X_{vv} X_u X_v)}{\sqrt{EG - F^2}} \quad (12)$$

Also, with definition of S_l in equation (9), all shapes can be mapped onto the interval $S_l = [0,1]$. Except for the planar shape, every different surface shape resembles to a different value of S_l. Because $k_{max} = k_{min} = 0$, the shape index of vertices on a planar surface is uncertain.

Texture energy approach

The application of simple filters to digital photographs is used in this feature extraction method. It consists of two steps. To create twenty-five 3×3 or 5×5 masks, numerous 1D arrays are convolved together in a combinatorial method. The texture field is then convolved with the latter to stress its microstructure. This yields a picture from which the microstructure's energy can be calculated. Second, large windows are used to obtain macro-statistic features. The identification of 1D array laws are,

- Edge E5 = [-2 -1 0 1 2]
- Spot S5 = [1 0 -2 0 1]

- Level L5 = [1 6 4 6 1]
- Ripple R5 = [-1 4 6 4 -1]
- Wave W5 = [1 -2 0 2 -1]

A set of attitude filters that are computed using a correlation matrix of basis filters. Local statistics (texture energy measures) are approximated at the output of this similar filter bank.

Fourier transform-based approaches

The picture I of size $H \times W$ under study is decomposed into its frequency components using a 2D discrete Fourier transform in the Fourier transform-based techniques.

$$\mathcal{F}(u,v) = \sum_{n=1}^{W} \sum_{m=1}^{H} I(n,m) exp\left(-j\pi\left(\frac{un}{W} + \frac{vm}{H}\right)\right) \quad (13)$$

where, u and v are the frequency of horizontal and vertical plane. The magnitude and phase, as well as the real and imaginary portions, can all be retrieved. The premise behind feature extraction is that spatial edges have a low frequency in one direction but numerous frequencies in other direction. The Fourier transform evaluates the average of the image for zero frequencies ($u = v = 0$). Typically, the Fourier transform output is presented as an amplitude spectrum matching to the complex values' modulus. For zero frequencies (u = v = 0), the Fourier transform calculates the image's average. The Fourier transform data is directly presented as an amplitude spectrum that matches the modulus of the complex values. The idea behind texture feature extraction is to represent the image as a weighted blend of vertical and horizontal sinusoids using the Fourier transform, each with its own frequency (u, v): the image is approximated using aa addition of sinusoidal plane waves of various frequencies.

Gabor decomposition-based approaches

Gaussian windowing in the 2D Fourier domain and continuing 4D filters most likely started the directional decompositions for image analysis. The consideration of the human visual system is typically linked to the employment of Gabor's wavelets for the image. Gabor wavelets are made up of isotropic Gaussian windowing of a complex plane wave in the direction.

$$\theta \ \psi^{\theta}(x) = \frac{e^{-\|x\|^2/2}}{2\pi} \ e^{-j(x^T \omega_0)} \quad (14)$$

where, $\omega_0 = F[\cos\theta; \sin\theta]^T$

The decomposition is determined by the number of fixed and evenly distributed orientations K in $[0, \pi]$.

$$\theta \epsilon \Theta = \left\{\frac{k\pi}{K}; \quad \theta \le k \le K\right\}$$

Then scalar product any real 2D signal (x) with the following atoms to decompose it.

$$\left\{\psi_{j,u}^{\theta}(x) = 2^{-j}\psi^{\theta}\left(2^{-j}(x-u)\right)\right\}_{\theta \epsilon \Theta, j \epsilon z, u \epsilon R^2} \quad (15)$$

Local graph structures

A dominant set for a graph $G = (V, E)$ is a subset D of V such that every vertex not in D is connected to at least one member of D by at least one edge. The number of vertices in a minimal dominating set for G is known as the domination number. LGS works with a pixel's six neighbors. Starting in the left area of the target pixel C, we move anticlockwise and assign a binary value equal to 1 on the edge connecting the two vertices if a neighbor pixel has a higher grey value than the target pixel (or the same grey value), otherwise assign a value equal to 0.Pause the target pixel C after finishing the left section of the graph, then travel horizontally (clockwise) to the right section of the graph and repeat the process until we reach the goal pixel C. A binomial weight 2^p is provided to each sign $s(g_d - g_n)$ to obtain the LGS for pixel $(x_d - y_d)$. Those binomial weights are added together.

$$LGS(x_d, y_d) = \sum_{k=0}^{7} s(g_d - g_n) 2^p \quad (16)$$

where, $s(x) = \begin{cases} 1 & x \ge 0 \\ 0 & x < 0 \end{cases}$

$p = 7. \quad p = 7,6, \ldots 0$

Variogram

A semi-variogram is a graphical depiction of a set of data's geographic variability. The variogram function $2\gamma(h)$, which is the mathematical prediction of the absolute deviations between two random variables separated by a distance h, can be used to calculate the relationship between two pixels. $Z(x)$ represents the value of the regionalized variable at point x, while $Z(x + h)$ represents the value at $x + h$. The semivariogram function is affected by the sample distance h as well as the position x. The fundamental hypothesis, which states that such variance of the difference in the two sample points depends only on the distance between them, must be adopted for the variogram to be based entirely on the distance between the sampling units (h).

$$2\gamma(h) = E\{[Z(x) - Z(x+h)]^2\} \quad (17)$$

The experimental semivariogram is defined for continuous variables as half a percent absolute difference between values separated by a certain lag, where lag is a vector in both direction and distance. It is calculated using equation 17, in which , $Z(x +$

$h)$ is the value of point $x + h$, (h) is the semivariance estimator for each distance h, $Z(x)$ is the value of the regionalized variable at point x, and $N(h)$ is the number of pairs of points separated by the distance h.

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x) - Z(x+h)]^2 \qquad (18)$$

The semi-variogram's graphical representation, spatial variance against distance h, allows estimating the variance number of distinct permutations of adjacent pixels. Three factors define the semi-variogram: sill (σ^2), range (φ), and nugget effect (T^2). The sill parameter indicates the amount of variation explained by the data's spatial structure. It is the plateau achieved by the semivariance values. The range variable indicates the distance at which the semivariogram reaches the sill, indicating when the data no longer correlate. The nugget effect is the result of a combination of collecting mistakes and fluctuations at sizes smaller than that of the distance across observed spots.

Local mapped patterns-based approaches

Each gray-level distribution inside an image neighborhood is assumed to be a local pattern according to the LMP approach. The gray-level variations around the central pixel can be used to illustrate this pattern. Equation 19 will be used to transfer each pattern defined by a $W \times W$ neighbourhood to a histogram bin h_b, where $P(k,l)$ represents the position of each pixel weighting matrix value within a neighborhood, $f_{g(i,j)}$ represents the mapping function, and B indicates the bin number of histograms. The weighted sum of each gray-level differential between nearby pixels and the central pixel, projected onto the [0,1] range by a scaling factor and rounded to B possible bins, is represented by this equation.

$$h_b = round\left(\frac{\sum_{k=1}^{W}\sum_{l=1}^{W}\left(f_{g(i,j)}P(k,l)\right)}{\sum_{k=1}^{W}\sum_{l=1}^{W}P(k,l)}\right)(B - 1) \qquad (19)$$

In LMP based texture analysis, sigmoid curve of the image is computed by,

$$f_{g(i,j)} = \frac{1}{1 + e^{\frac{[A(k,l) - g(i,j)]}{\beta}}} \qquad (20)$$

where, β is the slope curve as well as $[A(k,l) - g(i,j)]$ are the grey level variation in neighborhood center of $g(i,j)$.

Histogram of gradient (HOG) magnitudes

In this project, HOG is used to familiarize the program with the object that has to be detected. Based on vector theory, HOG is divided into three mandatory phases. The photograph in question is first roughed over to separate the object from the background. The magnitude difference in the image is what this method looks for. As a result, we only consider the magnitude part of the vector without the direction for the time being. Equation 21 could be used to calculate the image's magnitude.

$$m(u,v) = \sqrt{f_u(u,v)^2 + f_v(u,v)^2} \qquad (21)$$

The magnitude m of a feature vector at a point is obtained by using equation 21. The variables $f_u(u,v)$ and $f_v(u,v)$ make up $m(u,v)$. In the u-direction, the component is $f_u(u,v)$ and in the v −direction, the component is $f_v(u,v)$. After the system has approximated the position of the object in the image, we teach it to determine the object with greater precision. The direction of the vector is taken into account in the second phase. This is akin to the fine-tuning procedure.

$$\Theta(u,v) = \tan^{-1}\frac{f_v(u,v)}{f_u(u,v)} \qquad (22)$$

By the usage of arctangent to determine an angle as stated in equation 22 using $f_u(u,v)$ and $f_v(u,v)$. A large number of vectors are present at the same time. The object is represented as a continuous variation of gradient dependent on the direction of the vector with the help of vector direction, which is discussed in detail in equation 22. Then divided the image into a large number of pixels since the outcome, which appears as a gradient, is difficult to interpret. The entire result is shown as a histogram after the pixels have been defined. It indicates that pixels in the background accumulate at a far lesser rate than pixels in the item. In conclusion, this software can correctly and precisely discern between object and background by evaluating the histogram output.

Autocorrelation-based approaches

The video of N frame is represented by $V = 1,2,..N$. Calculate the mean value of intensity y_i. At the end of the mean value computation, the time sequence of the frame is identified for the video V that is $y_t = y_1, y_2 \ldots\ldots y_N$. It calculates the correlation between y_t and y_{t+k}, where $k = 0,\ldots,K$ and y_t is a random variable. As a result, autocorrelation for a delayed k is equal to,

$$r_k = \frac{c_k}{c_o} \qquad (23)$$

where, $c_k = \frac{1}{T}\sum_{t=1}^{T-k}(y_t - \bar{y})(y_{t+k} - \bar{y})$ and c_o represent the time series sample variance.

Locally encoded transform feature histogram (LETRIST)

By using a linear combination of many basis filters, any orientation of the first or second derivative of a Gaussian can be generated. Consider the following two-dimensional Gaussian function with circular symmetry:

$$G(x,y;\sigma)=\frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \quad (24)$$

where, σ represents for standard deviation. Gaussian's first and second derivaives at θ is computed by

$G_1^\theta=\cos(\theta)G_x+\sin(\theta)\,G_y$ and

$G_2^\theta=cos^2(\theta)G_{xx}-\sin(2\theta)\,G_{xy}+sin^2(\theta)G_{yy}$

where, G_x and G_y are first and second derivatives along x axis.

Then find the first and second order derivative for the raw image I is $L_x=G_x*I, L_y=G_y*I,\ L_{xx}=G_{xx}*I, L_{yy}=G_{yy}*I$, where, * represents intricacy. For first and second order Gaussian filter at θ is expressed as

$$I_1^\theta=G_1^\theta*I=\cos(\theta)L_x+\sin(\theta)L_y$$
$$=\sqrt{L_x^2+L_y^2}\sin(\theta+\phi) \quad (25)$$

where, $\phi=\arctan(L_x/L_y)$

$$I_2^\theta=\frac{1}{2}\left(L_{xx}+L_{yy}+\sqrt{\left(L_{xx}-L_{yy}\right)^2+4L_{xy}^2}\cos(2\theta-\psi)\right) \quad (26)$$

where, $\psi=\arctan(\frac{2L_{xy}}{L_{yy}-L_{xx}})$

The minimum and maximum value of I_2^θ is

$$I_{2\,max}^\theta=\frac{1}{2}\left(L_{xx}+L_{yy}+\sqrt{\left(L_{xx}-L_{yy}\right)^2+4L_{xy}^2}\right)$$
$$I_{2\,min}^\theta=\frac{1}{2}\left(L_{xx}+L_{yy}-\sqrt{\left(L_{xx}-L_{yy}\right)^2+4L_{xy}^2}\right)$$

Transform Feature Construction

A compact yet discriminative transform feature set can be generated by taking into account the eventual quantization and coding using the resulting extremum responses. The following is the structure of the transform feature set.

1. The first directional Gaussian derivative filter's greatest response, g is

$$g=I_{1\,max}^\theta=\sqrt{L_x^2+L_y^2} \quad (27)$$

2. The extrema difference d is the ratio between the second directional Gaussian derivative filter's maximum and minimum responses.

$$d=I_{2\,max}^\theta-I_{2\,min}^\theta=\sqrt{\left(L_{xx}-L_{yy}\right)^2+4L_{xy}^2} \quad (28)$$

3. Shape index s is defined as the curve of second order quantitative measure which is expressed as,

$$s=\frac{1}{2}-\frac{1}{\pi}arctan\left(\frac{I_{2max}^\theta+I_{2min}^\theta}{I_{2max}^\theta-I_{2min}^\theta}\right)$$
$$=\frac{1}{2}-\frac{1}{\pi}arctan\left(\frac{-L_{xx}-L_{yy}}{\sqrt{\left(L_{xx}-L_{yy}\right)^2+4L_{xy}^2}}\right) \quad (29)$$

Eigen value of $K_1$and K_2 hessian matrix is,

$$H=\begin{bmatrix}L_{xx} & L_{xy}\\ L_{xy} & L_{yy}\end{bmatrix}$$

and

$$s'=\frac{2}{\pi}arctan\left(\frac{k_2+k_1}{k_2-k_1}\right) \qquad (k_1\geq k_2)$$

The first- and second-order differential structures' correlation information is captured by the mixed extrema ratio r. The arctangent function as the rectifier to achieve a low dynamic range output:

$$r=\frac{2}{\pi}\arctan\left(c_i\sqrt{\frac{\left(L_{xx}-L_{yy}\right)^2+4L_{xy}^2}{L_x^2+L_y^2}}\right) \quad (30)$$

where, c represents scale factor at the ration of d and g. Transform features, $F=\{g,d,s,r\}$

A mean value based binary ratio quantizer$Q_1(.)$ for transform features g,d whose values are in a non-negative interval:

$$y=Q_1(x)=\begin{cases}0, & if\ x/m_x>k\\ 1, & otherwise\end{cases}$$

where, $x\in\{g,d\},k$ represents tuning parameter, as well as m_x represents the transform feature mean value of x.

We use a uniform quantizer $Q_2(.)$ for transform features s,r with values in the range [0, 1]:

$$y=Q_2(x)=\begin{cases}0, & x\in[0,\Delta]\\ 1, & x\in[0,2\Delta]\\ \ldots & \\ L-1, & x\in[(L-1)\Delta,1]\end{cases}$$

where, $x\in\{s,r\},L$ represents the level of quantization, as well as $\Delta=1/L$ which is the step of quantization.

ASC (adjacent-scale coding):

The transform features $g,d,$ and s are concurrently encoded over two neighbouring scales, such as $(\sigma_1,\sigma_2),(\sigma_2,\sigma_3)$ and so on. The ASC value of pixel (x,y) in picture I is computed as for the adjacent-scale pair (σ_i,σ_{i+1}) for $i=1,2,\ldots,N$.

$$c_i(x,y)=\sum_{j=1}^{2}(L\_s)^{j-1}y_s(x,y;\sigma_{i+j-1})+(L\_s)^2\sum_{j=1}^{2}(L\_d)^{j-1}y_d(x,y;\sigma_{i+j-1})+(L\_s)^2(L\_d)^2\sum_{j=1}^{2}(L\_g)^{j-1}y_g(x,y;\sigma_{i+j-1})$$

where, $y_s(x,y;\sigma_{i+j-1})$, $y_d(x,y;\sigma_{i+j-1})$ and $y_g(x,y;\sigma_{i+j-1})$ are texture codes

Full-scale coding (FSC):

All N scales $(\sigma_1,\dots,\sigma_N)$ are cooperatively encoded with transform characteristics r. In image I, the FSC value of pixel (x,y) is calculated by

$$cN_\sigma(x,y)=\sum_{j=1}^{N_\sigma}(L_r)^{j-1}\,y_r(x,y;\sigma_j) \quad (31)$$

where, $y_r(x,y;\sigma_j)$ represents the r feature texture code and L_r signifies the level of quantization.
Histogram of N_σ is created based on the image I

$$H_i(l)=\sum_{(x,y)\in I} f(c_i(x,y),l) \quad (32)$$

where $l\in\{0,1,\dots,c_i\}$

$$f(m,n)=\begin{cases}1, & if\ m=n\\ 0 & otherwise\end{cases} \quad (33)$$

$N=3, L_d=L_g=2$, and $L_r=5$ are the values we used in our calculations. H1, H2, and H3 are the 144, 144, and 125-dimensional histograms that result. The feature representations below can be derived from these three histograms:

- LETRIST ASC1: based on the ASC at two adjacent scale, the histogram H1 is $(\sigma_1,\sigma_2)=(1,2)$.
- LETRIST ASC2: H2 histogram based on ASC at two neighboring scales $(\sigma_2,\sigma_3)=(2,4)$.
- LETRIST ASC: The concatenated histogram [H1, H2].
- LETRIST FSC: H3 histogram based on FSC on all three scales $(\sigma_1,\sigma_2,\sigma_3)=(1,2,4)$.
- LETRIST: the [H1, H2, H3] concatenated histogram.

Fractal-based measures of texture

The object's self-similarity is a fundamental characteristic of fractal geometry. A self-similar item is a collection of reduced-scale duplicates of the original object that are identical or nearly identical. The repetitive tendency of similar patterns spread across the entire image can be thought of as an image's texture. The fractal dimension is a useful tool for determining the sophistication or irregularity of the grey level intensity distribution throughout an image pixels region.
The grey level intensity distribution of the input image was first used to determine a set of threshold values based on the parameter (m) value. Using the set of threshold values individually, the input grey scale image was broken into a set of binary images. The input image's grey level intensity distribution was first used to determine a set of threshold values based on the parameter (m) value. Using the set of threshold values individually, the input grey scale image was broken into a collection of binary images.

$$S_t(i,j)=\begin{cases}1 & if\ L_t<P(x,y)<U_t\\ 0 & otherwise\end{cases} \quad (34)$$

where, $S_t(i,j)$ signifies the binary image resultant, U_t and L_t denote the upper and lower values of threshold that is found from the raw image $P(x,y)$.

Gravitational models

Our suggested research includes simulating a simple gravitational system from an image. Each pixel (x,y) is treated as a particle with a mass equal to is its intensity, $m=I(x,y)$, because different pixels have different masses (x,y). Then placed a central mass M in the image's center. This mass functions like a black hole, drawing all particles in its direction. There is no interaction between the particles; just between the central mass and each particle. Depending on its distance and mass from center of the image, each particle moves in a unique way. As a consequence, for each time step t, an image can yield various collapse stages. Each collapse step generates suitable texture pattern depending on the particle placements. Each stage of collapse indicates a milestone in the system's progression. Each stage can be described using complexity descriptors like fractal dimension and lacunarity, resulting in a trademark for the picture in the contracting process. The gravitational system defined for color and grayscale images is described in greater depth.

Graph of tourist walk approach

The tourist walk approach can be thought of as a walker (tourist) who wants to visit N places on a d-dimensional map. These points can be regarded as tourist attractions, and tourists are welcome to visit them. A walker performs a person who chooses walk, where the identity is confined to the memory window = 1 and the tourist follows the deterministic rule of travelling to the closest place not visited in the preceding steps at each discrete time step.

Grey level co-occurrence matrix

The gray-level co-occurrence matrix (GLCM) is a popular way of defining texture by looking at the spatial correlation properties of grayscale. GLCM is created statistically detecting a state in which two pixels on the image are separated by a specific distance and have different grayscale levels. Over the

entire matrix, contrast is an indicator of the intensity between one matrix value and its neighbor. Equation (35) calculates contrast, while equation (36) calculates the correlation of a matrix value to its neighbor. In normalized GLCM, energy is the sum of squared elements. It's calculated using an equation (37). In the GLCM, homogeneity refers to the proximity of the element distribution. It's calculated using an equation (38). In the proposed work, a 90-degree GLCM is used to assess the vertical scratch pattern on the drug user's face.

$$Contrast = \sum_{i,j} |i-j|^2 \, p(i,j)^2 \quad (35)$$

$$Correlation = \frac{(i-\mu i)(j-\mu j)p(i,j)}{\sigma_i \sigma_j} \quad (36)$$

$$Energy = \sum_{i,j} p(i,j)^2 \quad (37)$$

$$Homogeneity = \sum_{i,j} \frac{(j-\mu)p(i,j)}{1+|i-j|} \quad (38)$$

The above fifteen feature extraction techniques are used to extract the features from video clips, then the outcome of fusion approach is classified.

Classification process is more crucial for activity recognition. In the proposed method, modified DNN is preferred for classification technique, which is explained below.

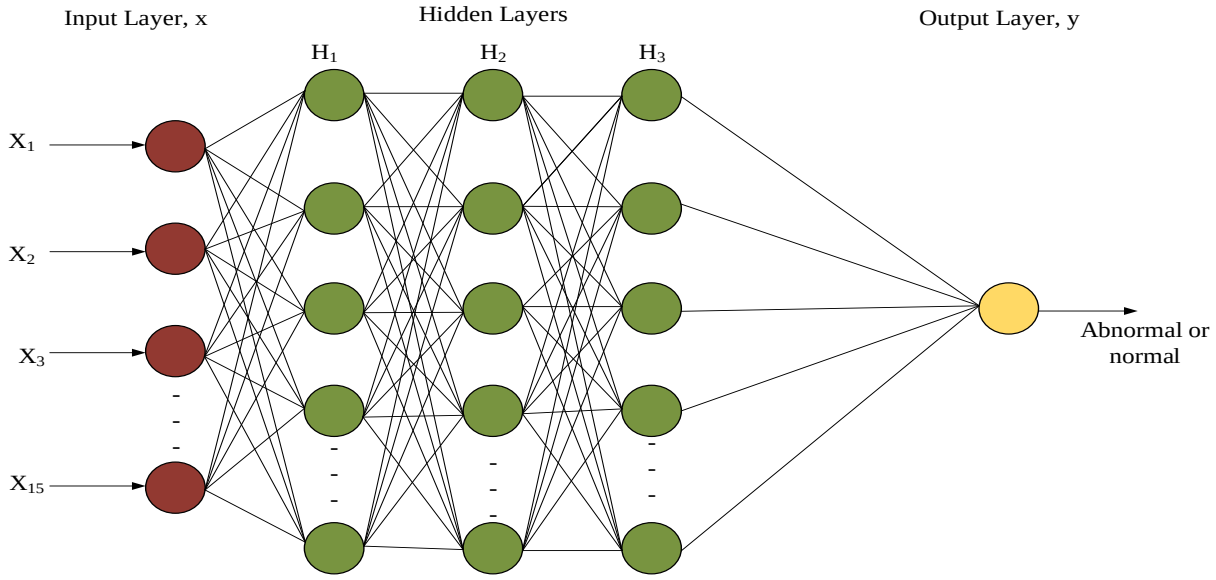


Figure 2 Structure of DNN

Same layer neurons are not allied in the forward propagation but in connection mode neurons are connected fully. Neurons outputs are specified by

$$y_q^{n+1} = \sigma(z) = \sigma\left(\sum_{i=1}^{m} \omega_{iq}^n y_i^n + b_q^{n+1}\right) \quad (39)$$

where, y_q^{n+1} denotes the output of q neuron in n+1 layer, $\sigma(z)$ denotes activation function and sigmoid can be used, b_q^{n+1} signifies the bias of linear relationship, ω_{iq}^n represents the weight among n layer's i neuron and n+1 layer q neuron.

Training the dataset

In the proposed work, weights are used to connect the layers. The weights are optimized by the use of sailfish optimization approach. Sailfish is a fastest swimming fish it can swim at a maximum speed of 100 km/h. The prey of these fish are smaller fishes. Certain behavior of sardines such as acceleration and maneuverability are considered as quite challenging for sailfish while hunting these sardines. To attack sardines sailfish attempt a slashing motion through injuring many sardines, these injured small fishes are detected to easily capture for its food.

The steps of sailfish optimization are as follows,

Step 1. Initialization:

Initiate the weight as an input,

$$weight = \{W_1, W_2 \ldots\ldots W_n\} \quad (40)$$

Step 2. Fitness function:

Select the fitness value, here the error is computed to find the fitness value.

$$fitness\ value = E(\theta) = -\frac{1}{N}\sum_{n}\sum_{q} t_{nq} \log y_{nq} \quad (41)$$

where, θ signifies the parameter of ω and b, t_{nq} actual values of q[th] sample n[th] element, N denotes the number of samples, y_{nq} is the projected value of q[th] sample n[th] element. Outfitting of neurons is

reduced by using a dropout mechanism, thus collapsing the neuron network structure with anyhow removal of neurons. On the other hand, the mechanism enhances the rate of constant learning with respect to the traditional gradient descent method. The optimal parameter of θ is computed by,

$$\begin{cases} g_t = \nabla_\theta E(\theta_{t-1}) \\ m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \\ V_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \\ \widehat{m_t} = \frac{m_t}{1-\beta_1^t} \\ \widehat{V_t} = \frac{v_t}{1-\beta_2^t} \\ \theta_t = \theta_{t-1} - \alpha \frac{\widehat{m_t}}{\sqrt{\widehat{v_t}+\varepsilon}} \end{cases} \quad (41)$$

$$\alpha = \alpha_0 \beta_3^{\frac{epoch-num}{N/batch-size}}$$

where, g_t denotes the parameter gradient, V_t signifies the average movement of the square of the gradient, m_t is the average movement of the gradient, $\widehat{V_t}$ and $\widehat{m_t}$ are corrected quantities, α_0 represents the learning rate initial value, β_1, β_2 and β_3 are exponential decay rates that are in use 0.9,0.999 and 0.95, $epoch - num$ signifies the current training times, $batch - size$ denote the batch processing parameter. The modified DNN is an effective method for classification, it categorizes the videos as either abnormal (fighting, fainting,) or normal (walking, running). An alert message will be raised to secure the person in the event of suspicious conduct.

Step 3. Updating the value:

Update the value to detect the best solution, the activity is analyzed based on the updated value.

$$X^i_{new\_S} = r \times \left(X^i_{elite\_SF} - X^i_{old_S} + AP\right) \quad (42)$$

Where $X^i_{old_S}$ is the sardine current position, $X^i_{elite\_SF}$ is elite sailfish best position formed so far, r is a random number between 0 and 1 as well as AP represent sailfish's Attack Power at each iteration.

$$Ap = A \times \left(1 - (2 \times Itr \times \varepsilon)\right)$$

where, A and ε are coefficients for reducing the value of attack power linearly from A to 0.

Step 4. Termination:

The final step is termination, when the best solution is obtained the process is terminated.

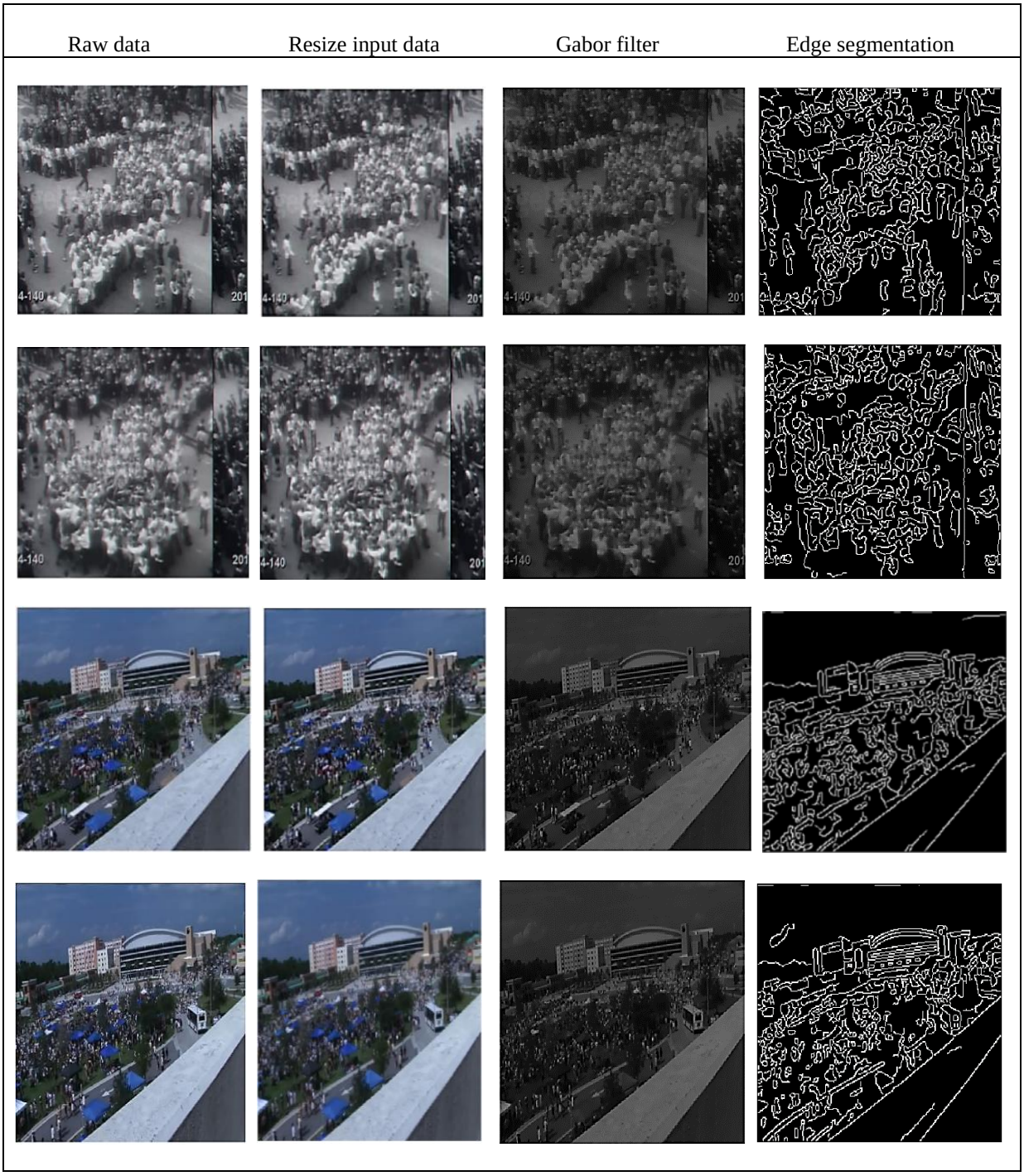
Testing the dataset

For testing purposes, the frames are taken from videos and kept in a single folder. The system classifies the frames as suspicious or normal based on our trained model. In the event of suspicious activity, a notice with the expected class will be forwarded to the appropriate authority. In the data set 80% data are trained 20% data are tested in the classification system.

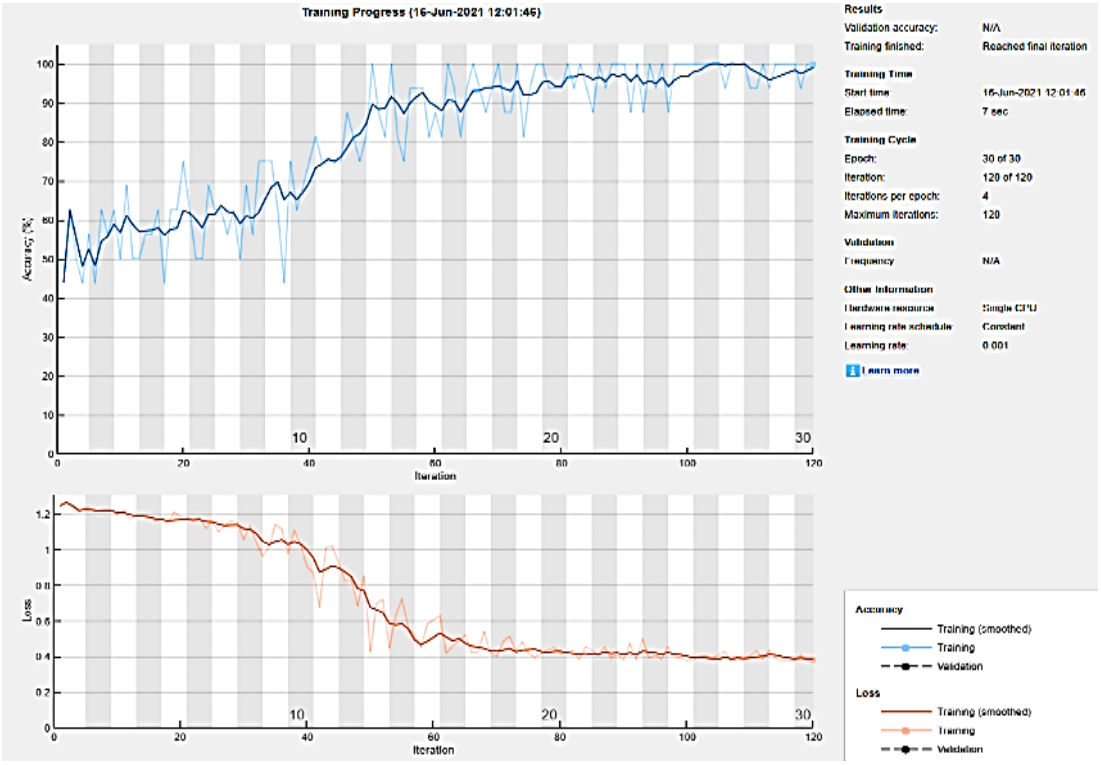
4 Results

The project's goal is to use CCTV footage to track suspicious activity on a campus and to warn security when any abnormal event happens. This was accomplished by extracting features from the frames with the help of M-DNN. The data set is used for the proposed approach is normal and abnormal activity of human beings in the form of video [22]. The video is segmented and pre-processed. The pre-processing clip is passed to the data fusion techniques. Data fusion technique improves the performance of the system and delivers an accurate output, in this technique the output of one extraction is fused to another extraction approach, so the outcome is very accurate. This accurate outcome is classified to identify the present situation of the place, M-DNN is used for classification approach. The proposed approach is well suited for dynamic and static circumstance. Table 1 shows the pre-processing section images.

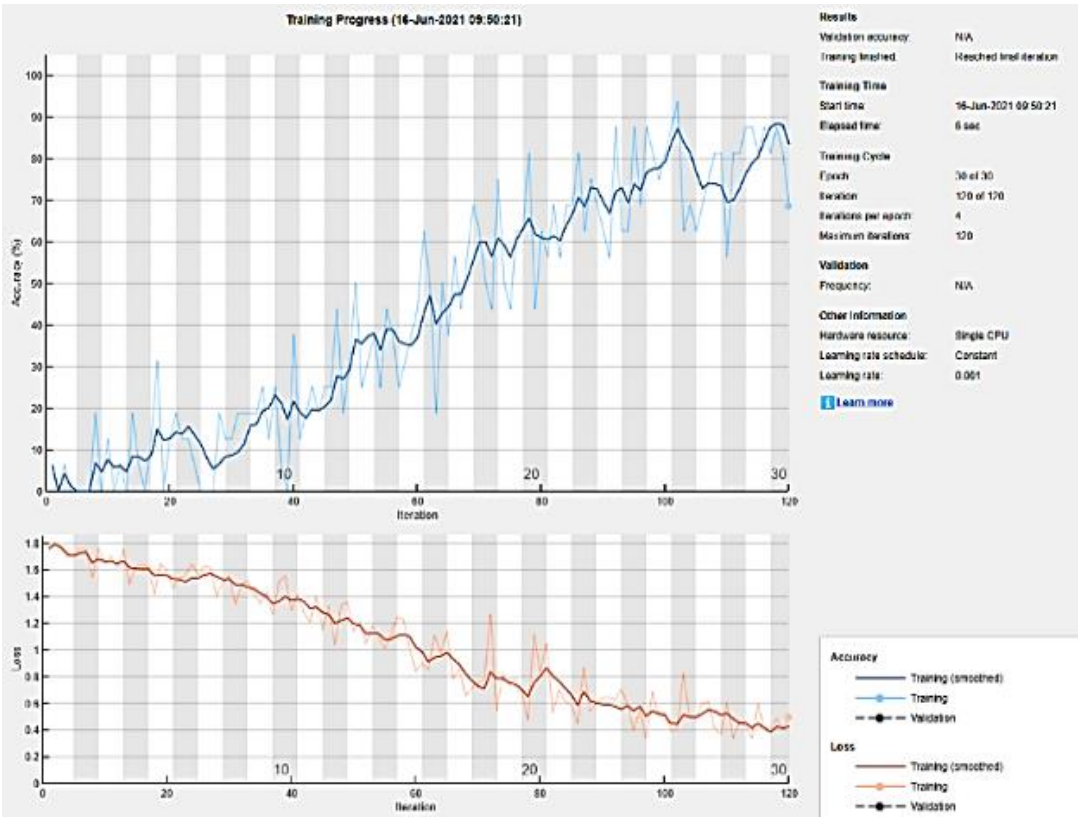
Table 1 Pre-processing technique in input data



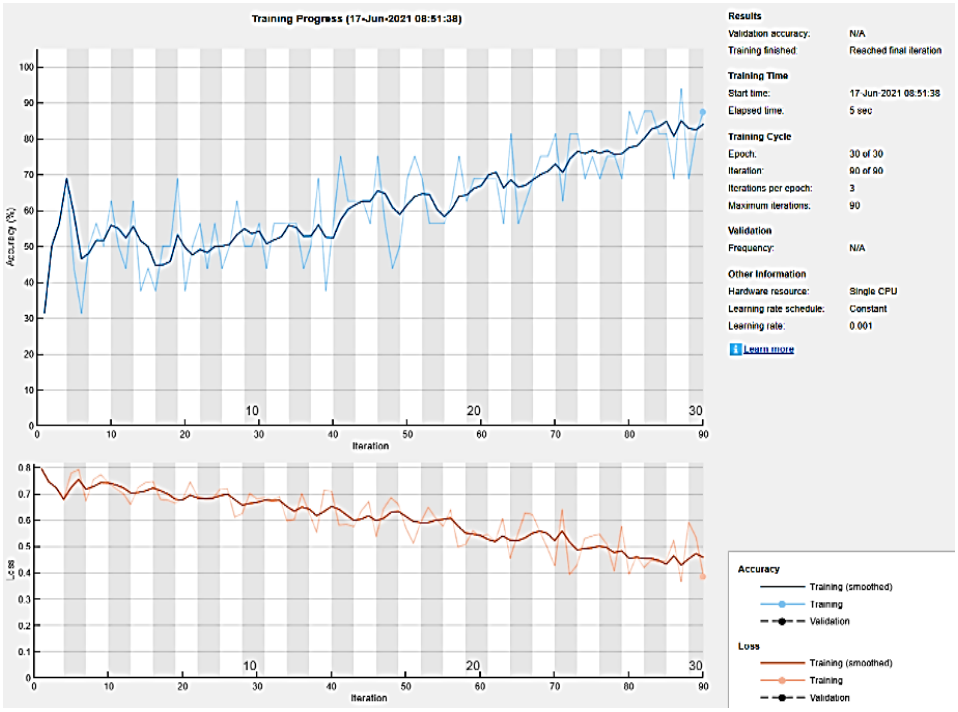
Finally, the dataset is trained for class prediction. Figure 3 presents the training process of the dataset. In proposed approach, the accuracy is validated for three various forms like accuracy in five features extraction, accuracy in eight features extraction and accuracy in fifteen features extraction. As compared to these three types accuracy, the proposed fifteen features extractions' accuracy is high with low error.


Training Progress (16-Jun-2021 12:01:46)
Results
Validation accuracy: N/A
Training finished: Reached final iteration
Training Time
Start time: 16-Jun-2021 12:01:46
Elapsed time: 7 sec
Training Cycle
Epoch: 30 of 30
Iteration: 120 of 120
Iterations per epoch: 4
Maximum iterations: 120
Validation
Frequency: N/A
Other Information
Hardware resource: Single CPU
Learning rate schedule: Constant
Learning rate: 0.001
Learn more
Accuracy (%)
Loss
Iteration
Accuracy
Training (smoothed)
Training
Validation
Loss
Training (smoothed)
Training
Validation

(a)


Training Progress (16-Jun-2021 09:50:21)
Results
Validation accuracy: N/A
Training finished: Reached final iteration
Training Time
Start time: 16-Jun-2021 09:50:21
Elapsed time: 6 sec
Training Cycle
Epoch: 30 of 30
Iteration: 120 of 120
Iterations per epoch: 4
Maximum iterations: 120
Validation
Frequency: N/A
Other Information
Hardware resource: Single CPU
Learning rate schedule: Constant
Learning rate: 0.001
Learn more
Accuracy (%)
Loss
Iteration
Accuracy
Training (smoothed)
Training
Validation
Loss
Training (smoothed)
Training
Validation

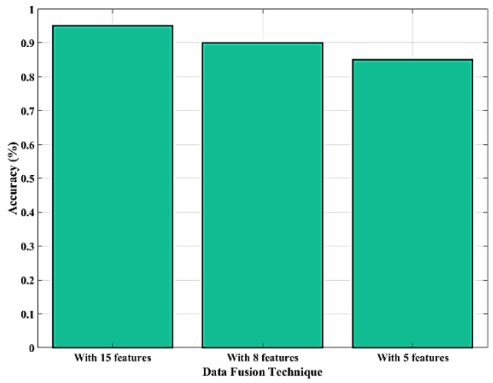
(b)



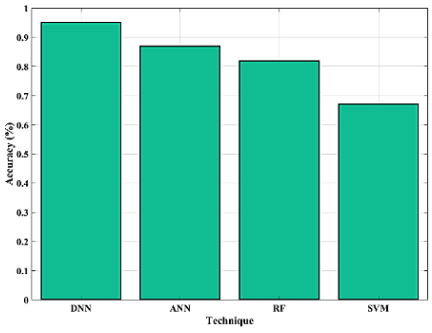
(c)

Figure 3 Training progress of (a) 15 feature (b) 8 feature (c) 5 feature

Accuracy is analyzed for three types, namely, for fifteen feature extractions, for five feature extraction and for eight feature extractions. There is a small variation in accuracy for these fusion techniques. In five features fusion techniques, the features used are GLCM, FFT, Local Binary Pattern, auto correlation and Gabor feature. Accuracy of the method is little low; it gets 85% accuracy. The other fusion approach is eight feature extractions, it contains eight techniques to extract the features from the video clips. The features are GLCM, FFT, Local Binary Pattern, auto correlation, Gabor feature, histogram-oriented gradient, shape index histogram and local mapped pattern. This technique gains high accuracy as contrast with five methods of fusion techniques, it gains 90% accuracy. The final type of fusion is fifteen feature extractions, i.e., proposed approach. The accuracy of the proposed approach is 95%.

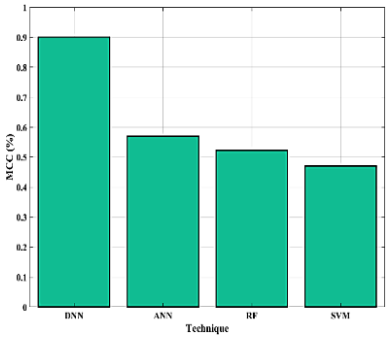


The degree to which the measured value is near to a known or standard value is referred to as accuracy. The proposed method provides an accuracy as high as 95% which is in contrast with the existing methods like ANN, RF and SVM with 89%, 81% and 69% of accuracy respectively which is shown in Figure 5 (a). Because it accounts for genuine and false positives and negatives, the coefficient is a balanced metric that can be employed even if the classes are of extremely different sizes. Figure 5 (b) shows the Matthews correlation coefficient comparison. The proposed method achieves 90% MCC, the ANN achieves 59% MCC, the RF achieves 53% MCC and the SVM achieves 49% MCC.

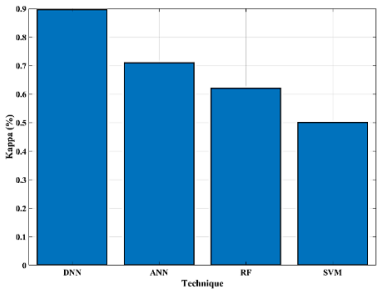


(a)

Figure 4 Comparative analysis of fusion technique



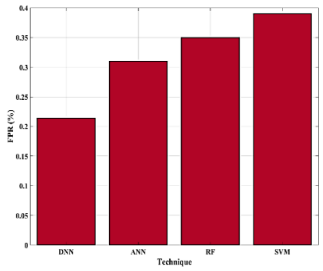
(b)



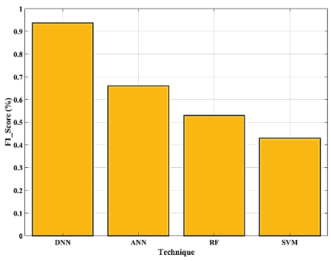
(c)

Figure 5 Comparison of proposed and existing methods (a) Accuracy (b) Matthews Correlation Coefficient Plot (c) Kappa

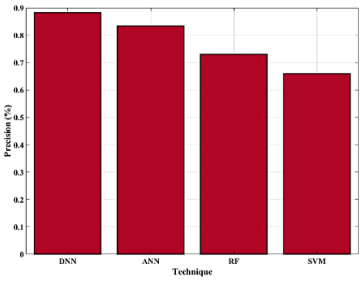
The statistical measurement of Kappa is described as comparing the observed values of a data set to the expected value. In the proposed work, Kappa is 90%. 50% in SVM, 74.1% in ANN, and 65% in the RF technique. Figure 5(c) depicts a comparison overview of the proposed and existing approaches.



(a)



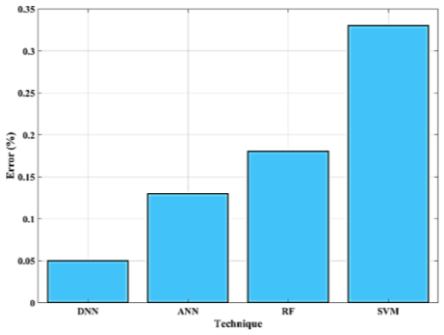
(b)



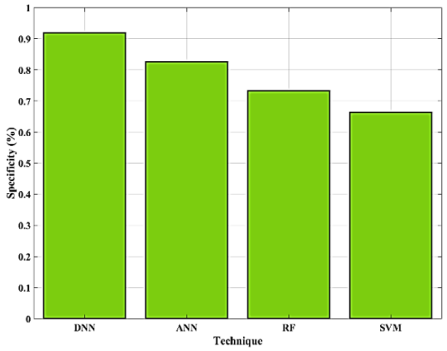
(c)

Figure 6. Comparison of proposed and existing methods (a) False positive rate (b) F1_score (c) Precision

F1 score is a machine learning metric that is used to calculate the system's binary categories and quantify the data set's accuracy. The recall and precision models have a well-defined harmonic mean. The F1 score value is 93% in the proposed study, which is greater than the prior methods. Precision is a technique that is used to label the positive dataset. The split of true positives into the addition of genuine positives and false positives yields precision. The proposed method has a precision value of 89%. When compared to existing approaches, the precision value of the projected method is extremely high. Figure 6 depicts a comparison analysis of the proposed and existing methods.



(a)



(b)

Figure 7 Comparison of proposed and existing methods (a) Error (b) Specificity

5 Conclusion

The aim of the paper is to recognize the activity of the human beings to alert the surroundings which is used to secure the humans form illegal issues. Video clips are gathered from the surveillance camera which is segmented into video frames then the segmented frame is pre-processed via gabor filter. After that, the frame is extracted by data fusion approach, which is again classified for M-DNN to identify the activity. The proposed approach offers the advantage of preventing crime before it occurs. CCTV footage is being tracked and analyzed in real time. The analysis outcome is a directive to the appropriate authority to take action if the result shows that an undesirable incident is likely to occur. This approach can be utilized in any circumstance where suspicious activity monitoring is required. Accuracy of the proposed approach is 95%, which proves that the advanced technique outperforms the existing methods in any circumstance.

Acknowledgment

The authors would like to thank S. G. Balekundri Institute of Technology, Belagavi for providing the necessary infrastructure to carry out this research work and also extend their gratitude to the reviewers at WSEAS transactions on systems and control.

References:

[1] Long, D., Liu, L., Xu, M., Feng, J., Chen, J. and He, L., 2021. Ambient population and surveillance cameras: The guardianship role in street robbers' crime location choice. Cities, 115, p.103223.

[2] Nasaruddin, N., Muchtar, K., Afdhal, A. and Dwiyantoro, A.P.J., 2020. Deep anomaly detection through visual attention in surveillance videos. Journal of Big Data, 7(1), pp.1-17.

[3] Pannirselvam, P.M., Geetha, M.K. and Kumaravelan, G., 2021. A Comprehensive Study on Automated Anomaly Detection Techniques in Video Surveillance. Annals of the Romanian Society for Cell Biology, pp.4027-4037.

[4] Ali, J.J., Shati, N.M. and Gaata, M.T., 2020. Abnormal activity detection in surveillance video scenes. Telkomnika, 18(5), pp.2447-2453.

[5] Meng, T., Jing, X., Yan, Z. and Pedrycz, W., 2020. A survey on machine learning for data fusion. Information Fusion, 57, pp.115-129.

[6] Wu, C., Guo, S., Wu, Y., Ai, J. and Xiong, N.N., 2020. Networked Fault Detection of Field Equipment from Monitoring System Based on Fusing of Motion Sensing and Appearance Information. Multimedia Tools and Applications, 79(23), pp.16319-16348.

[7] Gibert, D., Mateu, C. and Planes, J., 2020. The rise of machine learning for detection and classification of malware: Research developments, trends and challenges. Journal of Network and Computer Applications, 153, p.102526.

[8] Meng, T., Jing, X., Yan, Z. and Pedrycz, W., 2020. A survey on machine learning for data fusion. Information Fusion, 57, pp.115-129.

[9] Mahmoodzadeh, A., 2021. Human Activity Recognition based on Deep Belief Network Classifier and Combination of Local and Global Features. JOURNAL OF INFORMATION SYSTEMS AND TELECOMMUNICATION (JIST), [online], 9(1), p.33.

[10] Khalid, N., Gochoo, M., Jalal, A. and Kim, K., 2021. Modeling Two-Person Segmentation and Locomotion for Stereoscopic Action Identification: A Sustainable Video Surveillance System. Sustainability, 13(2), p.970.

[11] Shifa, A., Asghar, M.N., Fleury, M., Kanwal, N., Ansari, M.S., Lee, B., Herbst, M. and Qiao, Y., 2020. MuLViS: multi-level encryption based security system for surveillance videos. IEEE Access, 8, pp.177131-177155.

[12] Nawaratne, R., Alahakoon, D., De Silva, D., Kumara, H. and Yu, X., 2019. Hierarchical two-stream growing self-organizing maps with transience for human activity recognition. IEEE Transactions on Industrial Informatics, 16(12), pp.7756-7764.

[13] Popescu, A.C., Mocanu, I. and Cramariuc, B., 2020. Fusion Mechanisms for Human Activity Recognition Using Automated Machine Learning. IEEE Access, 8, pp.143996-144014.

[14] Heng, W., Jiang, T. and Gao, W., 2018. How to assess the quality of compressed surveillance videos using face recognition. IEEE Transactions on Circuits and Systems for Video Technology, 29(8), pp.2229-2243.

[15] Liu, W., Liao, S. and Hu, W., 2019. Perceiving motion from dynamic memory for vehicle detection in surveillance videos. IEEE Transactions on Circuits and Systems for Video Technology, 29(12), pp.3558-3567.

[16] Chen, B.H., Shi, L.F. and Ke, X., 2018. A robust moving object detection in multi-scenario big data for video surveillance. IEEE Transactions on Circuits and Systems for Video Technology, 29(4), pp.982-995.

[17] Zhang, H., Li, P., Du, Z. and Dou, W., 2020. Risk Entropy Modeling of Surveillance Camera for Public Security Application. IEEE Access, 8, pp.45343-45355.

[18] Singh, R., Kushwaha, A.K.S. and Srivastava, R., 2019. Multi-view recognition system for human activity based on multiple features for video surveillance system. Multimedia Tools and Applications, 78(12), pp.17165-17196.

[19] Tadic, V., Kiraly, Z., Odry, P., Trpovski, Z. and Loncar-Turukalo, T., 2020. Comparison of Gabor filter bank and fuzzified Gabor filter for license plate detection. Acta Polytechnica Hungarica, 17(1), pp.1-21.

[20] Shafiabadi, M., Kamkar-Rouhani, A., Riabi, S.R.G., Kahoo, A.R. and Tokhmechi, B., 2021. Identification of reservoir fractures on FMI image logs using Canny and Sobel edge detection algorithms. Oil & Gas Science and Technology–Revue d'IFP Energies nouvelles, 76, p.10.

[21] Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D. and Saeed, J., 2020. A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. Journal of Applied Science and Technology Trends, 1(2), pp.56-70.

[22] https://www.crcv.ucf.edu/projects/Abnormal_Crowd/Normal_Abnormal_Crowd.zip

https://www.crcv.ucf.edu/projects/Abnormal_Crowd/Crowd_Dataset_extra.zip

Contribution of individual authors to the creation of a scientific article (ghostwriting policy)

Mr. Shankargoud Patil proposed the idea, carried out the implementation and writing the manuscript.
Dr. Kappargaon S. Prabhushetty mentored, edited the manuscript and provided valuable suggestions.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)