Manuscript received December 8, 2024; Revised February 10, 2025; Accepted March 1, 2025; date of publication March 9, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.673

How to cite: Manojkumar K, Suji Helen L, "Categorizing Crowd Emotions based on Cross Division Expressions and Anomalies", Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 7, no. 2, pp. 341-351, April 2025.

Categorizing Crowd Emotions based on Cross Division Expressions and Anomalies

Manojkumar K, Suji Helen L
Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

Corresponding author: Manojkumar K[1] (e-mail: manojkumar.k@hotmail.com).

ABSTRACT The crowd emotion sensing is a critical element in surveillance and management of the crowd in different environments. With exploding populations, and developing nations, the crowd in urban cities mandate state of art surveillance methodologies involving continuous monitoring and reporting of criminal activities. The research article presents a novel technique to compute the spatial and temporal features obtained from the crowd environments and combine the novelty of neural networks for detecting the emotions of crowds with better accuracy and swiftness. The features are obtained from the continuous feed of surveillance videos typically categorized into the common features of human beings namely anger, sadness, disgust, surprise, fear, happiness and obviously neutrality. Such features are extracted after careful background separation which are typically difficult in crowded environments, using techniques namely SIFT, and FAST termed to be the visual descriptors. Once the features are extracted, spatial and temporal features are classified into individual and combined features as defined in the cross-division environment in order to portray the crowd dynamics and characteristics. Cross division environment computes the necessary features for identifying the anomalies in the crowded situations in a neural network, after a series of operations such as dimensionality reduction, and principal component analysis. From the semantic information, crowd behaviours are detected based on interactive features in a dynamic environment and the proposed technique has demonstrated effective results in terms of 98.9% accuracy in detecting especially violence in crowd datasets collected from UMN.

INDEX TERMS crowd emotion sensing, anomaly detection, spatial, temporal, behaviour, cross division

I. INTRODUCTION

In developing nations, the rapid urbanization has become the regions more crowded and prone to unsafe environments for vulnerable individuals. Despite have numerous measures for video surveillance and monitoring of the crowded environments, the crime rate has been alarmingly high in cities. Major events primarily the religious, sports, and festive seasons have been prone to increasing criminal occurrences in urban cities, according to reports in cities like Delhi, Chennai and Pune [1]. The chances of detecting the anomalies in a crowded environment become more challenging due to the dynamic features of a crowd and hence need for automated techniques to detect the anomalies becomes a primary concern for governments and other organizations. Since the crowd features are completely dynamic and unpredictable, determining the future actions of every individual becomes more challenging. The recent developments in machine learning algorithms involving deep learning, computer vision and automated techniques have opened up a new research domain named as crowd behaviour analysis [2]. Detecting the anomalies has been a booming area of research for monitoring the strange and abnormal activities of certain individuals in a crowded environment. As soon as the domain has recognized enough anomalies, public safety can be ensured by preventing the accidents, controlling the flow of individuals in the crowded environments and thereby protecting the innocent individuals in chaotic situations.

The detection of challenging situations has to be in real-time and thereby alerting the officials in order to prevent further damages to lives and properties. In most of the models, the feature engineering processes play a critical and inevitable role in determining the accuracy of predicting the chaotic situations in a crowded environment. Feature engineering is a process that identifies significant elements that are found to be potential features for representing the crowd behaviour [3]. These features clearly discriminate the chaotic movement and behaviour from a normal vs abnormal crowd. Yet, the accuracy of the determination is completely based on the quality of images, surveillance videos, background features along with the specificity of the abnormal event that has to be detected. This research work aims to consider the shortcomings of the existing methods, and contribute to the

betterment of quicker detection of abnormal events in the crowds.

The amount of information surrounding every individual is extremely huge and it takes only a glance for an average individual to process the information and put it to use. The ability to observe, contemplate, and perceive the information obtained from the surroundings has been extended to powerful ensemble algorithms. From the recovered information, statistical data such as mean, median, variance, distribution, relationship between the data and much more are estimated for understanding the situation of the current scenarios. The previous studies on ensemble algorithms have been extensively covered to illustrate the power of visual key descriptors for image and video analytics. The commonly observed attributes of primary importance are the size, orientation, hue, with respect to facial emotions, motions of the individuals and as a crowd, background information along with economic values [4]. The emotions of the individuals and the crowd depends on the features observed from the backgrounds too. From the existing research works, the ensemble algorithms considered the frequently changing features from a rich feature set obtained from the crowd surveillance videos. The threshold value for each emotion is identified around the mean value for each continuum of feature dimensions. The features may be considered as the immediate change of one emotion to another within a stipulated time, ranging from one to another, small to huge, low intensity to high intensity, thereby guiding the ensemble algorithms to determine the average emotion from emotion dimensions. Unique member identification from a set of emotions can thus be generated by approximating the information collected from the set of features, by recognizing the standalone set of emotions. Despite the measure of change of emotions being estimated in quantity, the stimuli play an important factor in the frequency of changes from happy to sad or to anger with respect to quantity again. All such facial expressions, are further enhanced to improve the quality of detections. Since the proposed work contemplates the cross dimension of various feature sets, every emotion is defined with a distinct emotion observed from the approximation from the feature sets along with the captured stimuli. However, the relationships between the individuals of the crowd were not considered in the ensemble algorithms based on the cross dimensions and categories. Given this scenario, the generalized feature extraction algorithms could not derive the minor overlapping of the feature sets, and hence the need for ensemble algorithms arise. The cross dimensions of the features are hence combined with a feature mapping across various dimensions forming a cross category for better evaluations of emotions.

From the previous research work, the models were able to determine the actions and emotions based on low level stimuli from the environments. The observations were made to classify the actions from circled emotions captured from simultaneous frames that are processed in a spatial-temporal sets of features [6]. The processes ensured that the individuals from the crowded environments are categorized into different classes based on the unique characteristics and thus forming appropriate subsets of feature dimensions. Statistics obtained from each subset depict the summary of various dimensions uniquely assigned to every emotion. Depending on the characteristics of individuals namely the tallness, shortness, fairness, presence of objects and other combinations, multiple cross category feature sets were computed for including monitoring numerous situations accordingly. From the updated cross category feature dimensions, it is observed that the quality of observations improved with considerable accuracy. The feature distribution was formed based on the combinations of different individuals from different dimensions including the category items. Numerically, the peak distributions were observed twice indicating the two-peaks of approximations, in place of a single distribution. The other works extended the previous research works to measure the number of hues collected from the continuous perceptions levels thereby establishing the relationships between various individuals and members of the feature dimension sets. A notable experiment was conducted to identify the hue value and categorize the dimension based on the unique feature set from every cross-category feature sets. Every individual hue property was compared to identify whether it is a new hue property or it exists in other dimension sets. From the investigative results of the previous works, the hue properties were the results of average extraction of individual and unique hue properties from actually derived hues of multiple hue properties. All these factors contributed to the need of ensemble algorithms for deriving multiple feature sets and cross category dimension sets [7]. The extent to which the category relationship of set members with high-level features, influences the combination of algorithm remains unclear wherever the facial features are used. While low-level features may differ fundamentally from processing of the high level features through the algorithms, it is challenging to directly apply findings from low-level feature studies to high-level feature scenarios. In everyday life, groups of interacting faces are often diverse in their categorical expressions, both spatially and temporally. For instance, individuals within a group may exhibit varying attitudes toward a specific event, or an individual's expression may shift significantly due to unforeseen circumstances. This leads to scenarios involving crowds with mixed-category emotions presented simultaneously or sequentially. Understanding whether perceivers can form averaged representations from such heterogeneous emotional expressions is essential, as the ability to derive statistical information from a group of emotional faces plays a critical role in daily life and overall well-being.

This study aimed to investigate whether perceivers could extract average expressions from sets of cross-category facial expressions through two experiments, where facial expressions were presented either spatially or temporally [8][9]. The focus was on happy and fearful expressions rather than the more commonly studied pair of happy and angry expressions. Happiness and anger are both associated with approach motivation—happiness with well-being and anger with attack. In contrast, fear belongs to a distinct emotional

category and is associated with avoidance motivation, representing the opposite end of the motivational spectrum from happiness. This distinction makes happy and fearful expressions more clearly categorized and widely used in studies on the categorical perception of facial expressions. Faces near the categorical boundary were selected as set members for two main reasons. First, in real-life interactions, individuals often display subtle and ambiguous expressions rather than distinct, prototypical emotions. Second, prior research indicates that increased variance within a set diminishes the ability to derive an average representation, which could obscure potential effects of ensemble coding in cross-category groups. Additionally, it is uncommon to encounter groups of individuals simultaneously expressing entirely different facial emotions or an individual rapidly transitioning between extreme emotional categories. While studies on categorical perception of emotional faces support basic emotion theory, ensemble coding of facial expressions aligns with dimensional emotion theory. This perspective suggests that observers perceive facial expressions as part of a continuous spectrum rather than discrete categories.

II. RELATED WORKS

The objective of the anomaly detection algorithms is to detect the changes in momentum, velocity and obviously the emotions of individuals in a crowded environment. A significant benefit of the anomaly detection algorithm is to detect the changes immediately and report the abnormal events to the concerned officials. Various methods have been proposed and the primary methods contemplate the anomaly detection based on objects and holistic approaches [10]. Object based approaches concentrate on the segmentation of individuals into smaller groups, monitoring the trajectories, predicting the trajectories and focus on object-based attributes to extract the probable behaviours of individuals and as a crowd. However, the factors such as occlusion and the presence of multiple target objects are inhibiting the visibility of anomaly detection algorithms. These factors greatly affect the accuracy of predictions and hence the need for better approaches. Holistic approaches, on the other hand, operate on a wider network of individuals where all are interconnected to provide more meaning to the scenes. The low level and mid-level features are extracted for monitoring the crowd behaviours. Holistic approaches concentrated on optical flow fields, that highlighted the low-level and mid-level features for predicting the outcome with better accuracy. An automated method for deriving the optical flow histograms were proposed for monitoring the overall motion of individuals, thereby connecting them in a wide network of people. In chaotic situations such as the stampedes [5], the proposed histograms were fruitful in identifying the potentially critical situations especially in a crowded environment. Another technique suggested the utilization of low to mid-level features in the motion aspect specifically to visualize the magnitude of the crowd along with the direction of dispersions. The segmentation algorithms proposed in the model were enabled to predict the motion of the crowd during chaotic situations by modelling the regions and probable motion of individuals during normal vs abnormal events. A probabilistic model based on Riemannian detection [11] for detecting crowd anomalies was proposed that handled the optical flow with respect to walking, running and dispersion at different velocities in various directions. Comparatively, the performance of models that have taken the regions of interest from the respective frames has been higher than the models that processed the entire video or frames of the videos [12]. The technique was to consider the particle advection specifically for quicker processing and deriving the prediction outcomes. Such techniques also considered the trajectory information of every individual with respect to spatial-temporal features and changes observed in different time frames. Motion patterns identified from the different video inputs have shown significant similarities in chaotic situations, thereby establishing a remarkable similarity in crowd anomaly detections. A technique named Histogram of Oriented Tracklets (HOT) was suggested based on the motion described by the histograms that illustrated the motion features, magnitude and the direction of individuals. In case of highly populated areas, the patterns were able to track the anomalies better than the conventional methods.

Crowd Anomaly Detection has gained popularity in the research domain in recent years owing to automated surveillance techniques, and numerous techniques have been introduced. The challenges in the current techniques are also accounted for and the advent of machine learning, deep learning and computer vision algorithms has simplified the entire process flow. Various articles contemplated the list of techniques used for monitoring the individual behaviours and specifically segmenting the abnormal events in crowded environments. The techniques contemplated in the literature review have been collectively describing the various crowded environments, anomalies and target objects that can be the reason for the anomalies [13]. Such techniques emphasized the need for machine learning, deep learning and computer vision algorithms for automated monitoring systems that can be a live-saving measure for crowded environments. Video analysis techniques played a significant role in anomaly detection models for adding more intelligence to the algorithms. Primarily, the intelligent systems for monitoring the anomalies and crowds used the spatial and temporal features along with perturbations, yet the models needed to address the uneven, non-repeating, unique and rare cases of abnormal events. A better approach implementing a Convolutional Neural Network with multiple optimization techniques was designed and delivered to improve the accuracy of crowd anomaly detection. The need for automated surveillance techniques and thus the accuracy of detection and prediction was justified in the article. Crowded environments were processed in a neural network as continuous streams of input video, right from the cameras and the algorithm detected the anomalies with various threshold features. Ability to process the crowded environments from surveillance videos was tested in a technique that enforced the optimized version of multiple Convolutional Neural Networks. Despite the huge

number of individuals in the crowded environments, and a diverse range of anomalies, the optimized Convolutional Neural Network ensemble performed considerably better than the previous techniques said in the literature survey. Another significant benefit of this approach was the considerably low computational cost and the efficiency when compared to the traditional techniques, apart from the difficulties faced by the other techniques in handling real-time surveillance videos. The next technique approached the anomalies detection problem with a two-step process, where the objects and individuals were identified using a You Only Look Once (YOLOv5) version 5 model and a model named DeepSORT was implemented for tracking the trajectories of the individuals [14]. Optical flow, yet again, was helpful in determining the features that highlight the abnormal behaviours and events of the individuals in the crowded environment. The spatial features were derived from the bounding boxes, compared with the threshold features and the abnormalities were determined. Support Vector Machines were the classifiers that compared the features of the captured events, and compared against the threshold features. From the experimental results, the SVM exhibited an Area Under the Curve metric of 88.9% and the same model has outperformed the other conventional approaches of detecting abnormal events [15][16]. The same model has proven to perform remarkably well while processing the videos of Hajj yatras, indicating the significant performance during the presence of a dense crowd. Moreover, the model has been able to identify seven distinct abnormal events captured during the yatras, even in the huge crowd. Since the further categorization has been narrowed down to seven more distinctive categories, the accuracy of classifying the abnormal events has improved significantly. Combination of multiple individual techniques have showcased the increase in accuracy and area under the curve of classification of abnormal events. Ensemble of state of art machine learning techniques [17] and multiple modalities of classification algorithms, bounding box techniques and feature extraction techniques have been adapting to fit the current landscape of crowd anomaly detection in the recent years [18]. The following TABLE 1 has detailed all the models and techniques discussed in the literature survey that have produced the models for crowd anomaly detections.

TABLE 1
Summary of Models and Techniques used for Crowd Anomaly Detection and their performance

Considered Datasets	Models / Techniques Used	Performance Metric	Outcome
ShangaiTech Violent Flow	Generative Adversarial Networks	Area Under Curve	73.8%
	Recurrent Neural Networks, 2D Convolutional Neural Networks	Accuracy	93.5%
	Optical Flow	Accuracy	73.56%
	Optical Flow GAN	Accuracy	79.6%
		Area Under Curve	98.1%
UCF Crime	CNN Residual Logn Short Term Memory	Area Under Curve	70.4%
	CNN, Random Forest	Area Under Curve	88.98%
Hajj Yatra	Optical Flow	Area Under Curve	88.96%
UMN	SVM	Area Under Curve	88.29%

III. MATERIALS AND METHODS

The proposed model in the research article contemplates the functioning of crowd anomaly detection methodology with two unique techniques namely, the illustration of the crowded environments, their attributes, definition of various behaviours, followed by the dimensionality reduction and classification of behaviours according to the attributes [19]. The surveillance video is processed in a series of frames, from which a set of feature sets are derived to form a list of probably trajectories and Delaunay triangles. From the frames, the visual attributes are further categorized into spatial and temporal features primarily being the velocity, density, motion descriptors and other feature attributes indicating the state of the crowd members. These features are critically acclaimed to the important features that define the quality background upon which the classifiers act upon and determine the abnormal events observed in the given environments [20]. The second part of the methodology involves the processing of the feature sets into a set of histograms after aggregation for computational purposes. This section applies the process of dimensionality reduction through the Principal Component Analysis and autoencoders enabling the classification process. Classification is a typical element of neural networks which categorizes and lists the captured events into two primary classes known to be the normal and abnormal events. The following FIGURE 1 illustrates the functionality of the proposed model and the following sections explain the processes in detail. Similar to any video analytics application, the input is obtained from the real time sources of surveillance and the same is processed by the proposed system. The advancements in surveillance technology have facilitated the storage and retrieval of information in real time.

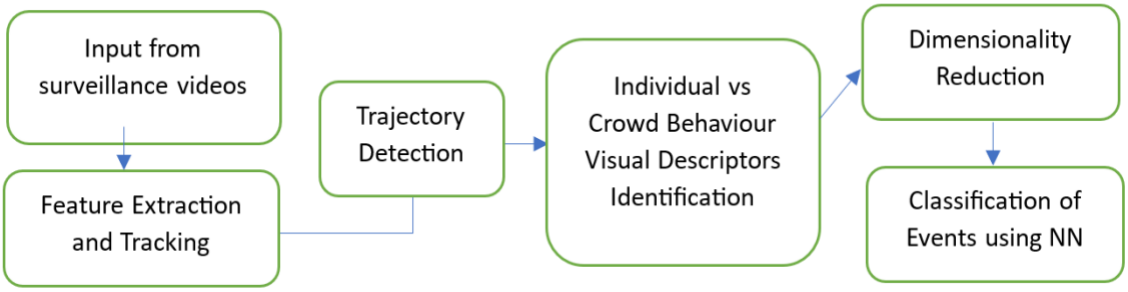


FIGURE 1. **Architecture of the Proposed System**

A. CROWD BEHAVIOUR ANALYSIS

The processes commence with the definition of visual descriptors that explain the characteristics of individuals and the entire crowd. Typically for an anomaly detection system, the input videos are processed directly from the surveillance videos captured in real-time. Given the installation of surveillance cameras in major cities, homes and crowded public places, the availability of surveillance videos from places has increased in recent years [21]. The frames are thoroughly analysed for regions of interest that may hold the potential of analysis followed by the definition of the feature points. Pixels of specific regions of interest with clear and concise events of abnormality are defined from such regions of interest and thus the other regions are removed from the processing of neural networks. The commonly applied techniques for identifying such events from the frames are FAST, SIFT and AKAZE [22], that are known for the quickness and accuracy of detecting such feature sets from subsequent images. Every frame is analysed through these renowned algorithms to narrow down the features of interest and following them in a series of images thereby ensuring that the abnormal event is prolonged for a specific duration. Such feature sets usually depict the characteristics of objects, individuals and the surroundings detected in a surveillance video. The identical feature sets in subsequent frames describe the valuable information that has to be processed for features tracking and predicting the trajectories of every individual in the crowd. In order to add more sense to the predictions, the objects are considered as well, to detect the abnormal events well in advance. The spatial features are better derived from the Delaunay triangulation strategy, where the spatial elements are highlighted and differentiated between the subsequent frames. Such features are further defined as the visual descriptors which are classified into individual and combined features accordingly [23].

1) FEATURE EXTRACTION AND TRACKING

The crowd emotion sensing is a critical aspect of the proposed model where the detection of various facial expressions of every individual in the crowd becomes mandatory. As soon as the expressions are detected, the next process is to monitor the collective gestures and movements which in turn reflect the diverse range of emotional states. The emotions have to be accurately predicted as much as the quickness in detection. Considering all the requirements, precision and performance have to be balanced in any complex environments. The techniques used in the proposed model are selected for assuring the balance between precision and performance. The Features from Accelerated Segment Test (FAST) technique is applied for assuring the quickness in corner detection even in real-time video inputs. Unlike the other feature detection techniques, FAST is known for lower overheads, and ensures quick approximation for detecting the crowd dynamics. The next method Scale Invariant Feature Transform (SIFT) is known for the resilience to noisy features and illumination issues found commonly in surveillance videos. Moreover, the ability to detect highly distinctive features even in the subtle environments was higher in SIFT. In order to improve the accuracy of detection in a non-linear scale space, Accelerated KAZE feature detection is applied for fine tuning the feature detection stage. Comparatively, the AKAZE feature detection technique is known for its performance and accuracy over the Modified Local Difference Binary approach (M-LDB) technique. On a broader perspective, SIFT and AKAZE techniques were much faster than Oriented FAST and Rotated BRIEF (ORB) technique. Speeded Up Robust Features (SURF) technique is known for its swiftness yet compromising on the robustness in handling the varying illumination issues. Generally used in object detection, Histogram of Oriented Gradients (HOG) may not be effectively applied in detecting emotions in real time surveillance videos. The following sections explains the techniques in detail, as applied in the proposed model.

In the proposed approach, the Features from Accelerated Segment Test (FAST) technique was applied to identify the pixels of interest from the frames of significant importance and the features were compared with the continuous frames to match for similarities [24]. The variations of intensity around the circled region of interest explains the continuous changes in the trajectories and on the other hand, the similar activities in the regions of interest in the neighbouring frames indicate the capture of abnormal events. In order to reduce the computational complexity, the corner criteria technique is applied to identify the predetermined areas of interest on a specific frame. As soon as a specific feature is identified using

FAST algorithm, Scale Invariant Feature Transform (SIFT) algorithm is applied to derive a 16x16 neighbourhood matrix around the detected feature thereby deriving a 128-bin value. The equation represents the Difference of Gaussians (DoG) [25], which approximates the Laplacian of Gaussian (LoG) for edge detection in image processing. This difference isolates features that vary in size between the scales, highlighting image details at a specific range of scales. The resulting difference is then convolved (*) with the input image I(x,y) to detect edges or features using the following Eq. (1)., where x, y are the source data to which the Difference of Gaussians filter is applied to detect features. Convolution integrates the filter with the image to detect the desired features at each pixel. The scales $k\sigma$ and σ indicate the difference between the two functions as shown in Eq. (1).

$$D(x,y,\sigma)=[G(x,y,k\sigma)-G(x,y,\sigma)]*I(x,y) \quad (1)$$

In this approach, the feature extraction capabilities of SIFT and FAST along with AKAZE were individually analysed, highlighting their unique contributions to anomaly detection in sparse feature tracking. SIFT is well-known for its ability to remain robust under variations in scale, rotation, and lighting conditions. This method is particularly effective in identifying distinct key points within crowd images, allowing it to capture intricate patterns that signal deviations from typical crowd behaviour. As a result, it is highly suitable for detecting anomalies of different sizes. FAST, on the other hand, is optimized for quick corner detection, enabling the rapid identification of key features in crowd images [26-28]. While it lacks the scale and rotation invariance of SIFT and FAST, its high speed makes it an excellent choice for real-time anomaly detection applications, where quick responses are essential. The efficient detection of key points by FAST enhances its utility in sparse feature tracking for recognizing unusual crowd behaviours. Sparse feature tracking is a widely utilized technique in computer vision for tracking a subset of key features across video frames. Unlike dense tracking methods, which monitor every single pixel, this approach zeroes in on distinctive features within the frames. This focus on unique features such as corners or structural elements enables it to handle challenges like occlusion (where parts of the crowd are hidden) and dynamic changes in crowd behaviours more effectively. Sparse tracking excels by identifying and following these prominent key points [29], ensuring robust performance even when parts of the crowd are obscured or when movement patterns shift unpredictably. One notable advantage of sparse feature tracking is its ability to re-detect and associate features over time. This ensures the system adapts to evolving crowd dynamics, allowing for accurate long-term tracking. By prioritizing features that are simple to identify and track, our approach maintains its effectiveness across various scenarios [30].

The Accelerated-KAZE (AKAZE) algorithm extends the original KAZE algorithm by using a computationally efficient framework known as Fast Explicit Diffusion (FED) to create its non-linear scale spaces. Built upon non-linear diffusion filtering, AKAZE utilizes the determinant of the Hessian matrix for feature detection. To enhance rotation invariance, it employs Scharr filters. The maximum responses from these detectors pinpoint specific feature point locations, forming the basis for AKAZE's strong and distinctive feature detection capabilities. The AKAZE descriptor [31] leverages the Modified Local Difference Binary (MLDB) algorithm, renowned for its power and efficiency. Due to the non-linear nature of AKAZE's scale spaces, the algorithm exhibits invariance to scale, rotation, and limited affine transformations. Furthermore, the distinctiveness of AKAZE's features is maintained and even enhanced as they are scaled up or down.

Our methodology capitalizes on these local features and incorporates the Lucas–Kanade optical flow algorithm. This algorithm is particularly adept at handling the challenges posed by unrestricted optical flow, making it ideal for applications like crowd anomaly detection. The Lucas–Kanade method stands out due to its precision, robustness, and adaptability to complex environments [32]. By emphasizing sparse feature tracking, we substantially reduce computational overhead while maintaining high accuracy, perfect for real-time applications. The Lucas–Kanade algorithm enhances its effectiveness in detecting anomalous behaviours by capturing subtle motion variations within crowded environments. Ensuring temporal coherence in feature tracking stabilizes the system and minimizes false positives, thereby increasing the reliability of anomaly detection mechanisms. This approach operates on the assumption that neighbouring pixels within a small, localized region share consistent optical flow values. Instead of analysing every pixel in the frame, it calculates optical flow based on these groups. This method provides several benefits, such as faster computations and the efficient generation of training data. Mathematically as expressed in Eq. (2), the optical flow constraint for a set of pixels moving at the same velocity can be expressed by a specific equation, ensuring a cohesive and efficient analysis of movement within the crowd [33]. According to the Eq. (2)., $I_x(x,y)$ and $I_y(x,y)$, the spatial intensity gradients are defined for any point (x,y). The changes in the pixel intensities are represented with respect to the directions x and y, where v depicts the velocity of a person or an object, t depicts the time for all the changes. By leveraging these principles, our methodology strikes a balance between computational efficiency and accuracy, making it a powerful tool for real-time crowd monitoring and anomaly detection.

$$\begin{aligned} I_x(x_1, y_1) \cdot v_x + I_y(x_1, y_1) \cdot v_y &= -I_t(x_1, y_1) \\ I_x(x_2, y_2) \cdot v_x + I_y(x_2, y_2) \cdot v_y &= -It(x_2, y_2) \\ &\cdots \\ I_x(x_n, y_n) \cdot v_x + I_y(x_n, y_n) \cdot v_y &= -I_t(x_n, y_n) \end{aligned} \quad (2)$$

This efficient and adaptive approach to sparse optical flow makes it a promising solution for crowd anomaly detection in diverse surveillance and monitoring scenarios [34].

2) TRAJECTORY DETECTION

The Lucas Kanade method may not function in case of object detection especially in motion and the trajectories may not be detected effectively. This raises the requirement to include the gradient transformation for considering the neighbouring pixels of the respective frames to detect the object motions. The proposed approach introduces a novel approach for considering the pixels from regions of interest in form of a pyramidal format. This technique contemplates the techniques of down-sampling, passing through a low-pass filtering technique and applying a factor of 2. The purpose of optical flow inclusion is justified when the low-quality images are processed first [35], followed by the high-quality frames from the same videos, in order to increase the optical flow accuracy. Once the spatial features are computed, the significant features are forwarded to the next stage of processing as tracklets. The series of frames, where the objects are defined to be the primary features of consideration, are transformed into a graph with mapped tracklets. Delaunay Triangulation graph technique [36] processes the features in omnidirectional nodes in the neighbouring frames in order to trace the local, spatial and temporal features or tracklets. The number of nodes depends on the tracklets and the node connections are represented as ϵ^n and the relationships between the tracklets are represented by $g^n(\vartheta^n, \epsilon^n, F^n)$. The number of triplet features are represented as F^n. Temporal features are effectively described with respect to the topographical features over the different states of time without affecting the shape of the triangle in a graph eliminating the noise and occlusion factors. A graph containing the tracklets derived from the previous stages, the local tracklets are identified as cliques and neighbouring nodes are identified as seed points, which forms the tracking points for objects and individuals in a crowded environment. The cliques are represented according to the following Eq. (3)., where $V_i^k$is an individual in a cluster C(V_i^k). In V_i^k, k indicates the frame and i denotes the position of the object/individual. The ϵ^n denotes the edges and relationship between individuals. This equation helps identify clusters of people moving together, which may indicate collective behavior.

$$C\left(V_i^k\right) = \left\{V_i^k\right\} \cup \left\{V_j^k, \forall \left(V_i^k, V_j^k\right) \in \epsilon^n\right\} \quad (3)$$

The connections between the cliques or the tracklets are connected through the short-term or long-term connections representing the probable trajectories. In terms of spatial features, the cliques are connected on varying terms increasing the dynamic ability of the crowd. With a more diversified crowd, the number of cliques with temporal and spatial features are comparatively higher than insignificant cliques.

3) INDIVIDUAL VS CROWD BEHAVIOUR VISUAL DESCRIPTORS IDENTIFICATION

Once the visual descriptors are defined and mapped in a graph, the characteristics of objects and individuals are further analysed for classification. Visual descriptors indicate the semantic information about the crowd participants, typically the spatial and temporal aspects. Information captured from the input videos are highly dynamic and critical for the further analysis [37]. The proposed system carefully analyses the individual and collective features, thereby contemplating the entire set of features from the surveillance videos. Such visual descriptors are further classified into individual and entire crowd-based features namely the collective features. Individual behaviours are processed to segment the participants of the crowd, understanding their activities by observing the dynamic properties such as direction of the flow and velocity of individuals. The direction of motions, describes the individuals with respect to normal and extreme conditions in varying tracklets motions [38]. A complete structure of the tracklets, neighbouring nodes, are represented by the F segment as described in the following Eq. (4). According to the equation, S_i^nsignifies the state of an individual i at the current time step n , $S_i^{n-\tau_2}$ denotes the state of an individual at an earlier time and so on.

$$\left\{S_i^n, S_i^{n-\tau_2}, \ldots . S_i^{n-(F-1)\tau_2}\right\} \quad (4)$$

The changes in the directions of the individuals are observed by monitoring the angular variations of every trajectory and tracklets. The following Eq. (5). provides the calculation for predicting the change of direction in a given trajectory.

$$D^{var}\left(V_i^k\right) = \frac{1}{F} \cdot \sum_0^{F-2} d_\theta\left(S_i^{n-f\tau_2}, S_i^{n-(f+1)\tau_2}\right) \quad (5)$$

Where θ represents the angular variations, F is represented by vectors of the graph, the states are considered as $S_i^{n-f\tau_2}$, separated by different time slots and $d_\theta$represents the angular distance. This could be useful in analyzing temporal patterns, detecting changes in dynamic systems, or extracting features from time-series data. On the other critical element for determining the individual characteristics, velocity of motion vectors in a specific direction explains the seriousness of the situation occurred in the crowded environment [39]. The accuracy of the visual descriptors depends on the considered frame in the history of frames. The motion vectors suitably describe the information about the motion of individuals using the following Eq. (6). Euclidean distance is the measure between the tracklets in any direction, preferably between the current node and origin, instead of measuring the sum of all the tracklets between the nodes.

$$D^{velocity}\left(V_i^k\right) = \frac{1}{\tau_1} \cdot \left\| \overrightarrow{V_i^{k-\tau_1} V_i^k} \right\| \quad (6)$$

In a crowded environment, it is extremely important to derive the collective characteristics of all the participants and the collective behaviours in this section explains the need for collective visual descriptors. The characteristics of the crowded collective features are derived from the stability, collectiveness, density and uniformity [40]. The concept of stability in crowd analysis reflects the degree of consistency in the crowd's topological structure over time. It evaluates how individuals within a crowd maintain their proximity to the same neighbours as time goes on. By examining the stability property, we can glean valuable insights into persistent

patterns and relationships within the crowd, thereby enhancing our understanding of its dynamics and behaviours.

The collectiveness property pertains to how pedestrians move as a cohesive group. This property is measured by calculating each individual's directional deviation from the overall movement of the group. Traditionally, coherent motion has been assessed using predefined collective transitions. However, in this approach, cliques are utilized for the local computation of this descriptor, offering a nuanced alternative. Conflict, an important property in crowd analysis, captures interactions among individuals, especially when they are in close proximity [41]. Similar to the computation of the collectiveness descriptor, the conflict property is also determined locally.

The local density descriptor focuses specifically on the spatial distribution aspect of the model. Unlike previous interactive descriptors, it emphasizes the spatial arrangement of individuals within the scene. This descriptor captures a key characteristic of crowd behaviours: the distribution of individuals. An approximate measure of local density can be obtained by evaluating the proximity of nearby features. This is based on the observation that when nearby features converge, it signifies a higher probability of a larger crowd forming in that area. The uniformity descriptor assesses the coherence of the spatial distribution of regional features. It indicates whether a group has a tendency to cluster together in a uniform manner or to fragment into smaller subgroups, reflecting non-uniform behaviours.

4) DIMENSIONALITY REDUCTION AND CLASSIFICATION

The previous section explained the list of descriptors that worked upon the features with spatial and temporal features that were predominant in identifying the abnormal events in the crowded environments. Dimensionality reduction [42] is a renowned technique for processing the potential list of features by reducing the rich feature sets and processing only the required critical information. The proposed approach reduced the dimensionalities using two remarkable techniques namely principal component analysis and autoencoding. These two approaches are common in computer vision applications and have been proven to reduce the computational cost associated with processing images and videos. The first technique employed was principal component analysis (PCA). PCA aims to transform the original features into a new set of uncorrelated variables, known as principal components, while preserving as much variance as possible. By projecting the data onto these components, we effectively reduce the information into a lower-dimensional space.

Neural networks are implemented in the proposed model as they are known for their ability to handle unstructured and complex information such as images and videos. Feature extraction and mapping were performed by three renowned techniques and neural networks enabled the automated feature engineering techniques eliminating the need for complex architectures for detecting individual and group dynamics extracted from the input videos. The neural networks are also known for the ability to handle complex data in huge datasets, making it suitable for large crowds as well. Comparatively, the neural networks are better performing than Support Vector Machines as they are bound to accuracy issues when complex and huge datasets are processed. On the other hand, clustering algorithms such as K-Means were restricted by their applicability over sequential information rather than temporal analysis. The computational overhead of random forests prevents the applicability owing to the computational time and lower effectiveness over time-series analysis.

The detection of anomalies within crowd dynamics is a critical challenge in visual crowd analysis. This paper presents a comprehensive approach that leverages dimensionality reduction techniques and neural network architectures to effectively identify unusual crowd behaviours. The process begins with the application of principal component analysis to reduce the dimensionality of the input data. By capturing the most significant features, PCA helps to prepare the data for the subsequent classification stage. The second approach involves using autoencoders, a type of neural network designed to learn efficient representations of the input data. Autoencoders consist of an encoder network that compresses the input into a latent space representation and a decoder network that reconstructs the original input from this representation. By training the autoencoder to minimize the reconstruction error, it learns to capture the most important features of the data in the latent space. After applying dimensionality reduction using PCA and autoencoders, we proceed to the classification stage. In this step, we utilize the reduced-dimensional feature set to train neural network classifiers. Specifically, we employ neural networks like the multi-layer perceptron. Neural networks are well-suited to handle non-linear problems and extract intricate patterns from the input data.

The adaptability of neural networks is especially useful in detecting anomalies in crowd behaviours. They excel at uncovering subtle relationships within the data and identifying unusual crowd dynamics that might be overlooked by traditional methods. Due to their hierarchical architectures, neural networks can capture both detailed and higher-level representations, providing a comprehensive understanding of crowd behaviours. Throughout our study, we evaluated neural networks with varying numbers of hidden layers to determine the optimal configuration for our analysis. Through systematic testing and performance comparisons, we identified the most effective approach for crowd anomaly detection (Favarelli & Giorgetti, 2020). This rigorous approach ensured that our methodology was robust and capable of addressing the complexities of detecting anomalies within crowd dynamics.

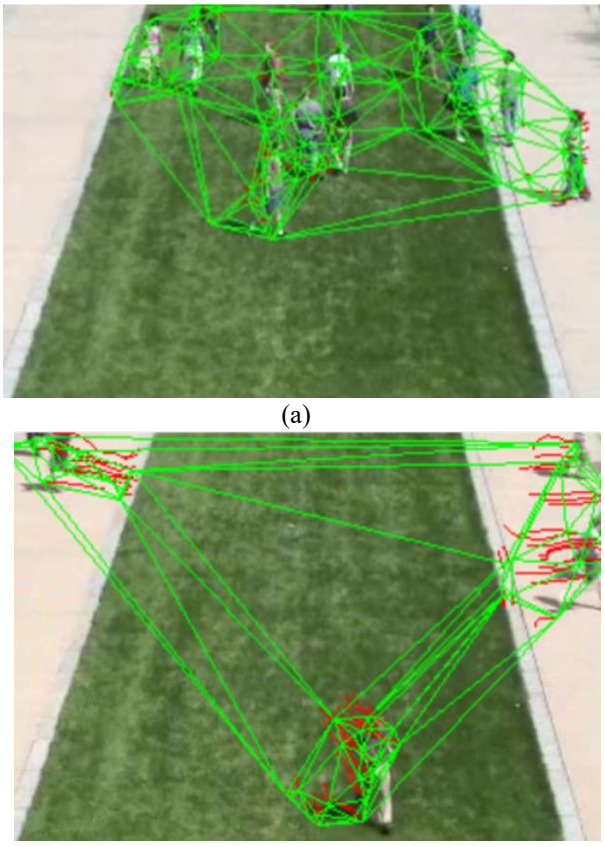
IV RESULTS AND DISCUSSIONS

The proposed approach exemplifies the performance of crowd anomaly detection through a series of investigations with state of art datasets known to have captured crowd activities. The datasets possess various crowd activities and is a standard dataset to be used in this specific research domain. Two primary datasets were considered for the study, University of

Minnesota (UMN) dataset with 11 videos comprising of three different scenes captured from a hallway, lawn and an open plaza. The videos clearly depict the variations from normal crowd behaviour and an escaping behaviour where the people depict panic. The videos are of 320x240 pixels in order to optimize the computational overhead. The other video dataset was sourced from YouTube videos that focused on crime and violent videos. The size of the datasets was around 24 videos of the same resolution with equal number of normal (peaceful situations) and panic situations. University of Minnesota (UMN) along with other datasets that have captured violent and unsettling events have been used for measuring the performance of the proposed approach. Various scenes in the 11 videos of the datasets depict the normal and abnormal (violent) behaviours in subsequent frames, thereby having scenes of people running in a single direction, different directions, moving from the same point of origin with high velocity, and other atypical scenes of abnormal nature. YouTube is also the repository that contains numerous videos of unsettling crowded environments with different tension among the participants, and other violent scenes. Surveillance videos are curated into a list of videos containing a balanced number of normal videos and violent videos. From the UMN datasets, the proposed approach has delivered the visual descriptors for sensing the different behaviours and the YouTube videos were used to measure the accuracy of predicting the abnormal events in challenging situations. It is understood that the UMN datasets are composed with higher number of crowd events with normal/peaceful environments and YouTube videos was curated with a higher number of violent incidents to simulate the proposed model. Given that the models perform inconsistently with such imbalances, the number of peaceful scenarios (majority class) and abnormal or panic situations (minority class) from UMN datasets are addressed using oversampling and random undersampling techniques. Synthetic Minority Oversampling Technique (SMOTE) generated synthetic samples of abnormal events by interpolating the existing minority class samples. And random undersampling was applied to remove the exceeding number of peaceful events from the UMN dataset. In case of the YouTube repository, violent videos were oversampled to match the number of peaceful situations using random oversampling technique.

The proposed model was built using a convolutional neural network after the dimensionality reduction techniques Principal Component Analysis and Autoencoders are implemented on the input datasets. The PCA reduced the features to 128 components. The first dense layer contemplated 256 neurons with ReLU (Rectified Linear Unit) for addressing the non-linearity of the input. Overfitting was avoided using a dropout measure of 0.2. The second dense layer consisted of 128 neurons with batch normalization and the third dense layer with 64 neurons. The output layer implemented a SoftMax classifier for multiclass identification. The learning rate of the proposed NN architecture was set at 0.001, as commonly used by Adam Optimizer and ReduceLROnPlateau scheduler was in place to avoid the validation loss plateaus. The batch sizes were tuned between 16 to 256, and 128 yielded the most promising results. Adam Optimizer is a commonly used optimizing in most of the image and video processing models owing to the need of minimal tuning. In case of epochs, the range was between 50 to 100 epochs, and early stopping was enabled to prevent validation losses.

The following FIGURE 2 illustrates the spatial attributes captured by the Delaunay triangulation technique which explains the distribution of crowd using the following metrics. Each individual in the crowd is marked as a distinct point. Triangles are formed such that no point lies inside the circumcircle of any triangle. The edges of these triangles represent the proximity of individuals to one another, capturing the spatial relationships within the crowd.



(a) (b)

FIGURE 2. The spatial attributes captured by the Delaunay triangulation technique, (a) Marking the individuals as distinct points, (b) Delaunay Triangulation showing distribution of crowd with anomaly detection represented as red coloured zones.

From the UMN dataset, the scene representing the spatial attributes between individuals and cluster of people will be tracked using the visual descriptors identified by the feature extraction techniques of the proposed model. As soon as the attributes are marked and identified to be potentially significant attributes, the Delaunay Triangulation technique will mark the presence of individuals as distinct points, a group of such distinct points marks a crowd with normal

activities. Upon the presence of abnormal activities, the events are marked as red, with varying velocity, the Delaunay triangulation will mark the edges in green-coloured lines representing different clusters of people/participants.

The cohesion among the group participants will be lesser in case of an abnormal event, thereby dispersing the crowd in different directions. The clusters will be less crowded owing to the events and the Delaunay triangulation clearly differentiates the normal vs abnormal crowds as shown in FIGURE 2(a) and FIGURE 2(b) respectively. When the crowd is found to be agitated due to an abnormal event, the distance between the nodes/distinctive points is found to be longer than the nodes with normal events. The length between the nodes will be consistent in a normal crowd and uniformity is a critical element. Length of the green lines will be longer and scattered when the participants are dispersed or scattered in different directions when a chaotic environment is found in the crowd. The proposed model was evaluated with respect to Area Under the Curve (AUC) as the metric is robust against the class imbalances in real-time scenarios. Typically, the situations such as panic and violence will be sensed much lesser than the peaceful events. The AUC has been proven to be a resilient condition in presence of different thresholds and diverse class distributions. On the other hand, accuracy of detection is a simple and straightforward metric that denotes the number of right vs wrong detection of events under a specific class. And accuracy metric is also proven to be resilient against class imbalances and this would be sufficient for initial classification results.

The following TABLE 2 and TABLE 3 has listed the performance measures in terms of accuracy and area under the curve (AUC) metrics based on the quality of feature extraction techniques applied in the proposed model in normal scenes and abnormal events. The configuration of the neural networks is tested for three different scenarios where a normal neural network is applied without any dimensionality reduction techniques, and an autoencoding technique is applied in the next configuration where the dimensions are reduced to 128 bits. The last variant is the performance of the proposed method where the contrast features are reduced based on Principal Component Analysis (PCA) where the variance value is considered to be 95. For testing purposes, a 5-fold cross-validation technique is applied for classifying the behaviours. The dataset was divided into 5 equal subsets where four subsets were used for the training purpose and one for testing. The process was iterated five times with each subset accounted for the testing and validation. The average performance of each fold was considered as the overall measure of the entire model. Cross validation is a standard approach to evaluate the performance of the models with repetitive and rigorous testing strategies, across the entire combinations of the data.

TABLE 2
Performance of different techniques for detecting chaotic scenes in input videos

Techniques	Neural Network		128-Bits Dimensionality Reduction NN		PCA – NN	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
SIFT	0.803	0.805	0.799	0.802	0.848	0.846
FAST	0.848	0.853	0.844	0.847	0.844	0.851
A-KAZE	0.858	0.862	0.865	0.890	0.923	0.898

TABLE 3
Performance of proposed model in detecting normal scenes vs panic scenes in input dataset

Normal Scene	Neural Network		128-Bits Dimensionality Reduction NN		PCA – NN	
	Acc	AUC	Acc	AUC	Acc	AUC
SIFT	0.995	0.90	0.985	0.976	0.989	0.984
FAST	0.987	0.98	0.943	0.967	0.981	0.98
A-KAZE	0.952	0.904	0.953	0.961	0.921	0.912

The model classified the group of people walking in a hallway as a panic situation when they started running towards the emergency exit, as given in the UMN dataset. The probability of predicting this event with high confidence was identified at 95%. The model was able to identify critical changes in crowd dynamics, velocity and density changes in the crowd. Likewise, from a violence event captured from a public place, as sourced from YouTube, the particular event was localized immediately and classified as an anomaly with a confidence level of 92%. The following TABLE 5 documents the confidence intervals for the accuracy and AUC obtained for the proposed model. Using bootstrap sampling, the testing dataset was repeatedly tested with the test dataset. The confidence level was defined at 95% within which the true metric value lies with a 95% probability. The performance of the different models considering the parameters accuracy and AUC is illustrated in FIGURE 4. The proposed model achieved an AUC of 89.67% and 97.50% for scenes with normal activities and abnormal activities, respectively when

the input from UMN dataset was processed. The model has been tested recursively with different scenarios for addressing the various complex situations and other factors limiting the processing ability of the proposed model.

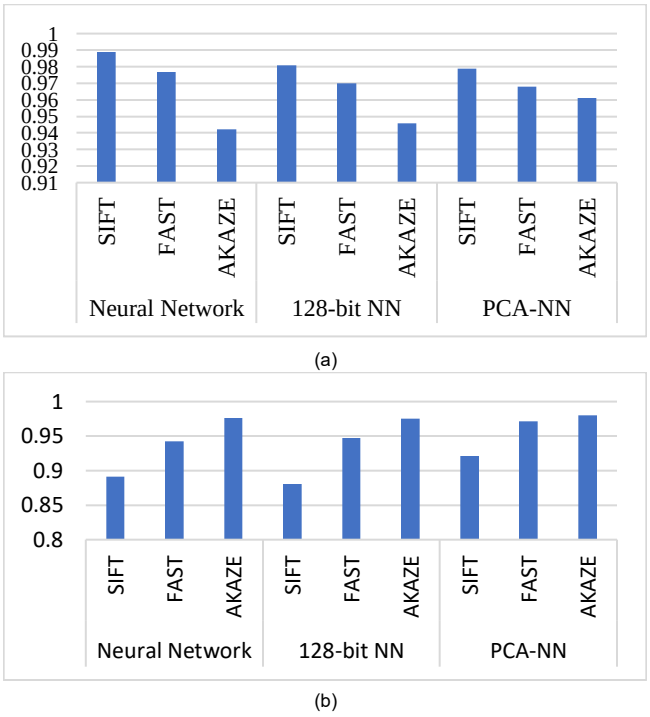


FIGURE 3. The visual descriptors presented in the neural network, (a): Performance of Visual Descriptors over UMN Dataset, (b) Performance of Visual Descriptors over YouTube Video List

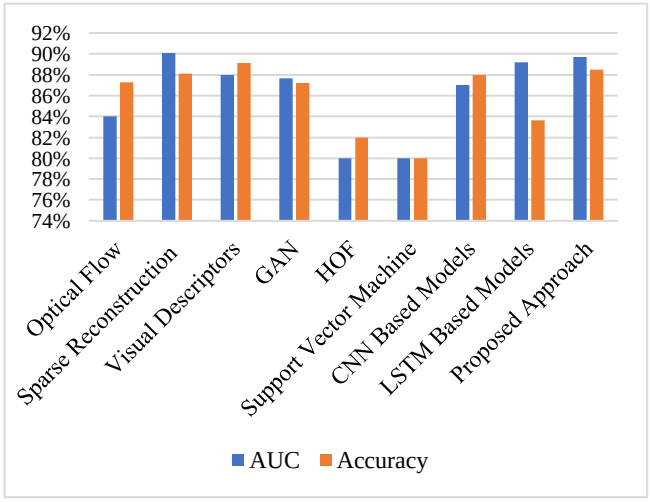


FIGURE 4. Performance of different models for classification

The input videos from the UMN dataset were processed, where the accuracy and AUC performance parameters are tabulated in the TABLE 3. Feature sensitivity was a major issue in the proposed model due to the limitations of the FAST, SIFT and AKAZE as thes techniques were sensitive to noise, occlusions, lighting and at times lead to inconsistent results. Since the videos were processed, the temporal dependencies

TABLE 4
Performance of different Models for classification

Authors	Models	AUC	Accuracy
Valentina Franzoni et al. [43]	Spectrogram-based analysis using deep learning	Audio recordings of crowd sounds	Accuracy: 80% in classifying emotions like cheering and booing
M. Rabiee et al. [44]	Lightweight Convolutional Neural Network (CNN) model for crowd behavior analysis	Video footage of crowds	Accuracy: 85% in emotion recognition tasks
S. K. Ghosh et al. [45]	Lexicon-based approach for emotion classification	Twitter data (textual content)	Accuracy: 75% in detecting emotions such as happiness, sadness, and anger
J. Long et al. [20]	Real-time crowd analysis using CNN for facial expression recognition	Images from surveillance cameras	Accuracy: 88% in real-time emotion prediction
T.H. Noor et al. [14]	Facial expression recognition for crowd monitoring	Video data	Accuracy: 82% in detecting emotions like happiness, surprise, and anger
X. Liu et al. [39]	Analysis of crowd behavior features from video data	Surveillance footage	Accuracy: 84% and focuses on feature extraction methodologies
S. Zhang et al. [23]	Cross-division expressions and neural networks for crowd emotion categorization	Video and image data	Accuracy: 87% in categorizing crowd emotions
Proposed Approach	Cross Category Expressions for classifying emotions	Video Data	Accuracy: 88.67% in categorizing crowd emotions

were also higher. Optimizing the recurrent layers of the NN architecture will address these issues. From the comparative analysis, our approach has shown superior performance relative to other methods as listed below. Specifically, our approach achieved 99.5%, 96.5%, and 99% for normal and abnormal activities of the UMN dataset, respectively. Moreover, our approach shows a more pronounced advantage on the YouTube videos, where the proposed approach has delivered 89.67% AUC and 88.5% accuracy in detecting the abnormal activities.

From the investigative results, consistent performance of our approach across both datasets indicates its high effectiveness in classification tasks, leveraging unique features and techniques that contribute to its superior performance. According to the following FIGURE 3 (a), the visual descriptors presented in the neural network, 128-bit neural network, and PCA-neural network are compared and listed for the dataset computed from UMN dataset. The next FIGURE 3(b) contemplates the performance of visual

descriptors of the same models over the dataset composed from the YouTube repository. The TABLE 4 lists down the comparative results of classification accuracy with other state of art techniques and models. The traditional state of art models such as CNN, SVM, HOF, and LSTM are included in the following comparison. The TABLE 5 lists down the comparison of recent approaches in crowd emotion sensing and how the proposed model has performed comparatively. The performance metric was chosen as the accuracy in the emotion detection and classifying them into the right emotions.

TABLE 5
Confidence Interval

Metric	AUC	Accuracy
Mean Value	0.89	89%
95% Confidence Interval	88.2% - 91.85%	0.89-0.92

V CONCLUSION

This study introduces an innovative model by leveraging the combination of visual descriptors, feature extraction and neural networks for detecting the abnormal events occurring in a crowded environment. The proposed method's effectiveness, demonstrated through the investigative results on the renowned UMN and crowd activities datasets on YouTube, highlights its potential in accurately detecting abnormal crowd behaviours. SIFT and AKAZE descriptors along with the neural networks, along with the impressive performance of the neural network configurations, underscores the robustness of our approach. The patterns of abnormal behaviours are processed collectively to provide meaningful insights to the models for future detection. The input datasets lacked the diversity in terms of environmental, surrounding settings, and cultural contexts. UMN dataset was limited to certain environments, and were subjected to controlled situations and thus may not exhibit the properties and complexities of real time situations. This challenge was addressed by including the scenes and situations from YouTube videos in order to increase the diversity. From the comparative results, the proposed model has delivered promising results and assured as accuracy of 88.5, which is higher than other state of art techniques. The AUC stands at 89.67%, proving that the model has exhibited better results in classification of normal versus abnormal events. The feature extractors proposed in the model were FAST, SIFT and AKAZE, which are manually handcrafted and shall be replaced with automated feature extraction techniques through CNNs. Models such as Long Short-Term Memory (LSTM) or Vision Transformers (ViTs) shall be incorporated to understand the spatio-temporal features with a better perspective. Public safety in highly crowded areas such as malls, stadiums and airports can be further enhanced with the implementation of these automated approaches.

REFERENCES

[1] Manoj Kumar. K, L. Sujihelen. (2022). Recognising Actions with Segmentation and Prediction Techniques in ROI based Deep Learning Framework. Mathematical Statistician and Engineering Applications, 71(4), 4072–4090. https://doi.org/10.17762/msea.v71i4.977

[2] M. K and L. Sujihelen, "Behavioural Analysis For Prospects In Crowd Emotion Sensing: A Survey," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2021, pp. 735-743, doi: 10.1109/ICIRCA51532.2021.9544607.

[3] W. Halboob, H. Altaheri, A. Derhab and J. Almuhtadi, "Crowd Management Intelligence Framework: Umrah Use Case," in IEEE Access, vol. 12, pp. 6752-6767, 2024, doi: 10.1109/ACCESS.2024.3350188.

[4] Liu, D.; Liu, W.; Yuan, X.; Jiang, Y. Conscious and Unconscious Processing of Ensemble Statistics Oppositely Modulate Perceptual Decision-Making. Am. Psychol. 2023, 78, 346–357.

[5] List of Human Stampedes and Crushes, Aug. 2022, [online] Available: https://en.wikipedia.org/wiki/List_of_human_stampedes_and_crushes.

[6] Aljuaid, H.; Akhter, I.; Alsufyani, N.; Shorfuzzaman, M.; Alarfaj, M.; Alnowaiser, K.; Jalal, A.; Park, J. Postures anomaly tracking and prediction learning model over crowd data analytics. PeerJ Comput. Sci. 2023, 9, e1355.

[7] Y. Wang, X. Luo and Z. Zhou, "Contrasting Estimation of Pattern Prototypes for Anomaly Detection in Urban Crowd Flow," in IEEE Transactions on Intelligent Transportation Systems, vol. 25, no. 8, pp. 10231-10245, Aug. 2024, doi: 10.1109/TITS.2024.3355143.

[8] A. M. Al-Shaery et al., "Open Dataset for Predicting Pilgrim Activities for Crowd Management During Hajj Using Wearable Sensors," in IEEE Access, vol. 12, pp. 72828-72846, 2024, doi: 10.1109/ACCESS.2024.3402230.

[9] L. Luo, Y. Li, H. Yin, S. Xie, R. Hu and W. Cai, "Crowd-level abnormal behavior detection via multi-scale motion consistency learning", Proc. AAAI Conf. Artif. Intell., pp. 8984-8992, 2023.

[10] L. Luo, S. Xie, H. Yin, C. Peng and Y. -S. Ong, "Detecting and Quantifying Crowd-Level Abnormal Behaviors in Crowd Events," in IEEE Transactions on Information Forensics and Security, vol. 19, pp. 6810-6823, 2024, doi: 10.1109/TIFS.2024.3423388.

[11] X.-C. Liao, W.-N. Chen, X.-Q. Guo, J. Zhong and X.-M. Hu, "Crowd management through optimal layout of fences: An ant colony approach based on crowd simulation", IEEE Trans. Intell. Transp. Syst., vol. 24, no. 9, pp. 9137-9149, Sep. 2023.

[12] F. Abdullah, M. Abdelhaq, R. Alsaqour, M. H. Alatiyyah, K. Alnowaiser, S. S. Alotaibi, et al.,"Context aware crowd tracking and anomaly detection via deep learning and social force model", IEEE Access, vol. 11, pp. 75884-75898, 2023.

[13] X. Ding, F. He, Z. Lin, Y. Wang, H. Guo and Y. Huang, "Crowd density estimation using fusion of multi-layer features", IEEE Trans. Intell. Transp. Syst., vol. 22, no. 8, pp. 4776-4787, Aug. 2021.

[14] T. H. Noor, "Behavior analysis-based IoT services for crowd management", Comput. J., vol. 66, no. 9, pp. 2208-2219, Sep. 2023.

[15] S. Alsubai et al., "Design of Artificial Intelligence Driven Crowd Density Analysis for Sustainable Smart Cities," in IEEE Access, vol. 12, pp. 121983-121993, 2024, doi: 10.1109/ACCESS.2024.3390049.

[16] L. Luo, Y. Li, H. Yin, S. Xie, R. Hu and W. Cai, "Crowd-level abnormal behavior detection via multi-scale motion consistency learning", Proc. AAAI Conf. Artif. Intell., pp. 8984-8992, 2023.

[17] R. Zhao et al., "Dynamic crowd accident-risk assessment based on internal energy and information entropy for large-scale crowd flow considering COVID-19 epidemic", IEEE Trans. Intell. Transp. Syst., vol. 23, no. 10, pp. 17466-17478, Oct. 2022.

[18] T. Yang, C. Wang, T. Zhou, Z. Cai, K. Wu and B. Hou, "Identification of anomalous behavioral patterns in crowd scenes", Comput. Mater. Continua, vol. 71, no. 1, pp. 925-939, 2022.

[19] S. Yadav, P. Gulia, N. S. Gill and J. M. Chatterjee, "A real-time crowd monitoring and management system for social distance classification and healthcare using deep learning", J. Healthcare Eng., vol. 2022, pp. 1-11, Apr. 2022.

[20] J. Long, W. Liang, K.-C. Li, Y. Wei and M. D. Marino, "A regularized cross-layer ladder network for intrusion detection in

industrial Internet of Things", IEEE Trans. Ind. Informat., vol. 19, no. 2, pp. 1747-1755, Feb. 2023.
[21] Y. Li, Z. Xie, B. Li and M. Mohiuddin, "The impacts of in situ urbanization on housing mobility and employment of local residents in China", Sustainability, vol. 14, no. 15, pp. 9058, Jul. 2022.
[22] G. Yu, S. Wang, Z. Cai, X. Liu, E. Zhu and J. Yin, "Video anomaly detection via visual cloze tests", IEEE Trans. Inf. Forensics Security, vol. 18, pp. 4955-4969, 2023.
[23] M. Zhang, T. Li, Y. Yu, Y. Li, P. Hui and Y. Zheng, "Urban anomaly analytics: Description detection and prediction", IEEE Trans. Big Data, vol. 8, no. 3, pp. 809-826, Jun. 2022.
[24] R. Lalit and R. K. Purwar, "Crowd abnormality detection using optical flow and GLCM-based texture features", J. Inf. Technol. Res., vol. 15, no. 1, pp. 1-15, Jun. 2022.
[25] T. N. Nguyen and S. Zeadally, "Mobile crowd-sensing applications: Data redundancies challenges and solutions", ACM Trans. Internet Technol., vol. 22, no. 2, pp. 1-15, May 2022.
[26] S. Zhang, Y. Yang, W. Liang, V. K. A. Sandor, G. Xie and K. R. Choo, "MKSS: An effective multi-authority keyword search scheme for edge-cloud collaboration", J. Syst. Archit., vol. 144, 2023.
[27] P. Martí, L. Serrano-Estrada, A. Nolasco-Cirugeda and J. L. Baeza, "Revisiting the spatial definition of neighborhood boundaries: Functional clusters versus administrative neighborhoods", J. Urban Technol., vol. 29, no. 3, pp. 73-94, Jul. 2022.
[28] C. Chen et al., "Comprehensive regularization in a bi-directional predictive network for video anomaly detection", Proc. AAAI Conf. Artif. Intell., vol. 36, no. 1, pp. 230-238, Jun. 2022.
[29] L. Deng, D. Lian, Z. Huang and E. Chen, "Graph convolutional adversarial networks for spatiotemporal anomaly detection", IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 6, pp. 2416-2428, Jun. 2022.
[30] T. Yang, C. Wang, T. Zhou, Z. Cai, K. Wu and B. Hou, "Identification of anomalous behavioral patterns in crowd scenes", Comput. Mater. Continua, vol. 71, no. 1, pp. 925-939, 2022.
[31] Y. Liu, X. Liu, X. Li, M. Li and Y. Li, "Participants recruitment for coverage maximization by mobility predicting in mobile crowd sensing", China Commun., vol. 20, no. 8, pp. 163-176, Aug. 2023.
[32] Y. Chen and A. Deng, "Using POI data and Baidu migration big data to modify nighttime light data to identify urban and rural area", IEEE Access, vol. 10, pp. 93513-93524, 2022.
[33] C. Cao, Y. Lu and Y. Zhang, "Context recovery and knowledge retrieval: A novel two-stream framework for video anomaly detection", IEEE Trans. Image Process., vol. 33, pp. 1810-1825, 2024.
[34] C. H. Liu et al., "Modeling citywide crowd flows using attentive convolutional LSTM", Proc. IEEE 37th Int. Conf. Data Eng. (ICDE), pp. 217-228, Apr. 2021.
[35] J. Zhang and X. Zhang, "Multi-task allocation in mobile crowd sensing with mobility prediction", IEEE Trans. Mobile Comput., vol. 22, no. 2, pp. 1081-1094, Feb. 2023.
[36] H. Wang, S. Zeng, Y. Li and D. Jin, "Predictability and prediction of human mobility based on application-collected location data", IEEE Trans. Mobile Comput., vol. 20, no. 7, pp. 2457-2472, Jul. 2021.
[37] T. Liu, C. Zhang, K.-M. Lam and J. Kong, "Decouple and resolve: Transformer-based models for online anomaly detection from weakly labeled videos", IEEE Trans. Inf. Forensics Security, vol. 18, pp. 15-28, 2023.
[38] G. Woo, C. Liu, D. Sahoo, A. Kumar and S. Hoi, "CoST: Contrastive learning of disentangled seasonal-trend representations for time series forecasting", Proc. Int. Conf. Learn. Represent., pp. 1-18, 2022.
[39] X. Liu and J. Liu, "A truthful double auction mechanism for multi-resource allocation in crowd sensing systems", IEEE Trans. Serv. Comput., vol. 15, no. 5, pp. 2579-2590, Sep./Oct. 2022.
[40] P. Wu et al., "VadCLIP: Adapting vision-language models for weakly supervised video anomaly detection", Proc. AAAI Conf. Artif. Intell., pp. 6074-6082, 2024.
[41] Y. Yang, C. Zhang, T. Zhou, Q. Wen and L. Sun, "DCdetector: Dual attention contrastive representation learning for time series anomaly detection", Proc. 29th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), pp. 3033-3045, 2023.
[42] S. Zhang, J. He, W. Liang and K. Li, "MMDS: A secure and verifiable multimedia data search scheme for cloud-assisted edge computing", Future Gener. Comput. Syst., vol. 151, pp. 32-44, 2024.
[43] Franzoni, V., Biondi, G. & Milani, A. Emotional sounds of crowds: spectrogram-based analysis using deep learning. Multimed Tools Appl 79, 36063–36075 (2020). https://doi.org/10.1007/s11042-020-09428-x
[44] Jignesh Vaniya, Safvan Vahora, Uttam Chauhan, Sudhir Vegad, "Crowd Emotion and Behavior Analysis Using Lightweight CNN Model," SSRG International Journal of Electrical and Electronics Engineering, vol. 11, no. 10, pp. 30-46, 2024. Crossref, https://doi.org/10.14445/23488379/IJEEE-V11I10P104
[45] R. Kamal, M. A. Shah, C. Maple, M. Masood, A. Wahid and A. Mehmood, "Emotion Classification and Crowd Source Sensing; A Lexicon Based Approach," in *IEEE Access*, vol. 7, pp. 27124-27134, 2019, doi: 10.1109/ACCESS.2019.2892624.
[46] Monish, L. et al. (2024). Emotion Prediction Based on Real-Time Crowd Analysis Using Deep Network. In: Mahapatra, R.P., Peddoju, S.K., Roy, S., Parwekar, P. (eds) Proceedings of International Conference on Recent Trends in Computing.

Manojkumar K is currently pursuing his Doctor of Philosophy (Ph.D.) at Sathyabama Institute of Science and Technology, Chennai, where he is actively engaged in advanced research in the field of Computer Science and Engineering. He obtained his Master of Engineering (M.E.) degree in Computer Science and Engineering from Anna University in 2013, equipping him with a strong foundation in computing technologies and research methodologies. His primary areas of interest encompass Data Science, Machine Learning, Crowd Emotion Sensing, and Image and Video Analytics. Demonstrating his commitment to professional development and global research communities, he has been a member of the International Society for Research and Development (ISRD) since 2016 and the International Association of Engineers (IAENG) since 2017. His active participation in these organizations reflects his passion for knowledge sharing and academic excellence.

Dr. L. Sujihelen holds both a Master of Engineering (M.E.) and a Doctor of Philosophy (Ph.D.) in Computer Science and Engineering from Sathyabama University, India. With a strong academic background and extensive research experience, she has made significant contributions to various domains of computer science. Her primary research interests include Image Processing, Wireless Sensor Networks (WSN), the Internet of Things (IoT), and Cloud Computing. She has actively engaged in research projects and publications that explore advancements in these fields. Currently, she serves as an Associate Professor in the Department of Computer Science and Engineering at the Faculty of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai. Her dedication to academics and innovation continues to shape the next generation of engineers and researchers in the field of computing.