



*Research article*

## **Audio-visual multi-modality driven hybrid feature learning model for crowd analysis and classification**

**H. Y. Swathi<sup>1,\*</sup> and G. Shivakumar<sup>2</sup>**

<sup>1</sup> Department of Electronics and Communication Engineering, Malnad College of Engineering, Visvesvaraya Technological University, Belagavi, India

<sup>2</sup> Department of Electronics and Communication Engineering, AMC Engineering College, Visvesvaraya Technological University, Belagavi, India

\* **Correspondence:** Email: [hys@mcehassan.ac.in](mailto:hys@mcehassan.ac.in); Tel: 9964258515.

**Abstract:** The high pace emergence in advanced software systems, low-cost hardware and decentralized cloud computing technologies have broadened the horizon for vision-based surveillance, monitoring and control. However, complex and inferior feature learning over visual artefacts or video streams, especially under extreme conditions confine majority of the at-hand vision-based crowd analysis and classification systems. Retrieving event-sensitive or crowd-type sensitive spatio-temporal features for the different crowd types under extreme conditions is a highly complex task. Consequently, it results in lower accuracy and hence low reliability that confines existing methods for real-time crowd analysis. Despite numerous efforts in vision-based approaches, the lack of acoustic cues often creates ambiguity in crowd classification. On the other hand, the strategic amalgamation of audio-visual features can enable accurate and reliable crowd analysis and classification. Considering it as motivation, in this research a novel audio-visual multi-modality driven hybrid feature learning model is developed for crowd analysis and classification. In this work, a hybrid feature extraction model was applied to extract deep spatio-temporal features by using Gray-Level Co-occurrence Metrics (GLCM) and AlexNet transferrable learning model. Once extracting the different GLCM features and AlexNet deep features, horizontal concatenation was done to fuse the different feature sets. Similarly, for acoustic feature extraction, the audio samples (from the input video) were processed for static (fixed size) sampling, pre-emphasis, block framing and Hann windowing, followed by acoustic feature extraction like GTCC, GTCC-Delta, GTCC-Delta-Delta, MFCC, Spectral Entropy, Spectral Flux, Spectral Slope and Harmonics to Noise Ratio (HNR). Finally, the extracted audio-visual features were fused to yield a composite multi-modal feature set, which is processed for classification using the

random forest ensemble classifier. The multi-class classification yields a crowd-classification accuracy of (98.26%), precision (98.89%), sensitivity (94.82%), specificity (95.57%), and F-Measure of 98.84%. The robustness of the proposed multi-modality-based crowd analysis model confirms its suitability towards real-world crowd detection and classification tasks.

**Keywords:** multi-modal crowd analysis; deep-spatio-temporal features; acoustic features; ensemble learning; audio-visual crowd classification

---

## 1. Introduction

The last few decades have witnessed significantly high-paced growth in software technologies, decentralized computing, cloud-infrastructures, low-cost hardware and sensing technologies to cope with monitoring and surveillance purposes. Whether it is industrial monitoring and control or any strategic surveillance task(s) serving defense purposes or civic surveillance etc. vision computing has a vital role to play [1]. Undeniably, the development in decentralized cloud computing and wireless communication technologies has broadened the horizon for vision-based surveillance systems to serve real-time decision making. Majority of these surveillance systems employ vision-computing, also called vision-based computing techniques by applying camera(s) as sensor to collect a real-time video stream for continuous monitoring and control [2]. However, there are a large number of application environments where there can be significantly large sensors deployed across the region of interests such as defense systems, broader security, civic surveillance, industrial surveillance etc., where detecting certain specific kind of activity by manual mapping and detection becomes infeasible [2,3]. In other words, to cope with the customized activity detection and reactive decision making, an automated video analysis approach is inevitable [3].

On the other hand, there are a number of surveillance environments where abnormal event detection or crowd analysis becomes vital [1]. The recent innovations in vision techniques have enabled activity detection such as moving pedestrian detection, moving vehicle detection and tracking, moving object detection etc. These vision-based object detection and tracking models apply spatio-temporal motion vector or features (say, cues) to perform intended task [1,3]. However, there are purposes such as abnormal activity detection, crowd detection, crowd type identification which require time-efficient and accurate video analysis [1,4,5]. In the last few years, the events of terrorism, bomb-blasts, communal violence, lynching etc. have increased significantly [5]. On the other hand, the manual video analysis approaches over largely deployed cameras can be a difficult task and can undergo delayed response or even human assessment related errors giving rise to the disaster [5]. Identifying abnormal crowd with different crowd nature indoor or outdoor (i.e., marketplace, public gathering spots, religious places, sports auditorium, stadium, entertainment destinations, schools, social gathering places etc.) is a highly difficult task [5–7]. In sync with these problems, in the last few years a research domain called crowd-behavior analysis has gained widespread attention across academia-industries.

Crowd behavior analysis is the process to assess crowd inputs such as video streams or audio samples statistically to identify or classify event types or crowd types, so that proactive and timely decisions could be made [1,4–7]. Undeniably, in reference to the aforesaid events (terrorism, bomb-

blasts, communal violence, lynching, prayer, celebration, community activities etc.) classifying crowd or allied event types can help security agencies to make optimal decision making [7].

To keep up with aforesaid contemporary demands, in the past a few efforts have been made towards crowd behavior analysis and anomaly event detection [4–7]. A majority of the existing systems are developed for video-based analysis where the spatio-temporal or deep features from input video streams are examined to detect abnormal events or event types [1,4–7]. The at hand solutions intend to assess high-level descriptive feature assessment to detect and identify a person's movement pattern and crowd likelihood assessment. However, detecting crowd over randomly moving with high-level occlusion becomes a mammoth task [7,8]. Moreover, under such complex environment, most of the classical spatio-temporal feature driven models undergo high false positive performance. In the past, a few efforts have been made where authors have exploited inter-personal behavioral changes to perform anomaly detection [3–9] and crowd-likelihood prediction [10,11].

Typically, crowd analysis or crowd-type prediction can be achieved by means of two approaches; first video-feature driven methods, and second acoustic cues driven models. As the name indicates, video feature-based method exploits aforesaid spatio-temporal cues to perform person's movement pattern analysis and allied crowd likelihood assessment. On the contrary, acoustic-based methods are typically audio-feature based approaches. However, so far, these two modalities have been addressed distinctly. Though, relatively larger number of efforts are made towards vision-based crowd analysis methods; the fraction of acoustic driven approaches is lesser. In vision-based approaches, authors [12] have performed target segmentation, which is followed by segmented targets behavioral analysis to perform crowd identification or prediction [12]. Unfortunately, such approaches can be limited especially over the fast-moving multiple people under occlusion. Other approaches like optimal flow analysis [8,9,13–15] have been applied to estimate optical flow histogram to detect global motion vector, which is later applied for crowd behavior analysis. However, computational complexities of these methods cap their scalability. Though, to alleviate it, authors [8] suggested to use low-level motion features (for region segmentation); however, its efficacy over the different crowd types with random movement patterns or ambiguous pattern can be limited [16]. In sync with such at-hand limitations, authors [16] developed a probabilistic crowd event detection concept where authors trained their model over the different user's activity patterns like running, walking, merging, splitting, local dispersion and evacuation. Though, the use of crowd motion representation and allied tracklets models performed better [17,18]; yet, detecting multiple descriptive spatio-temporal features (say, motion features for the different persons simultaneously) over streaming video can yield ambiguous feature vector and hence can yield false positive performance [17,19].

A few researchers found that the use of inter-personal visual interaction feature can help predicting crowd types [4,18,20,21]. However, assessing correlation dimensions over the multiple moving objects and disjoint as well as correlated feature extraction and learning [21] can confine their suitability. Such approaches can be well-suited for small and fixed volume of human presence in search space or field view. Action-energy [22] approaches too can be limited and highly false-prone under dense crowd with different person's behave or actions [20,23]. Despite the use of static and dynamic agent driven group behavior assessment [23], their efficacy remains suspicious and ungeneralizable over real-time applications. Approaches like local motion patterns [18] or mid-level spatio-temporal features [24] are found limited over the different crowds having similar or near-similarity posture but different intends (Ex. Prayer and cheering up events in indoor facilities). Similar to these methods, the classical template-based models too can be limited and inaccurate over realistic application

environments [25]. However, entropy- and texture-driven methods have exhibited higher accuracy [25]. Recently, it was reported that a mixture of dynamic textures [5] and other spatio-temporal textural feature compositions [6,26] can yield superior performance for fast-moving-object analysis in video streams. This can help improve crowd behavior analysis and classification. Studies have revealed that merely exploiting the spatio-temporal features, entropy, or allied correlation information of a frame cannot yield accurate crowd analysis and specific crowd detection [27]. Moreover, audio-driven approaches as standalone solutions can produce false positives, especially for crowd classification, owing to noise, ambiguity, and ambiguous acoustic similarity [27]. Neither vision-based nor audio-based standalone solutions are optimal for crowd analysis and counting, especially under extreme conditions. However, scientific neurobiological investigations indicate that ear and eye sensing together can provide an accurate perception that can be used in decision making [27]; therefore, the use of time-domain acoustic signals and corresponding spatio-temporal video features can yield a superior and realistic solution for crowd analysis or crowd classification. The use of audio-visual cues (together) can retain auditory information that can act as an auxiliary cue for each video input to ensure accurate crowd analysis [27,28]. Unfortunately, no significant effort has been made to study audio-visually driven crowd analysis tasks.

Considering the above inferences as motivation, in this study, highly robust and efficient multi-modality (audio-visual)-driven crowd analysis model was developed. As the name indicates, the proposed model embodies both audio signals and video spatio-temporal features as composite feature vectors for performing crowd analysis. To ensure optimal intrinsic information-rich feature learning, deep spatio-temporal features from input video sequences and the corresponding acoustic features were used to perform feature learning and classification. Specifically, the input surveillance video sequences were transformed into video frames and audio (\*.mp3) samples for each video frame or sequence. Along with video feature extraction, unlike in classical textural feature-driven models, deep spatio-temporal features were exploited using the gray-level co-occurrence matrix (GLCM) and AlexNet. The purpose of applying the hybrid deep spatio-temporal textural feature (STTF) model was to retain the maximum feature diversity with a deep intrinsic feature vector to guarantee optimal performance and reliability. The use of the GLCM made it possible to extract six different STTF features that were amalgamated (horizontal concatenation fusion) with AlexNet deep features obtained with five convolutional layers (CONV) and three fully connected (FC) layers. Thus, the use of the deep STTF feature model provides visual feature vectors for further learning. For each video sequence, acoustic samples were obtained in the form of a \*.mp3 audio frame, which was processed for static (fixed size) sampling, pre-emphasis, block framing, and Hann windowing. These processes act as preprocessing steps. Subsequently, for each segmented audio sample, different acoustic features were obtained, including the gammatone cepstral coefficient (GTCC), GTCC-Delta, GTCC-Delta-Delta, Mel frequency cepstral coefficient (MFCC), pitch, harmonics-to-noise ratio (HNR), and other spectral features. After the aforementioned acoustic features were extracted from the complete audio samples over the input video sequence, the acoustic features were averaged to generate a composite (acoustic) feature vector. Once the visual (i.e., deep STTF features) and acoustic (composite acoustic features) features were extracted, they were horizontally concatenated to model a fused audio-visual feature vector. This audio-visual feature vector was processed for multiclass classification using a random forest ensemble algorithm. To assess the efficacy of the proposed multi-modal crowd analysis system, different input samples were considered from various categories, such as prayers, sports game cheers, and quarreling.

The MATLAB-based simulation model revealed that, compared with classical standalone feature-based crowd analysis models, the proposed model exhibited superior accuracy (98.26%), precision (98.89%), sensitivity (94.82%), and specificity (95.57%), as well as an F-measure of 98.84%. The proposed multi-modality-driven approach was found to be superior to previous methods, including vision- or audio-based standalone solutions. The robustness of the proposed multi-modality-driven crowd analysis model confirms its suitability for real-world crowd detection and classification solutions for reliable automated surveillance purposes.

The remainder of this article is organized as follows. In Section II, related work on different crowd analysis models is discussed, and the research questions are presented in Section III. In Section IV, the research method and allied implementation are described. The simulation results and inferences based on them are discussed in Section V. Finally, the conclusions and scope of future work are discussed in Section VI. The references cited in this article are provided at the end.

## 2. Related work

In a previous study [17], the histogram of oriented tracklet descriptors (HOTD) from an input video stream was used to perform abnormal event detection. After HOTD feature vectors were extracted, a support vector machine (SVM) classifier was applied to detect abnormal events or crowds. In another study [20], instead of HOTD, structural analysis, such as a structural context descriptor, was employed to classify abnormal events or crowds [20]. A motion-oriented gradient was used to perform crowd analysis over sparse crowd input. Similar to the other study [17], it was found that the SVM classifier outperformed neural-network methods [29]. In another research [29], the learning of dynamic and temporal cues was applied to a 3D representation, where a convolutional neural network (CNN) was applied (3DS-CNN) to extract features and learn. Similarly, in another work [30], a CNN with a recurrent CNN (RCNN) was employed as a hybrid deep model for crowd analysis. The main motive behind using the hybrid deep model was to reduce the computational time. An improved deep and spatio-temporal-feature-driven crowd analysis method was proposed [31]. In this approach, a deep spatio-temporal perspective was applied in the form of displacement information of crowd motion patterns. The displacement information was applied as a high-level feature representation to train a convolutional network for crowd analysis. In another study [32], CNNs with auto-encoders and R-CNNs were used to perform crowd classification or analysis. Similarly, in other research [33], a fine-tuned residual network was used to extract features. The extracted features were trained using a one-nearest-neighbor classifier to classify the crowd type.

In a previous study [34], the efficacy of a linear SVM in crowd analysis was assessed, and it was found that the linear SVM was superior to the K-nearest-neighbor (KNN) and random forest classifiers. The moving object or person trajectory information was processed using a correlation clustering algorithm, which was later employed using a structural SVM to perform trajectory analysis [35]. However, the study failed to address crowds in the real world. Despite using random forest classification, another proposed model [36] could not address real-time crowd analysis problems. In another work, a machine-learning-driven crowd analysis concept was designed [37]. Other researchers [38] applied a cubic kernel SVM and subspace KNN separately to perform crowd analysis; however, they did not address the need for a more feature-intensive approach, which could have yielded superior accuracy and reliability. In another study [39], the focus was mainly on reducing the time required for rejecting motion outliers or the allied crowd analysis.

In other research [40], a spatio-temporal sparse coding representation of input video sequences was applied for crowd analysis. In this approach, sparsely coded features [41] were trained using K-means clustering and an SVM to perform crowd analysis. It was concluded that SVM classification is superior to the K-means-clustering-driven analysis model. In another study [42], the Spearman distance information was applied to a naive Bayesian classifier to perform crowd analysis. Spatio-temporal features were extracted from an input video to detect crowd abnormalities [43]. Here, a nonlinear SVM was applied to perform the classification. Specifically, the authors applied machine learning and a threshold-driven model for crowd analysis [43]. Regions with a CNN (RCNN) model were applied to input video streams for crowd analysis and classification (normal or abnormal crowds) [44]. Other researchers [44–48] exploited spatio-temporal features over a three-dimensional grid structure to perform crowd analysis. To fine-tune performance, an artificial bacterial colony was used as a feature selection model [49]. Using different spatio-temporal features, an artificial bacterial colony heuristic was used to retain the most significant features for crowd analysis and classification. In another study [50], change detection was performed, in which motion feature extraction was applied to the input video for crowd analysis and classification. It used a triplet network for motion feature extraction. However, its reliability remains questionable under dynamic and multiperson presence and movements within the same frame.

A local directional strength pattern (LDSP) with local directional rank histogram pattern (LDRHP) features, which were trained with a CNN to perform crowd analysis, was used in other research [51]. In another study [52], locally consistent scale priors and global occlusion reasoning features were employed for crowd analysis in video analysis [52]. Despite these efforts, the main emphasis was pedestrian behavior analysis [53]. Other researchers [54] developed an interactive crowd behavior learning model to assess crowd anomaly detection during virtual-world training [54]. Spatio-temporal features were used in other work [55], focusing on crowd sensing and behavior analysis. A similar effort was made [56] in which spatio-temporal feature descriptors were obtained from input videos to perform crowd behavior analysis. In other research [56,57], coherent motion patterns were applied by employing a structural trajectory learning concept. These patterns were subsequently used for learning crowd behavior analysis. A user's individual acceleration features were applied to perform crowd analysis; however, the scalability remains questionable in real-world applications [58]. To alleviate such issues, other researchers [59] focused on granular computing (GrCS), which helps in crowd segmentation, followed by feature learning for classification. To simplify this task, others considered global features as crowd descriptors [60]; however, they failed to characterize the type of crowd so it acted merely as a threshold-based crowd-sensing model. In another study [61], divergent centers were estimated over consecutive frames to detect crowd anomalies; however, the scalability under real-time operating conditions remains questionable. In contrast to these approaches, a social-attribute-aware force model (SAFM) was designed to detect abnormal crowd behavior in video sequences [62]. To improve efficacy, researchers applied multiple cameras for group crowd analysis [63]. Micro-Doppler (MD) signatures retrieved from a low-power radar device were used to identify universal crowd preference sensing for real-time decision making [62]. In other work [64] a transfer deep-learning method was used to perform activity detection and analysis. The swarm intelligence concept was used to detect the event of interest from a crowd spatio-temporal feature space [65]. However, to improve the computational efficacy, other researchers [41] applied sparse features; however, they could not address multiple crowd identification under different crowd patterns.

In one study [27], it was concluded that extracting significant spatio-temporal features from low-quality video streams is a highly difficult task that can confine the efficacy of the at-hand solution. To alleviate such problems, an audio-visual multi-scale network (AVMSN) was developed that exploits both audio and video cues from each input stream to perform crowd analysis. In this approach, sample convolutional blocks were used as the multi-scale vision-end branch to extract features that made it possible to estimate the weighted-visual features. In addition, audio features were obtained in the temporal domain by exploiting the spectrogram information, and the audio features were learned. In this study, a separate audio-VGG network was applied to learn the audio data. The weighted visual and audio features were amalgamated using the multi-modal fusion concept by applying a cascade fusion paradigm for the estimated density map calculation. The overall approach for real-time solution fitting is highly complex and exhaustive.

Researchers designed a CapsNet-based approach for crowd analysis or crowd counting [66]. They claimed that, unlike CNN-based methods, CapsNet can enable high-capacity feature representation; hence, a high-quality density map and accuracy can be guaranteed. In another study [67], the focus was on amalgamating the short- and long-time analyses of audio signals to detect impulsive and sustained events in a real-time crowded environment. In another work [68], a feature called the “mixture of kernel dynamic textures” was extracted from the input video stream. In this approach, the feature models the appearance and dynamics across the scene, and the result is later used for abnormal crowd detection or outlier detection. Because of the limitations of feature extraction and learning over the audio-visual feature space, researchers [28] proposed the notion of auxiliary and explicit image patch-importance ranking (PIR) and patch-wise crowd estimate (PCE) information to generate a run-time solution. These audio-visual modalities undergo transformer-inspired cross-modality co-attention mechanisms, resulting in crowd estimation. However, despite the claim of superior (33.8%) efficacy, the computational burden over multiple object-driven streams appears limited or time consuming.

In another report [69], it was suggested that the use of audio features, such as the zero-crossing rate, MFCC, and its combination with a hidden Markov model (HMM) and an SVM classifier can yield multiple crowd (type) classifications. The F-measure observed in the study was only 80.6%, which must be improved further to make it a realistic solution. In other work [19], the use of multiple modalities, including audio, video, pictures, and text, to perform crowd analysis for social security was proposed. Despite the claim that the performance was superior to that of genetic algorithms (GA) based methods, the robustness was not assessed using real-time audio-visual datasets. Other researchers [70] obtained the acoustic feature MFCC from an input soccer sound sequence, which was followed by windowing to achieve homogenous component segmentation. It was asserted that windowing eliminated the need to define a heuristic set of rules for audio segmentation. Each segmented audio was labeled using a series of HMM classifiers, each a representation of one of six predefined semantic content classes found in the soccer video. It is a typical audio-based model, but it fails to address ambiguities caused by similar acoustic disturbances. In this case, the use of both audio and video features can yield superior performance, provided that computational cost efficiency and high accuracy are maintained. This is the driving force in the present study.

### 3. Research question

Considering the overall research intentions and scope of this work, a few questions were framed to be addressed in this study. The questions addressed in this study are as follows.

- RQ1: Can the use of a multi-modal feature environment (audio and video features) yield superior feature vectors for accurate and reliable crowd analysis and classification?
- RQ2: Can the strategic amalgamation of visual features encompassing GLCM features and transferrable deep features or AlexNet features with acoustic features (possessing MFCC, GTCC, GTCC-Delta, GTCC-Delta-Delta, pitch, harmonics, and other spectral information) yields an optimal (accuracy, scalability, veracity and realizability) multi-modality-driven crowd analysis and classification system?
- RQ3: Can the use of the random forest classifier rather than the aforementioned (RQ2) multi-modality feature environment yield optimal solutions for crowd analysis and classification?
- This research was conducted to obtain the optimal answers to these questions.

#### 4. System model

Unlike classical vision-based crowd analysis methods, in this study, it was hypothesized that the use of visual features and acoustic cues or allied features can make a learning model more efficient and reliable. This hypothesis assumes that perceptions with the ear and eyes together make a more accurate real-time decision possible than one based on a standalone approach. Moreover, unlike traditional shallow spatio-temporal feature-driven video analysis approaches, the use of deep spatio-temporal features is more appropriate for crowd analysis. In this study, these hypotheses were the driving force for designing a state-of-the-art novel and robust audio-visual multi-modality-driven hybrid feature-learning model for crowd analysis and classification. As the name indicates, “multi-modality” implies a combination of two different feature sets or cues to perform crowd detection and classification.

In this work, audio and video spatio-temporal features from a streaming input video were used for crowd analysis and classification. To ensure the optimal reliability and scalability of the system under different crowd conditions, unlike traditional shallow feature-driven models, in this study, the focus was on exploiting the maximum possible significant features over video data and corresponding speech (audio) samples. Realizing that, under extreme dynamic conditions, such as crowds with multiple objects or persons moving randomly at different speeds and occlusions (with unpredictable behavior), textural or energy information alone cannot provide an optimal target behavioral analysis; hence, such approaches can make crowd analysis error-prone or inaccurate. To alleviate such issues, different features were obtained separately from audio and video inputs. They were later fused or horizontally concatenated to yield a composite feature vector for learning-based prediction. The composite feature vector encompassing audio-visual feature instances was trained using a random forest to perform multiclass classification. To assess scalability or reliability, different types of crowd-related video sequences including prayers, sports crowd, praising, and quarreling were considered.

Once different crowd video streams were collected, the input video sequences were processed for allied audio separation. In other words, the input \*.mp4 video sequences were processed to extract \*.mp3 audio streams, which were processed for feature extraction and learning. Because it is a multi-modality-driven approach, features were extracted separately from the video streams and the corresponding audio samples. Considering feature heterogeneity and its impact on the overall learning and classification results, deep features and spatio-temporal textural features, such as the GLCM, were extracted. In other words, from each input video data, AlexNet CNN-driven high-dimensional deep features were extracted along with the GLCM-based STTF features. In the proposed model, the AlexNet deep network was designed with five CONV and three FC layers to extract 4096-dimensional



features from each input video sequence. Similarly, the GLCM, a well-known STTF feature extraction model, was applied to the video sequence to extract different features, including contrast, correlation, energy, homogeneity, mean, standard deviation, skewness, and kurtosis. Thus, after the different STTF GLCM features with AlexNet deep features were extracted, they were amalgamated to generate a composite feature vector for the video input. However, the audio samples extracted from the video streams were processed for acoustic feature extraction. To ensure noise-free and accurate feature learning, basic pre-processing tasks were performed, including pre-emphasis, blocking, and Hann-windowing, to segment the audio samples. Subsequently, different acoustic features were extracted, including the GTCC, GTCC-delta, GTCC-delta-delta, MFCC, spectral entropy, spectral flux, spectral slope, and HNR to train the classifier model. After the deep-STTF features (i.e., GLCM and AlexNet-CNN features) were extracted from the videos, as well as the corresponding acoustic features, horizontal concatenation was performed to achieve a complete composite feature vector, which was later processed using the random forest ensemble classifier for multi-class classification. Thus, the proposed work encompasses the following steps:

*Step 1: Data acquisition*

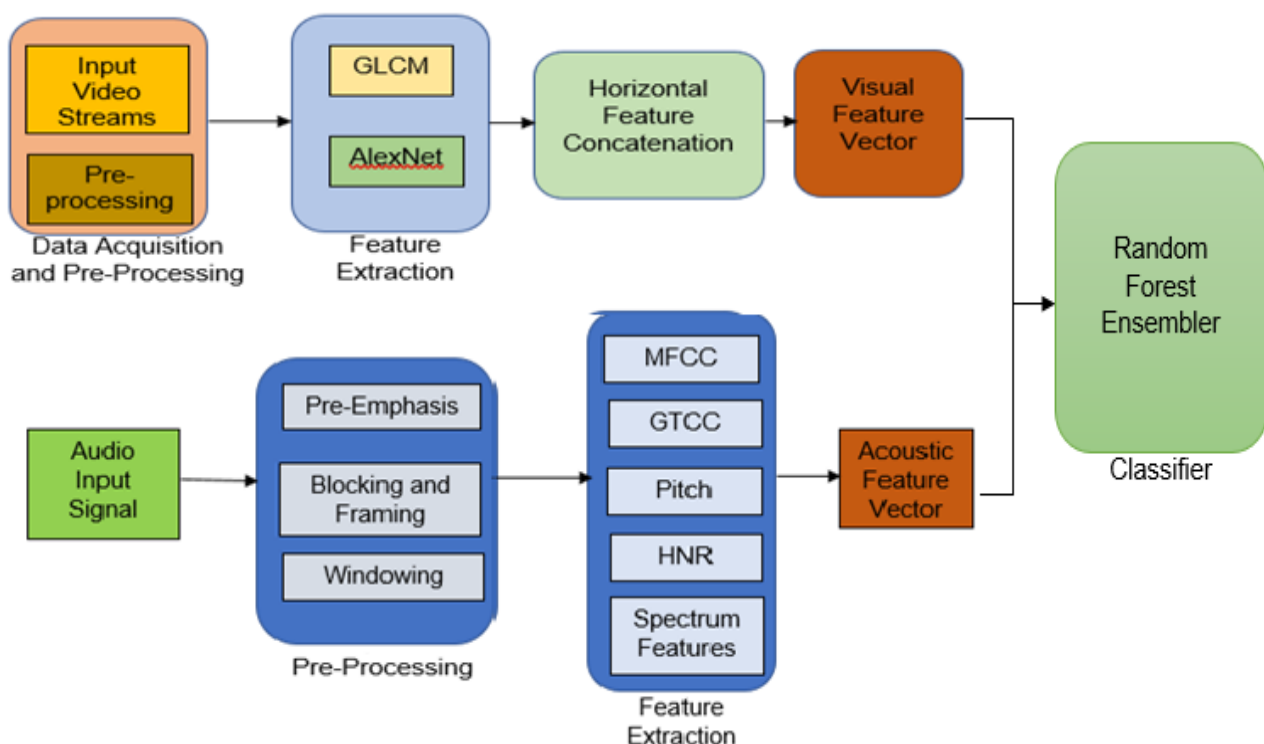
*Step 2: Video and audio data separation*

*Step 3: Deep-STTF (AlexNet-CNN and GLCM) feature extraction*

*Step 4: Acoustic feature extraction*

*Step 5: Random Forest ensemble learning.*

A block diagram of the proposed model is depicted in Figure 1.







**Figure 1.** Proposed model.

The proposed model is discussed in detail in the subsequent sections.




#### 4.1. Data acquisition

As stated above, to ensure the scalability of the proposed crowd analysis, crowd samples from different categories were considered, including praising, sports, cheering, prayers, violence, and quarrels. Most existing video analysis models have applied distinct types of single video stream - for example, news data and pedestrian movement video traces. However, training a model with single type of crowd behavior cannot be suitable for a “fit-to-all” crowd analysis problem. Therefore, the model was trained using datasets of different types (i.e., different crowd characteristics). To obtain these inputs, different benchmark datasets were considered, including the UMN and PETS databases. In addition to the aforementioned data traces, other datasets prepared from YouTube videos by performing basic video conversion or cropping tasks were considered. These all-input videos were considered at 30 frames per second (fps) to make it possible to use the proposed model with real-time supervision cameras. Snippets of different crowd datasets and human behavioral presentations are presented in Table 1.

**Table 1.** Snippets of random frames from the different crowd video datasets and respective ground truth.

Sample	Random Frame	Ground Truth Crowd Type
1.		Praising
2.		Normal
3.		Violence
4.		Cheering

*Continued on next page*

Sample	Random Frame	Ground Truth Crowd Type
5.		Prayer
6.		Protest
7.		Cheering

#### 4.2. Video and audio separation

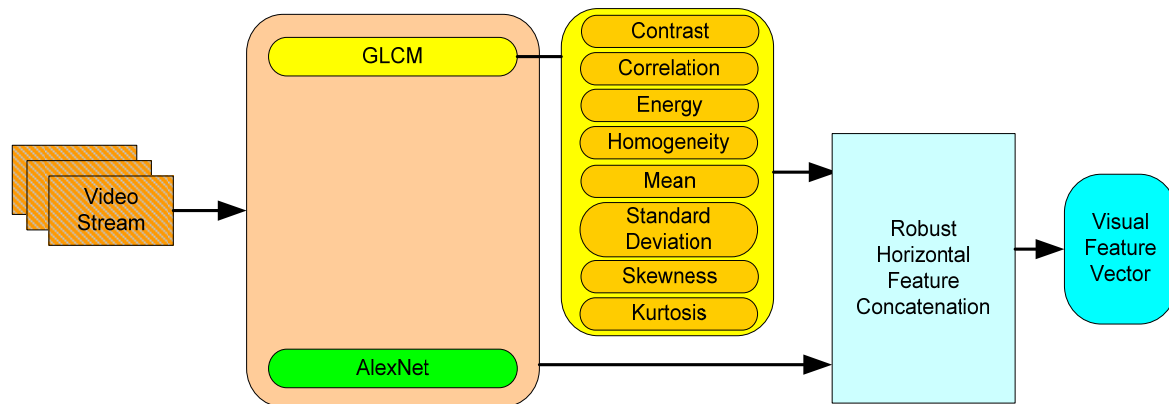
In this study, the input data used were in the typical \*.mp4 format, which is processed for audio segmentation. In this study, the audio extraction concept was used, in which video streams are converted into audio files. Numerous software packages that can convert video into audio files are available. Specifically, the input video streams (\*.mp4 and \*.AVI) were converted to \*.mp3 formats. Once the set of videos and corresponding audio samples was obtained, feature extraction was executed for the inputs (i.e., video and audio files) separately. Because the feature concept is multi-modal, different feature extraction methods were applied for the video and audio inputs. After feature extraction, the extracted features were merged or fused to obtain a composite feature vector for further learning and classification. A detailed discussion of the feature extraction methods used is provided in the subsequent sections.

#### 4.3. Deep-STTF (AlexNet-CNN and GLCM) driven visual feature extraction

As stated, being a multi-modal crowd-analysis concept, we consider video streams (say, consecutive frames) as well as corresponding audio (acoustics) for feature extraction and learning. However, the feature extraction methods for visual and audio data or samples are different. Here, we have proposed to use a hybrid deep-spatio temporal feature extraction model for video related feature extraction, while acoustic features are obtained from the audio-samples. The overall feature extraction method involved in the proposed research is depicted in Figure 2.

Because the crowd analysis concept is multi-modal, video streams (e.g., consecutive frames) were considered, as well as corresponding audio (acoustics), for feature extraction and learning. However, the feature extraction methods for visual and audio data or samples are different. Herein, the use of a

hybrid deep spatio-temporal feature extraction model for video-related feature extraction is proposed, and acoustic features are obtained from the audio samples. The overall feature extraction method is shown in Figure 2.



**Figure 2.** Deep-STTF feature vector preparation for visual data.

#### 4.3.1. GLCM STTF feature extraction

In this study, GLCM features act as descriptive feature models characterizing the likelihood of the pixel's gray-scale values over the input video frames representing crowd scenes. In this function, high-dimensional descriptive features encompassing energy, contrast, correlation, homogeneity, mean, standard deviation, kurtosis, and skewness are extracted. To extract the STTFs using the GLCM, the video sequences were converted into multiple frames. To accommodate real-time surveillance systems and their demands, videos were processed at 30 fps. Similar to GLCM feature extraction, it was hypothesized that the aforementioned textural (descriptive) features are distributed uniformly across the input video frame(s). Thus, for each consecutive frame, the GLCM algorithm was applied to extract various STTF features, including energy, contrast, correlation, homogeneity, mean, standard deviation, kurtosis, and skewness, which were subsequently horizontally concatenated to obtain a composite feature vector. In this study, STTF GLCM features were obtained in the form of a matrix that signifies the pixel intensities  $I(x, y)$ , centered on the  $(x, y)$  pixel. Here, the feature extraction takes place in such a manner that, over each consecutive input video frame, it provides a distinct probability matrix  $P_{i,j}$ , signifying the intensity disparity between the  $i$  – th and the  $j$  – th pixels that makes it possible to detect motion pattern(s) in each input frame. In the GLCM method, the gray scale signifies the pair relationship in the same direction; hence, obtaining the gray-scale values can result in a matrix depicting the relationship matrix among the different pixels in the direction of the target. In this method, the symmetric matrix  $S$  is retrieved by combining gray-scale information with the associated transpose values. This makes it possible to estimate the cumulative relationship among the pixels in one direction. To achieve this, the symmetric association matrix  $S$  is first normalized according to Eq (1) to measure the probability matrix  $P_{i,j}$ .

$$P_{i,j} = \frac{S_{i,j}}{\sum_{i,j=0}^{N-1} S_{i,j}} \quad (1)$$

After the probability matrix  $P_{i,j}$  was extracted, different STTF features were obtained. The primary motive was to retain the diversity of features encompassing the energy, textural, and orientational feature components. Training a machine-learning model over such diversified and heterogeneous features can yield higher accuracy and reliability. Brief descriptions of these different STTF features are provided below.

#### 1) Contrast

In the GLCM STTF feature prospect, contrast is defined as the changes in gray-scale values over the input frames. Once the probability matrix in Eq (1) has been derived, the pixel pairs signifying the diagonal elements provide the difference in contrast values. In this approach, the texture contrast signifies the overall changes in local pixel intensities throughout the input frame(s). Typically, the non-linearity over the input frames is assessed using a certain statistical assessment and an allied textural continuity examination. In this study, Equation (2) was used to perform contrast estimation and associated feature retrieval.

$$CONT = \sum_{i,j=0}^{N-1} P_{i,j} (i - j)^2 \quad (2)$$

#### 2) Energy

The energy distribution throughout the input frame(s) was extracted. To achieve this, the angular second moment (ASM) signifying the rotational acceleration over the input feature space was estimated. Equation (3) was used to estimate the ASM value for each input frame. The ASM value increases linearly as the gray-level values increase over the input video frame. After the ASM value per input frame was estimated, the energy parameter was measured using Eq (4).

$$ASM = \sum_{i,j=0}^{N-1} P_{i,j}^2 \quad (3)$$

$$ENR = \sqrt{ASM_{i,j}} \quad (4)$$

#### 3) Homogeneity

Generally, in image processing, homogeneity is assessed in terms of the inverse different moment (IDM); hence, a higher IDM is interpreted as higher homogeneity. In other words, it is also observed in terms of contrast, where a lower contrast is assumed to result in higher homogeneity. Here, Equation (5) was applied to estimate the homogeneity distribution across the input video frame. Similar to the linear magnitude distribution across the input video frame, a smaller contrast often results in higher homogeneity.

$$HOM = \sum_{i,j=0}^{N-1} \frac{P_{i,j}}{1+(i-j)^2} \quad (5)$$

#### 4) Correlation

In the case of GLCM feature extraction, the correlation parameter refers to a descriptive statistic (feature) across the input frame. To ensure STTF diversity, three statistical (descriptive) features were obtained: mean, standard deviation, and correlation. To estimate these feature values, the probability matrix value  $P_{i,j}$  in Eq (1) obtained for each input video frame was used. Mathematically, Equations (6)–(9) were used to measure the mean and standard deviation values for the different pixels.

$$\mu_i = \sum_{i,j}^{N-1} i(P_{i,j}) \quad (6)$$

$$\mu_j = \sum_{i,j}^{N-1} j(P_{i,j}) \quad (7)$$

The mean values derived in Eqs (6) and (7) were used to estimate the standard deviation, as in Eq (9).

$$\sigma_i^2 = \sum_{i,j}^{N-1} P_{i,j} (i - \mu_i)^2 \quad (8)$$

$$\sigma_i = \sqrt{\sigma_i^2}$$

$$\sigma_j = \sqrt{\sigma_j^2} \quad (9)$$

Correlation information was obtained using the estimated values of the mean and variance, as in Eq (10).

$$CORR = \sum_{i,j}^{N-1} P_{i,j} \left[ \frac{(i-\mu_i)(j-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}} \right] \quad (10)$$

#### 5) Skewness

In addition to the textural features, different orientation-related features encompassing skewness and kurtosis were obtained. These features are referred to as “symmetrical statistical features.” Skewness is sometimes referred to as a “lack of symmetry.” Here, skewness is defined in the form of a shade feature, in which a higher cluster shade indicates the asymmetric nature of the feature space (pertaining to the specific input frame). In this study, Equation (11) was used to measure the skewness over each input video frame.

$$SKEW = \sum_{i,j}^{N-1} P_{i,j} (i - \mu_i + j - \mu_j)^4 \quad (11)$$

#### 6) Kurtosis

Kurtosis (Kurt) refers to the “peakedness” of the input gray-level values distributed across the input video frame(s). Generally, a higher kurtosis value indicates that the magnitude of the feature distribution is primarily strenuous toward the tail(s) compared with the mean value. Conversely, a lower kurtosis indicates that the feature distribution remains strenuous in the direction of the spike, which is closer to the mean value. In this work, the kurtosis was estimated over the complete input image because there is no specific target (feature) region, and the complete input video frame serves as an input textural feature.

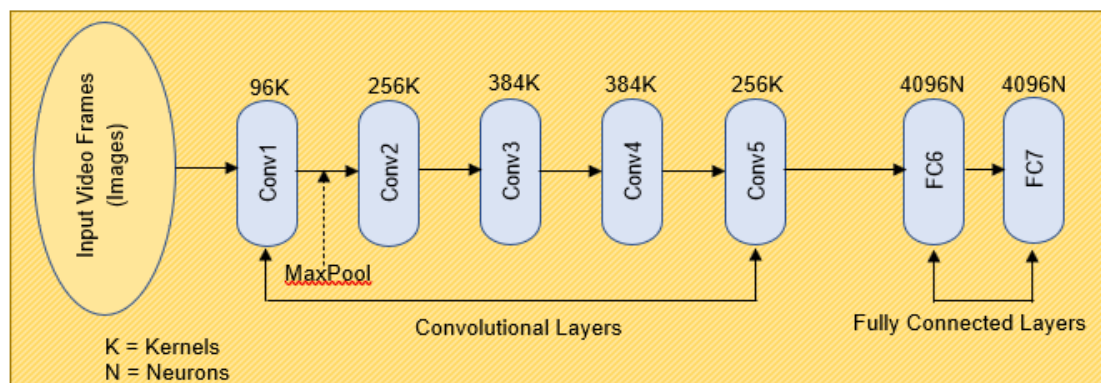
After the eight different STTF GLCM features were extracted, a horizontal concatenation-based feature fusion (Figure 2) was performed. Thus, the proposed STTF GLCM feature model results in Eq (12) as a composite STTF feature vector, which can be used for learning and classification.

$$GLCM_{Feat} = Conc(CONT, ENE, HOM, CORR, Mean, STD, Kurt, Skew) \quad (12)$$

### 4.3.2. AlexNet deep feature extraction

Based on the hypothesis that the strategic amalgamation of deep features with STTF GLCM features can yield superior feature vectors for learning and classification, in addition to the aforementioned GLCM features, deep features were extracted using AlexNet. Unlike the classical CNN, which applies 256 dimensional features, the AlexNet deep-learning model employs 4096-dimensional features at the fully connected layer (FC6 and FC7). Consequently, it can provide significant and in-depth intrinsic features for learning and classification. A detailed discussion of the AlexNet model for feature extraction is provided below.

The AlexNet deep-learning model, which is often included under the umbrella of a high-performance transferable deep-learning structure, provides high-dimensional features for learning and classification. Originally developed for target detection and classification problems, the AlexNet model has evolved significantly during the past few years. The capability of AlexNet to handle varied data and deep-learning structures makes it suitable for object detection and classification. To ensure the maximum possible feature efficacy and potential, the optimal network structure encompassing five convolutional layers (CONV1, CONV2, CONV3, CONV4, and CONV5) and three fully connected layers (FC6, FC7, and FC8) was retained. The overall design of the AlexNet deep-learning model is shown in Figure 3.



**Figure 3.** AlexNet deep-learning model.

A description of the different layers and associated characterization follows.

#### 1) Input Layer

In this study, each (target) video frame was input to the AlexNet input layer. A linear activation function was maintained at the input layer; hence, the input of the first convolutional layer (CONV1) was the same as the original input or preprocessed input frame. To ensure uniform feature extraction over the visible range, each input frame was resized to  $217 \times 217$ . Thus, the input of CONV1 can be stated to be of the  $217 \times 217$  dimension. Moreover, to exploit the maximum possible spatio-temporal features, the total number of channels was set to be three. Once each input frame was assigned as an input, AlexNet executed a CONV layer that applies two distinct filters (horizontal and vertical) to perform feature extraction.

#### 2) Convolutional Layers (CONV)



CONV is a combination of different filters designed to extract deep features or patterns from the input data (video frames). More specifically, it applies two filters at each CONV layer (horizontal and vertical filters) to extract features over each cross section. As depicted in Figure 3, CONV1 applies 96 kernels or neurons to extract the features. The neurons in the extracted feature map shared an equivalent set of weight ( $W$ ) and bias ( $b$ ) values. This enabled the neurons to detect a pattern(s) with similar features. In the proposed method, the AlexNet CONV layer filters the input video frame to estimate a single feature map, which signifies the output of the CONV layer. In this study, AlexNet was designed with five layers: CONV1, CONV2, CONV3, CONV4, and CONV5. Zero-padding and a stride of 2 were applied for feature extraction. The kernel configurations (i.e., neurons) at different CONV layers are shown in Figure 3. In the proposed model, to ensure timely computation and real-time decision-making for crowd detection and classification, a drop-out of 0.50 (i.e., 50%) was applied after the convolutional layer. Consequently, only significant feature sets were retained by dropping less significant feature elements. It not only helps in learning over the significant features but also reduces delay and the computational burden.

### 3) Max-Pooling Layer

To retain significant feature sets for learning, a dropout layer with a value of 0.5 was used; however, to guarantee data presentation in the optimal shape and size, a max-pooling layer was used before the fully connected layer. Max-pooling acts as a feature selection layer, where it is intended to minimize the spatial resolution of each feature map retrieved from the CONV layers. In addition, it reduces the computational cost by performing local averaging and subsampling. This also alleviates the issue of overfitting. A max-pooling layer was applied in its native structure, which retained the translation-invariant representation across the input video frames. This approach down samples the latent representation by applying a constant factor as the maximum value over a non-overlapping subspace. It retains sparsity rather than hidden representation by dropping all nonmaximal values throughout the non-overlapping subspace. Thus, the use of max-pooling avoids insignificant solutions and ensures that it does not permit a suitable solution to be carried forward. In the proposed AlexNet model, a max-pooling layer was applied after each CONV layer, where each layer was defined as a  $3 \times 3$  receptive field with a stride of 4.

### 4) ReLU Layer

In the deployed AlexNet deep model, a rectified linear unit (ReLU) was applied as an activation function. This layer contains a nonlinear element-wise operating function. In the proposed deep model, three ReLU layers were employed, in which, with input  $y$ , the output for the neuron  $q(y)$  is calculated as  $y$  if  $y > 0$  and  $(\delta \times y)$  if  $y \leq 0$ . Here,  $\delta$  signifies whether negative values must be ignored by performing a multiplication with a slope (Ex. 0.01...) or fixing it to 0. In this work, a value of  $\delta = 0$  was set. Consequently, the ReLU acts as a default activation or ReLU function  $q(y) = \max(0, y)$ .

### 5) Fully Connected (FC) Layer

In the deployed AlexNet layer, the FC layer performs high-level reasoning for feature learning and classification. Functionally, this layer receives a set of neurons (representing the feature vector) from the previous layers (i.e., the CONV layers) and maps them to connected neurons, thus generating a one-dimensional feature vector. In accordance with the AlexNet deep-learning model (Figure 3), the one-dimensional features were retained in the FC6 layer, which provided a 4096-dimensional feature vector for further feature learning and classification. Thus, the final feature vector obtained is  $AlexNet_{Feat}$ . After the visual features were extracted from GLCM ( $GLCM_{Feat}$ ) and AlexNet



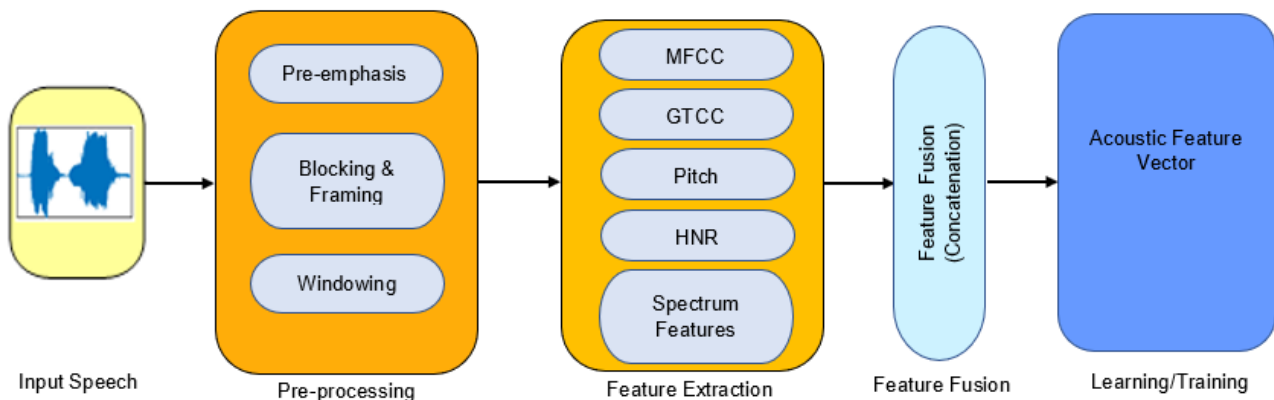
( $AlexNet_{Feat}$ ), they were concatenated to yield a visual feature vector for the input video data. In other words, the cumulative feature vector for the deep spatio-temporal Fourier transform (STFT) feature extraction model was retained, as shown in Eq (13).

$$Feat_{visual} = [GLCM_{Feat}; AlexNet_{Feat}] \quad (13)$$

After the deep STTF features were extracted from the video inputs, audio feature extraction was performed from the same video inputs. Details of the acoustic feature extraction are provided in Section 4.3.3.

#### 4.3.3. Acoustic feature extraction

Because of the multi-modal crowd analysis approach, in addition to visual features ( $Feat_{visual}$ ), acoustic features were also extracted. The majority of previous audio-based person identification or crowd analysis approaches have applied Mel-frequency Cepstral Coefficient (MFCC) acoustics to perform classification; however, these solutions failed to address the acoustic ambiguity primarily caused by overlapping sounds, mixed-gender sounds with different pitches, and allied temporal characteristics. Considering such a complex operating environment (i.e., crowd analysis under extreme conditions), especially with multiple types of crowd (singing, praying, cheering, quarreling, violence, etc.), merely applying the MFCC cannot yield a generalizable solution. Considering this, in this study, the goal was to extract (and retain) different acoustic features, including MFCC, GTCC (gammatone frequency cepstral coefficient (GFCC)), HNR, pitch, spectral centroid, spectral flux, spectral slope, and GTCC-DELTA. The main motive behind applying different acoustic features was to ensure high feature heterogeneity to guarantee improved learning and classification for crowd detection and classification. The proposed acoustic feature extraction model is shown in Figure 4.



**Figure 4.** Acoustic feature extraction.

##### 4.3.3.1. Audio pre-processing

For improved acoustic analysis, preprocessing should be performed. In this study, three subtasks were performed: pre-emphasis, blocking or framing, and windowing. The purpose of these preprocessing steps was to obtain a suitable set of audio samples with minimum disturbances and optimal spatio-temporal cues encompassing higher cohesion and continuity. Here, pre-emphasis and

windowing help segment the complete video sample. In other words, a continuous audio signal (or sample) is split into multiple blocks or smaller frames. In this study, the Hamming window method was used.

Instead of traditional continuous wavelet transform or Fourier transform methods, the Spatio-temporal Fourier Transform (STFT) method was used for acoustic feature extraction to ensure the maximum possible spatio-temporal cue retention. This approach converts the input spatial-domain data into frequency-domain values. As stated in the previous section, before extracting the MFCC, preprocessing was executed in the form of blocking or windowing without imposing aliasing effects or noise components. In this study, the Hann window method was used, which converts continuous input audio signals into multiple blocks or smaller audio frames encompassing  $N$  samples. An attempt was made to generate the frames in such a manner that consecutive frames remained distinguished or separated by  $M$  samples ( $M < N$ ). This helped reduce the likelihood of overlapping (e.g., adjacent frame overlapping) by  $N - M$  samples. In crowd analysis, in which voices can come from multiple sources speaking the same way or differently, ensuring overlap-free inputs can provide more-accurate acoustic cues for further learning and classification. When the frame size is smaller than the overlapping speech component, the information or data contained in the frame do not have sufficient intrinsic information to help with crowd identification or the associated classification. Therefore, the Hann window concept was used to convert the input audio stream into multiple blocks of frames. Subsequently, the Hamming window method was employed to minimize the disruptions at the start and end of each block or frame. To achieve this, the window function (here, Hann windowing) was multiplied by each block or frame. In the case of a window function defined as  $W_n(m)$ ,  $0 \leq m \leq N_m - 1$ , where  $N_m$  represents the sample quality within each retrieved frame, the output after windowing would be Eq (14):

$$Y(m) = X(m)W_n(m), 0 \leq m \leq N_m - 1 \quad (14)$$

In Eq (14),  $Y(m)$  states the windowed signal output. Equation (15) is used as the Hamming window function in Eq (14) to obtain a windowed signal.

$$W_n(m) = 0.54 - 0.46 \cos\left(\frac{2\pi m}{(N_m - 1)}\right), 0 \leq m \leq N_m - 1 \quad (15)$$

#### 1) Mel-Frequency Cepstral Coefficient

Human perception of sound frequency does not follow a linear scale. Therefore, in the case of a crowd audio sample(s), every tone encompassing the actual frequency  $f$  (Hz) signifies a subjective pitch that is often measured on a scale. This scale is called the ‘‘Mel scale’’ [79], which is defined as Eq (16).

$$f_{MEL} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (16)$$

In the above derived function,  $f_{MEL}$  signifies the subjective pitch (Mel) in the form of frequency (Hz). Consequently, the MFCC is defined as a baseline acoustic feature (set) for speech recognition [71]. In general, the MFCC coefficient signifies a set of discrete cosine transform (DCT) decorrelated parameters that are measured by applying a transformation of logarithmically compressed filter output energies [72]. In classical approaches, the MFCC is measured by applying a perceptually spaced triangular filter bank that applies a discrete fourier transform (DFT) to an input speech sample. An  $N$ -point DFT over the input speech signal  $y(n)$  is obtained using Eq (17).

$$Y(k) = \sum_{n=1}^M y(n) \cdot e^{\left(\frac{-j2\pi nk}{M}\right)} \quad (17)$$

In Eq (17), the condition  $1 \leq k \leq M$  is followed. A filter bank with linearly spaced filters is applied in Mel scale over the acoustic spectrum. Thus, the filter response  $\psi_i(k)$  of the  $i^{\text{th}}$  filter is obtained using Eq (18).

$$\psi_i(k) = \begin{cases} 0 & \text{for } k < k_{b_{i-1}} \\ \frac{k-k_{b_{i-1}}}{k_{b_i}-k_{b_{i-1}}} & \text{for } k_{b_{i-1}} \leq k \leq k_{b_i} \\ \frac{k_{b_{i+1}}-k}{k_{b_{i+1}}-k_{b_i}} & \text{for } k_{b_i} \leq k \leq k_{b_{i+1}} \\ 0 & \text{for } k \leq k_{b_{i+1}} \end{cases} \quad (18)$$

If  $Q$  is the total number of filters in the filter bank, the boundary points of each filter can be derived using Eq (19).

$$\{k_{b_i}\}_{i=0}^{Q+1} \quad (19)$$

Equation (19) signifies the boundary points of the filter, and  $k$  is the coefficient index. In general, the boundary points for each filter  $i$  ( $i = 1, 2, \dots, Q$ ) are measured as equidistant points on the Mel scale. In the case of the MFCC, the boundary points for each filter are obtained according to Eq (20).

$$K_{b_i} = \left(\frac{M}{f_s}\right) \cdot f_{mel}^{-1} \left[ f_{mel}(f_{mel}) + \frac{i\{f_{mel}(f_{high})-f_{mel}(f_{low})\}}{Q+1} \right] \quad (20)$$

In Eq (20),  $f_s$  is the sampling frequency (Hz),  $f_{low}$  and  $f_{high}$  are the low- and high-frequency boundaries of the filter bank, respectively, and  $f_{mel}^{-1}$  signifies the inverse of the transformation in Eq (16), which is mathematically obtained as Eq (21).

$$f_{mel}^{-1}(f_{mel}) = 700 \left[ 10^{\frac{f_{mel}-1}{2595}} - 1 \right] \quad (21)$$

In this manner, the energy outputs  $e(i)$  ( $i = 1, 2, \dots, Q$ ) pertaining to the Mel-scaled bandpass filters are measured as the addition of energy  $|Y(k)|^2$ , falling into a predefined Mel-frequency band, weighted by the corresponding frequency response  $\psi_i(k)$ .

$$e(i) = \sum_k |Y(k)|^2 \psi_i(k) \quad (22)$$

The STFT method is applied to the log filter bank energies  $\{\log[e(i)]\}_{i=1}^Q$  to decorrelate the energies. Thus, the MFCC coefficient  $C_m$  is obtained using Eq (23).

$$C_m = \sqrt{\frac{2}{N} \sum_{i=0}^{Q-1} \log[e(l+1)] \cdot \cos \left[ m \cdot \left( \frac{2l+1}{2} \right) \cdot \frac{\pi}{Q} \right]} \quad (23)$$

where  $m = 0, 1, 2, \dots, R-1$ , and  $R$  is the target number of MFCCs. In this study, the MFCC was considered as an acoustic feature for learning.

## 2) Gammatone Cepstral Coefficient (GTCC)

Typically, the GTCC is thought to have the potential to address the shortcomings of the MFCC or other Mel-frequency acoustic representations. Unlike the MFCC representations discussed above,

the GTCC spectrogram applies multiple asymmetric filters rather than triangular filters. The use of gammatone or asymmetric filters makes it possible to approximate the filters in a superior manner, particularly when basilar-membrane-driven filtering is applied. The final processing element of this approach is cubed-root compression, which is performed on the filter response outputs. The filter bandwidth, also called the “equivalent rectangular bandwidth” (ERB), is estimated according to Eq (24), and the central frequency  $f_c$  is measured using Eq (25).

$$ERB(f_c) = 24.7 * \left( \frac{4.37 * f_c}{1000} + 1 \right) \quad (24)$$

$$f_{c,i} = (f_H + 228.7) * \exp\left(\frac{v_i}{9.26}\right) - 228.7, 1 \ll i \ll 128 \quad (25)$$

where  $i$  is the applied filter, and  $f_H$  is the cutoff frequency of the  $i^{\text{th}}$  filter. In the GTCC, as many as 128 filters can be deployed to achieve personalized performance. The variable  $v_i$  in Eq (25) is the placement location or step at which the filter is applied or placed. In the proposed model, Eq (24) is used as the GTCC acoustic feature extraction model because it retains low computation with a sufficiently large feature space to perform learning and classification. A gammatone filter bank is applied over the Fourier transform of the input signal, emphasizing the vital audio signal frequencies or acoustic cues. A gammatone filter bank is applied with the filter order in Eq (20) and the corresponding ERB concept. Finally, a log function is employed, along with a discrete cosine transform, over the ERB to model the loudness perception of the crowd or video. Finally, decorrelation is performed over the outputs from the logarithmic compressed filter, providing superior acoustic energy information. By applying the above-stated GTCC function, the acoustic cues can be estimated for further learning and classification.

Two other variants of the GTCC model (GTCC-delta and GTCC-delta-delta) were used. Recalling that the GTCC belongs to a biologically inspired variant and by applying gammatone filters with ERB, different GTCC coefficients were obtained. To improve the intrinsic temporal information, the GTCC was improved using first- and second-order derivatives. Applying first- and second-order derivatives of the GTCC coefficients resulted in three different sets of acoustic features (GTCC, GTCC-delta, and GTCC-delta-delta).

### 3) Pitch

In addition to the above discussed MFCC and GTCC acoustic feature variants, the pitch information over the input audio samples was estimated. In acoustics, pitch is generally considered a measure of sound frequency (Hz), where a higher frequency results in a higher pitch value. “Pitch” is defined as the perceived magnitude pertaining to the “frequency of vibration.” This definition works as perceived because some people cannot recognize or identify a specific pitch value. The frequency at which the vocal cords start vibrating is a function of the cord shape, airflow, and elasticity or tension across the vocal tract. In summary, pitch is a quantifiable parameter that signifies the extent of the frequency perceived from the input audio signal. In this study, the pitch information was estimated from each audio sample or input. To extract these acoustic features “audioFeatureExtractor” a MATLAB (2020) function, was used.

### 4) Harmonics to Noise Ratio (HNR)

In general, the HNR is defined as the degree of acoustic periodicity, which is estimated in decibels. Mathematically, it is derived as the ratio of the energy of the periodic part to the noise energy. The HNR also measures the ratio between the periodic and nonperiodic components of speech sounds. As

the name indicates, the HNR estimates the association between harmonics and noise components. This makes it possible to identify the voiced component(s) in the speech signal; however, it requires quantification of the association between the periodic and nonperiodic components (dB). Because the sound signal, depending on pronunciation, speech pattern, or a specific kind of sound or voice, can have different harmonics, the HNR can be applied to crowd analysis. In other words, the harmonics embedded in singing or prayer are different from those pertaining to quarrels or violence. HNR estimation is related to the energy transformed by the voiced signal through glottal impulses and the energy of the glottic noise fraction after filtering through the vocal tract. Typically, noise occurs with disturbances caused when airflow passes through the glottis (especially during phonation). Thus, by quantifying these important parameters, the HNR can be estimated and used to detect and classify crowds. Although numerous approaches have been developed for HNR estimation (previously, both cepstrum and autocorrelation information have been used to estimate the harmonic and noise components), in this study, Equations (26) and (27) were applied.

$$X(w) = H(w) + N(w) \quad (26)$$

In Eq (26),  $X(w)$  represents the speech signal in the frequency domain, while  $H(w)$  and  $N(w)$  are harmonic and noise components, respectively. In general, the HNR is defined as the logarithmic measure of the energy ratio pertaining to the harmonic and noise components. Therefore, the model derived in Eq (28) can integrate the spectral power over the audible frequency range.

$$HNR = 10 \times \log_{10} \frac{\int_w |H(w)|^2}{\int_w |N(w)|^2} \quad (27)$$

Thus, once the acoustic features (i.e., MFCC, GTCC, GTCC-delta, GTCC-delta-delta, pitch, and HNR) are estimated, they can be concatenated to yield the acoustic or audio features  $Feat_{Audio}$ .

$$Feat_{Audio} = [MFCC, GTCC, GTCC - Delta, GTCC - Delta - Delta, Pitch, HNR] \quad (28)$$

Finally, the visual features  $Feat_{Visual}$  and the acoustic features  $Feat_{Audio}$  were concatenated to yield a composite feature vector for further learning and classification.

$$Feat_{Composite} = [Feat_{Visual}; Feat_{Audio}] \quad (29)$$

#### 4.4. Random forest ensemble learning

Random forest is one of the most successful ensemble-learning algorithms. It structurally encompasses multiple tree-based classifiers and behaves as an ensemble-learning model. In the proposed tree model, each tree provides a corresponding vote for the most probable class for each crowd type. When the number of training samples (crowd video samples) is  $N$ , a sample encompassing  $N$  cases is randomly selected from the original data. The selected samples are employed as a training set to form a new tree. When there are  $M$  input variables, the best split on these,  $m$ , is applied to split the node. The value of  $m$  is kept constant during forest development, which is also called the “growing phase.” Thus, each tree develops to the greatest extent. Unlike classical machine-learning methods, which use a large number of hyperparameters to be tuned during training, random forest requires the minimum number of parameters to be estimated, even with a larger number of decision trees involved in tree formation (forest growth) for classification. This feature makes it computationally more efficient than other state-of-the-art machine-learning methods. In addition, its reliability renders it suitable for real-time applications. A complete random forest algorithm can

eventually be defined as a combination of different tree structures, as shown in Eq (30).

$$\{h(x, \theta_k), k = 1, 2, \dots, i \dots\} \quad (30)$$

where  $h$  is the classifier function, and  $\{\theta_k\}$  means the random vector is distributed identically. Each tree contains a vote for the most probable class for a specific video query as the input  $x$ . The dimensionality of  $\theta$  primarily depends on its use in tree formation. The most important reason for the success of random forest is its ability to form each of the decision trees that constitute the forest. A bootstrapped subset of training samples was employed to train each tree throughout the constructed forest, which enabled almost 70% of the training data to be used, whereas the remaining dataset was considered out-of-bag samples that were later used to perform inner cross validation to examine the classification results and enhance them.

## 5. Results and discussion

The emphasis of this study was on exploiting audio and video features or cues to perform crowd analysis and classification. However, the majority of available solutions applied just video features, including spatio-temporal or energy-related information, to detect certain events in vision-based surveillance systems. Characterizing crowd types has not been explored in depth. Most solutions either use visual (spatio-temporal or energy) or acoustic features as standalone cues to perform event detection. However, the aforementioned features (visual and acoustic) can differ for different types of crowd. For instance, the features of prayers differ from those representing violence or quarrels. In this case, adopting a classical feature model does not yield a generalizable solution. Moreover, ensuring the scalability of a solution (i.e., a universal solution with the ability to detect different crowds or events accurately) is necessary for contemporary surveillance systems. Training a model with the maximum number of possible feature traits or cues can be of great significance in achieving this. In this study, a multi-modality-driven hybrid feature-learning concept was developed for crowd analysis and classification. Because it is a multi-modal system, audio and visual features are both considered when performing crowd analysis. Because the intrinsic features pertaining to audio inputs can be quite different from video features, applying the same feature extraction model is not viable. Therefore, the crowd analysis problem and feature extraction were decoupled into audio- and vision-based features. To extract vital features or cues, different feature extraction models were applied to these modalities. More specifically, spatio-temporal and deep features from video inputs were used, while different acoustic cues were obtained from audio samples (from the same video input) to perform learning and classification. Based on previous research [73] in which it was found that the amalgamation of deep features with the GLCM can yield superior performance for crowd analysis, the AlexNet deep model was used, along with the GLCM method, to extract deep spatio-temporal features from the video inputs. AlexNet was employed using five CONVs and three FC layers. The only objective was to retain the high-dimensional features at the FC layer to make learning more accurate. At the FC6 layer, 4096-dimensional features were retained, and AlexNet was deployed with an adaptive learning capability (ADAM) with a learning rate of 0.0001. To improve the computational efficacy, a dropout layer with a value of 0.5 was used. Moreover, native ReLU was retained, as well as a single max-pooling layer (single max-pooling post-CONV), which not only helped in reducing the spatial resolution but also helped in retaining significant features for learning and classification. By applying the AlexNet deep model discussed above, a set of visual features ( $AlexNet_{Feat}$ ) were obtained. In the case of GLCM-

driven STTFs, eight features were extracted: contrast, correlation, energy, homogeneity, mean, standard deviation, skewness, and kurtosis. The purpose was to retain the maximum possible number of diversified features (including orientational, energy, and textural features) to improve the overall accuracy and reliability. The aforementioned GLCM features were concatenated to yield  $GLCM_{Feat}$  features. To retrieve a composite feature vector from the video inputs, robust horizontal concatenation was performed; thus, the final video features were obtained using  $Feat_{visual}$ —see Eq (13).

To extract acoustic feature cues from the same video input, audio signals were first obtained from the input video. The audio data were saved in \*.mp3 format; however, they can also be saved in other formats, such as \*.wav and \*.AVI, provided the computing machine has the ability to read the specific format. The audio inputs were first processed for pre-emphasis, framing, and windowing. Hamming windowing (periodic implementation) was applied following the Nyquist criteria to alleviate any aliasing problem. Subsequently, different acoustic features were retrieved, including the MFCC (static), GTCC (static), GTCC-delta, GTCC-delta-delta, pitch, HNR, spectral entropy, spectral flux, and spectral slope. The GTCC is thought to be superior to MFCC. However, because almost all existing methods for audio-based crowd analysis have applied MFCC acoustic cues, MFCC was used as well. To extract the aforementioned acoustic features, a window size of 20 ms, a step size of 10 ms, and a sampling rate of 22,050 Hz were considered. The number of speakers was either infinite or unknown. The noise component of the input samples was white Gaussian noise. Thus, to obtain a set of audio features or acoustics, concatenation was performed to yield a composite feature vector,  $Feat_{Audio}$ . When the final composite feature vector was obtained, it was projected to the random forest (RF) ensemble-learning classifier, which performed multiclass classification, thus categorizing the input crowd video as cheering, praying, protesting, violence, etc. Crowd video samples from the UMN [74] and PETS09 [75] benchmark data sources were used. The model was also tested using normal YouTube video samples pertaining to the crowd categories mentioned above. The proposed model was developed using MATLAB 2020b. The simulation was performed over a central processing unit arranged in a Microsoft Windows operating system with an i5 Intel processing unit, 8-GB RAM, and a 3.6-GHz processor.

**Table 2.** Performance parameters.

Parameter	Mathematical Expression
Accuracy	$\frac{(TN + TP)}{(TN + FN + FP + TP)}$
Precision	$\frac{TP}{(TP + FP)}$
Sensitivity	$\frac{TN}{(TP + FN)}$
Specificity	$\frac{TN}{(FP + TN)}$
F-Score	$\frac{2 * TP}{(2 * TP + FP + FN)}$

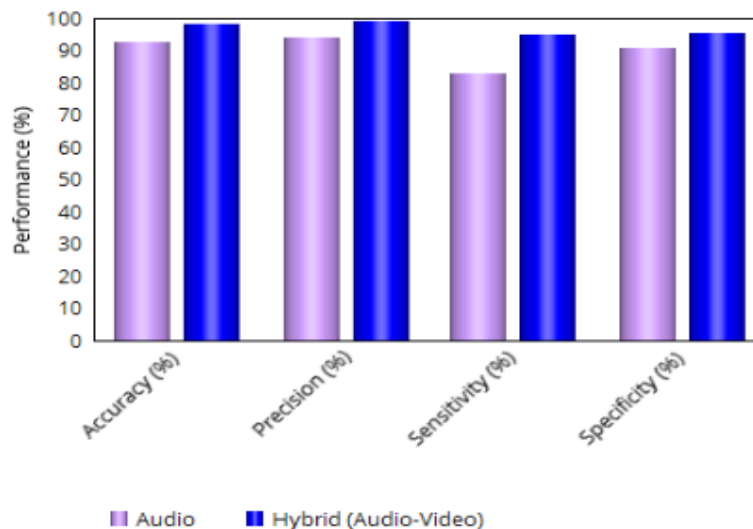
To assess the performance, a statistical performance analysis was conducted in terms of accuracy, precision, sensitivity, specificity, and F-score. To achieve this, confusion metrics were obtained in

terms of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The mathematical formulation employed for performance parameter derivation is presented in Table 2.

The overall assessment was performed using two broad approaches: intramodal assessment and intermodal assessment. An intramodal assessment was performed to examine the performance of the proposed crowd analysis and classification with audio and hybrid audio–video features, whereas an intermodal assessment was performed to examine the performance relative to other state-of-the-art methods. A detailed discussion of the results obtained, and inferences made follows.

### 5.1. Intramodel assessment

It was hypothesized that the inclusion of hybrid features or multi-modal systems encompassing audio and video features can yield superior and more-reliable solutions for crowd analysis. The efficacy of the proposed model was assessed using different feature sets. In other words, the performances of the proposed model on audio and hybrid features (audio and video) were examined separately. A random forest classifier was applied separately to the  $Feat_{Visual}$  and  $Feat_{Audio}$  features as the common classifier. The performance outputs were obtained in terms of accuracy, precision, sensitivity, specificity, and F-scores. The objective was to assess whether multi-modal (audio–video) features yield superior performance or whether audio-feature-driven methods can be convincing. The statistical performance outputs are listed in Table 3. A graphical depiction of the results is shown in Figure 5.



**Figure 5.** Intra-model assessment.

The results show that the proposed hybrid feature-driven multi-modal crowd analysis concept yields superior performance to the audio-driven crowd classification system (Table 3 and Figure 5). The simulation results reveal that the random forest ensemble-learning model with only acoustic or audio features yields a crowd classification accuracy of 92.67%, while the proposed multi-modal model (audio–video hybrid features) yields a superior accuracy of 98.26%. This clearly indicates the superiority of hybrid features over audio cues as standalone feature vectors for crowd analysis. The other performance parameters also confirm that the use of hybrid features (multi-modal audio–video



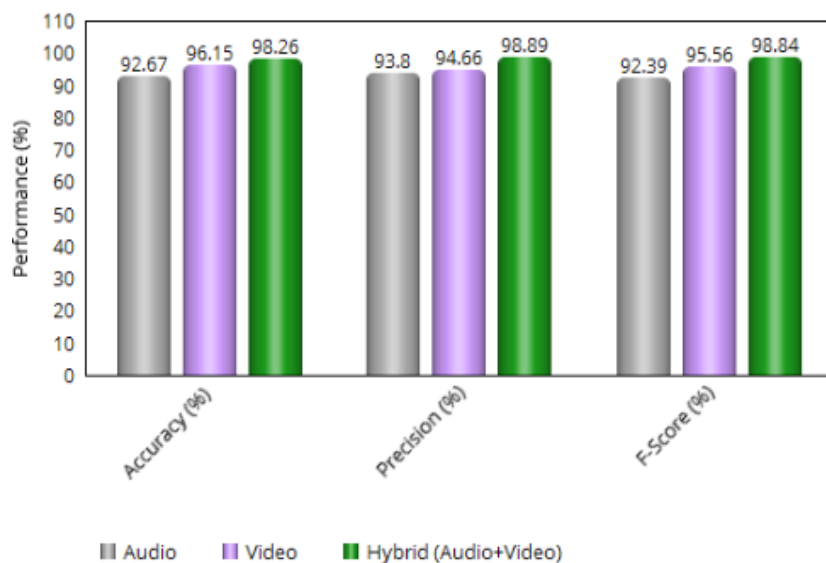
features) can be more efficient and reliable for crowd analysis tasks. The audio-driven approach yielded a precision of 93.80%. A sensitivity of 82.91% and specificity of 90.48% were obtained. In contrast, the proposed multi-modal or hybrid feature exhibited a precision of 98.89%, sensitivity of 94.82%, and specificity of 95.57%, which were significantly higher than those of the audio-driven crowd analysis concept. In the F-score analysis, which is often employed to assess the reliability of a system under different class imbalance conditions or real-world application scenarios, the audio-driven approach had an F-score or F-measure value of 0.9239, whereas the proposed hybrid or audio–video feature model yielded an F-score of 0.9884. This demonstrates the robustness of the proposed multi-modal crowd analysis system for real-time applications. The sensitivity of the audio-driven approach was 82.91%, whereas that of the proposed multi-modal-based crowd analysis system was 94.82%. This result shows that the proposed approach is almost 11.91% more sensitive, demonstrating the robustness of the hybrid feature in crowd analysis and classification, which can have a diversity of speech or audio–video differences as well as near similarity. Such high sensitivity makes the proposed system more efficient in real-world applications, where it makes swift and more-accurate crowd classification possible so that decisions can be made more quickly and effectively. Regarding the first research question (RQ1), it can be confirmed that the inclusion of hybrid features (audio and video) yields superior crowd analysis and classification compared with either video- or audio-based solutions.

The robustness of the proposed model in video-based crowd analysis was assessed further by referring to previous work [73], in which the GLCM with AlexNet (deep-spatio-temporal features) was used for crowd analysis. However, a heuristically driven neurocomputing concept was implemented as a classifier. More specifically, in the previous study, a bat-based algorithm multilayer feedforward neural network (BBA-MFNN) was used for crowd classification, where GLCM and AlexNet features were considered.

**Table 3.** Intra-model performance comparison.

Feature Set	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F-Score (%)
Audio	92.67	93.80	82.91	90.48	92.39
Hybrid (Audio-Video)	98.26	98.89	94.82	95.57	98.84

With numerous innovations in terms of computational optimization measures and feature engineering, the BBA-MFNN achieved an accuracy of 96.15%, a precision of 94.66%, and an F-measure of 95.56%. In comparison, the proposed method, which applied hybrid features (audio and video), exhibited an accuracy of 98.26%, a precision of 98.89%, and an F-score of 0.9884. Although the previous discussion already confirmed that the use of multi-modal features yields superior performance over an audio-driven crowd analysis system, a comparison with a video-based approach [73] indicates that the proposed model outperforms the video-based crowd analysis system by applying the same video features. The multi-modal concept yields 5.59% higher accuracy than the audio-based model and 2.11% higher accuracy than the deep-STTF GLCM-driven vision-based solution. This confirms that the use of multi-modal features (audio and video features together) yields superior performance compared with standalone feature-driven systems (either audio or video).



**Figure 6.** Performance with different feature combination.

A comparison of the proposed multi-modal crowd analysis system with the standalone feature-driven concept is shown in Figure 6. Considering the above conclusion that the proposed approach yields superior performance, it was compared with other state-of-the-art methods in terms of efficacy. This is the intermodal performance assessment or characterization. Details of the intermodal assessment are provided in the next section.

### 5.2. Intermodel assessment

For the intermodel assessment, the performance of the proposed multi-modal (audio–video feature-driven) crowd analysis method was compared with those of other state-of-the-art methods. Unlike for other vision-based classification problems or audio-driven personal recognition tasks, little effort has been devoted to crowd analysis. Moreover, research on multi-modal crowd concepts has been rare. Because this research focused on designing a hybrid feature-driven solution (unlike previous approaches that apply either audio or video), the model was compared with previous approaches by applying audio as well as video as a standalone solution. However, based on an in-depth literature survey, a few approaches were identified for crowd analysis and classification. The relative performances of the previous solutions and the proposed model are as follows.

Recently, a Gaussian mixture model (GMM) was used for feature extraction [76]. It was later processed using an SVM for crowd detection. Numerous real-time aspects were not addressed, such as crowd (type) heterogeneity, and the system was designed only for specific human movement patterns. However, the highest accuracy obtained (over the UMN dataset) was 85.53%, which was significantly lower than that of the proposed model, which exhibited an accuracy of 98.26%. Another approach was suggested [77], in which a substantial derivative concept was used to exploit fine-grained features or cues. However, its highest accuracy was 85.43%, which is less than that of the proposed model (98.26%). In another study [78], tracklet descriptive features were employed to perform crowd behavior analysis (a two-type classification problem for detecting normal and abnormal behaviors). This approach was not designed to perform multiple types of crowd detection; rather, it focused on

classifying crowd scenes behaviorally. The highest accuracy was 82.3%, which is significantly lower than that of the proposed hybrid feature-driven model (98.26%). In other studies [79,80], crowd analysis and classification were addressed; however, the highest accuracies were 81.3% and 81.5%, which are almost 17% lower than the results of the proposed multi-modal-driven crowd analysis and classification concept. This demonstrates the robustness of the proposed model over existing approaches. Here, the roles of the acoustic and visual features as composite feature indicators cannot be ruled out. Other researchers [81,30] applied different machine-learning methods to assess crowd analysis behavior; however, the highest accuracy obtained was 96%, which is still 2.2% lower than that of the proposed approach. As discussed in the previous section, the proposed multi-modal feature-driven crowd analysis model (98.26%) outperformed a previous vision-feature-based method (96.15%). In another study [69], the MFCC acoustic feature was employed to perform crowd analysis; however, the highest accuracy was 80.6%, which is significantly lower than that of the proposed solution. The proposed audio-feature-driven solution exhibited a crowd classification accuracy of 92.67 %, which is higher than that of the previous MFCC-based model [69]. This shows that the amalgamation of MFCC with other acoustics, such as GTCC and variants, pitch, HNR, and other spectrum cues, helped the proposed model achieve superior and more-reliable crowd analysis and classification results. A comparison of the results is presented in Table 4.

**Table 4.** Inter-model performance comparison.

Reference	Feature	Accuracy (%)
[76]	Visual	85.53
[77]	Visual	85.43
[78]	Visual	82.30
[79]	Visual	81.30
[80]	Visual	81.50
[81]	Visual	96.00
[69]	Audio	80.60
[73]	Visual	96.15
Proposed	Audio	92.67
	Hybrid (Audio-visual)	98.26

## 6. Conclusions

A robust multi-modality-driven crowd analysis and classification system was developed. To ensure high accuracy and reliability, the focus was on exploiting audio and visual features from input video sequences to perform crowd detection, analysis, and classification. Based on previous crowd analysis or event detection solutions, it was hypothesized that the strategic amalgamation of acoustic and visual features with information-rich deep spatio-temporal features can provide superior performance for crowd analysis and classification. The multi-modal feature concept was decoupled, and visual and acoustic features were obtained separately. Eight STTF GLCM features (contrast, correlation, energy, homogeneity, mean, standard deviation, skewness, and kurtosis) were obtained. This feature heterogeneity helped retain textural, orientational, and other statistical descriptive feature cues to achieve superior learning and classification. In addition to the GLCM features, the AlexNet deep-learning model was applied to the FC6 (4096-dimensional) feature set. The GLCM and AlexNet

features were fused to generate visual features. The use of these two concepts enriched the visual features to achieve high heterogeneity and diversity, helping the proposed model achieve superior classification efficacy. For audio inputs, different acoustic features were retrieved, including MFCC, GTCC, GTCC-delta, GTCC-delta-delta, pitch, HNR, and other spectrum-related features. This ensured that the acoustic cues retained an optimal set of features for further learning. Finally, the extracted visual and audio features were horizontally concatenated to provide a composite (multi-modal) feature vector for further learning and classification. The proposed multi-modal features were trained using a random forest ensemble classifier that classified input videos into different categories, including prayer, protest, violence, and cheering. A performance assessment revealed that the proposed hybrid feature-driven approach (i.e., audio-visual multi-modal system) achieved the highest crowd analysis and classification accuracy (98.26%), precision (98.89%), sensitivity (94.82%), specificity (95.57%), and F-score (0.9884). The performance of the approach was superior to the performances of audio-driven methods and visual-feature-based crowd analysis models. The information-rich features were the most important reason for the superior performance. Moreover, the diversity of features, including deep-STTF features, Nyquist-conditioned noise, and overlapping free acoustic features with different cues, helped the proposed model exhibit the best performance for crowd detection, analysis, and classification. To ensure the scalability of the model, it was trained on different crowd videos, including violence, prayer, cheering, and quarreling. The model was tested on different crowd datasets, and it exhibited superior performance compared with other crowd analysis and classification methods. The robustness and scalability of the model render it suitable for real-time applications.

## Acknowledgments

The authors thank the Management, the Principal, and the authorities of Malnad College of Engineering, Hassan, for extending their full support for this research. We would like to thank Editage ([www.editage.com](http://www.editage.com)) for English language editing.

## Conflict of interest

The authors declare there is no conflict of interest.

## References

1. J. C. S. Jacques Junior, S. R. Musse, C. R. Jung, Crowd analysis using computer vision techniques, *IEEE Signal Process. Mag.*, **27** (2010), 66–77. <https://doi.org/10.1109/MSP.2010.937394>
2. Z. Jie, D. Gao, D. Zhang, Moving vehicle detection for automatic traffic monitoring, vehicular technology, *IEEE Trans. Veh. Technol.*, **56** (2007), 51–59. <https://doi.org/10.1109/TVT.2006.883735>
3. H. Qiu, X. Liu, S. Rallapalli, A. J. Bency, K. Chan, R. Urgaonkar, et al., Kestrel: Video analytics for augmented multi-camera vehicle tracking, in *2018 IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IoTDI)*, IEEE, Orlando, FL, (2018), 48–59. <https://doi.org/10.1109/IoTDI.2018.00015>
4. R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2009), 935–942. <https://doi.org/10.1109/CVPR.2009.5206641>

5. D. Y. Chen, P. C. Huang, Motion-based unusual event detection in human crowds, *J. Visual Commun. Image Represent.*, **22** (2011), 178–186. <https://doi.org/10.1016/j.jvcir.2010.12.004>
6. C. C. Loy, T. Xiang, S. Gong, Detecting and discriminating behavioural anomalies, *Pattern Recognit.*, **44** (2011), 117–132. <https://doi.org/10.1016/j.patcog.2010.07.023>
7. P. C. Ribeiro, R. Audigier, Q. C. Pham, RIMOC, a feature to discriminate unstructured motions: Application to violence detection for video-surveillance, *Comput. Vision Image Understanding*, (2016), 1–23. <https://doi.org/10.1016/j.cviu.2015.11.001>
8. Y. Benabbas, N. Ihaddadene, C. Djeraba, Motion pattern extraction and event detection for automatic visual surveillance, *EURASIP J. Image Video Process.*, **2011** (2011), 1–7. <https://doi.org/10.1155/2011/163682>
9. B. Krausz, C. Bauckhage, Loveparade 2010: Automatic video analysis of a crowd disaster, *Comput. Vision Image Understanding*, **116** (2012), 307–319. <https://doi.org/10.1016/j.cviu.2011.08.006>
10. V. Kaltsa, A. Briassouli, I. Kompatsiaris, M. G. Strintzis, Timely, robust crowd event characterization, in *19<sup>th</sup> IEEE International Conference on Image Processing*, IEEE, (2012), 2697–2700. <https://doi.org/10.1109/ICIP.2012.6467455>
11. Y. Zhang, L. Qiny, H. Yao, P. Xu, Q. Huang, Beyond particle flow: Bag of trajectory graphs for dense crowd event recognition, in *IEEE International Conference on Image Processing*, ICIP, 2013. <https://doi.org/10.1109/ICIP.2013.6738737>
12. W. Choi, S. Savarese, A unified framework for multi-target tracking and collective activity recognition, in *Computer Vision-ECCV 2012*, (2012), 215–230. [https://doi.org/10.1007/978-3-642-33765-9\\_16](https://doi.org/10.1007/978-3-642-33765-9_16)
13. M. Hu, S. Ali, M. Shah, Learning motion patterns in crowded scenes using motion flow field, in *2008 19th International Conference on Pattern Recognition*, (2008), 1–5. <https://doi.org/10.1109/ICPR.2008.4761183>
14. S. Ali, M. Shah, A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis, in *2007 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2007. <https://doi.org/10.1109/CVPR.2007.382977>
15. C. C. Loy, T. Xiang, S. Gong, Modelling multi-object activity by gaussian processes, in *Proceedings of the British Machine Vision Conference, BMVC*, 2009.
16. A. S. Rao, J. Gubbi, S. Marusic, M. Palaniswami, Crowd event detection on optical flow manifolds, *IEEE Trans. Cybern.*, **46** (2016), 1524–1537. <https://doi.org/10.1109/TCYB.2015.2451136>
17. H. Mousavi, S. Mohammadi, A. Perina, R. Chellali, V. Murino, Analyzing tracklets for the detection of abnormal crowd behavior, in *IEEE Winter Conference on Applications of Computer Vision*, (2015), 148–155. <https://doi.org/10.1109/WACV.2015.27>
18. H. Fradi, J. L. Dugelay, Spatial and temporal variations of feature tracks for crowd behavior analysis, *J. Multimodal User Interface*, **10** (2016), 307–317. <https://doi.org/10.1007/s12193-015-0179-2>
19. Y. Zhao, Z. Li, X. Chen, Y. Chen, Mobile crowd sensing service platform for social security incidents in edge computing, in *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, (2018), 568–574. [https://doi.org/10.1109/Cybermatics\\_2018.2018.00118](https://doi.org/10.1109/Cybermatics_2018.2018.00118)
20. Y. Yuan, J. Fang, Q. Wang, Online anomaly detection in crowd scenes via structure analysis, *IEEE Trans. Cybern.*, **45** (2015), 548–561. <https://doi.org/10.1109/TCYB.2014.2330853>

21. S. Wu, B. E. Moore, M. Shah, Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes, in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (2010), 2054–2060. <https://doi.org/10.1109/CVPR.2010.5539882>
22. X. Cui, Q. Liu, M. Gao, D. N. Metaxas, Abnormal detection using interaction energy potentials, in *Conference on Computer Vision and Pattern Recognition, CVPR, 2011*, (2011), 3161–3167. <https://doi.org/10.1109/CVPR.2011.5995558>
23. S. Cho, H. Kang, Abnormal behavior detection using hybrid agents in crowded scenes, *Pattern Recognit. Lett.*, **44** (2014), 64–70. <https://doi.org/10.1016/j.patrec.2013.11.017>
24. E. de-la-Calle-Silos, I. González-Díaz, E. Díaz-de-María, Mid-level feature set for specific event and anomaly detection in crowded scenes, in *2013 IEEE International Conference on Image Processing*, Melbourne, (2013), 4001–4005. <https://doi.org/10.1109/ICIP.2013.6738824>
25. N. Ihaddadene, C. Djeraba, Real-time crowd motion analysis, in *2008 19th International Conference on Pattern Recognition*, (2008), 1–4. <https://doi.org/10.1109/ICPR.2008.4761041>
26. A. N. Shuaibu, A. S. Malik, I. Faye, Behavior representation in visual crowd scenes using space-time features, in *2016 6th International Conference on Intelligent and Advanced Systems (ICIAS)*, (2016), 1–6. <https://doi.org/10.1109/ICIAS.2016.7824073>
27. R. Hu, Q. Mo, Y. Xie, Y. Xu, J. Chen, Y. Yang, et al., AVMSN: An audio-visual two stream crowd counting framework under low-quality conditions, *IEEE Access*, **9** (2021), 80500–80510. <https://doi.org/10.1109/ACCESS.2021.3074797>
28. U. Sajid, X. Chen, H. Sajid, T. Kim, G. Wang, Audio-visual transformer based crowd counting, in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, (2021), 2249–2259. <https://doi.org/10.1109/ICCVW54120.2021.00254>
29. A. N. Shuaibu, A. S. Malik, I. Faye, Adaptive feature learning CNN for behavior recognition in crowd scene, in *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, Malaysia, (2017), 357–361. <https://doi.org/10.1109/ICSIPA.2017.8120636>
30. K. Zheng, W. Q. Yan, P. Nand, Video dynamics detection using deep neural networks, *IEEE Trans. Emerging Top. Comput. Intell.*, **2** (2018), 224–234. <https://doi.org/10.1109/TETCI.2017.2778716>
31. Y. Li, A deep spatio-temporal perspective for understanding crowd behavior, *IEEE Trans. Multimedia*, **20** (2018), 3289–3297. <https://doi.org/10.1109/TMM.2018.2834873>
32. L. F. Borja-Borja, M. Saval-Calvo, J. Azorin-Lopez, A short review of deep learning methods for understanding group and crowd activities, in *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, (2018), 1–8. <https://doi.org/10.1109/IJCNN.2018.8489692>
33. B. Mandal, J. Fajtl, V. Argyriou, D. Monekosso, P. Remagnino, Deep residual network with subclass discriminant analysis for crowd behavior recognition, in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Athens, (2018), 938–942. <https://doi.org/10.1109/ICIP.2018.8451190>
34. A. N. Shuaibu, A. S. Malik, I. Faye, Y. S. Ali, Pedestrian group attributes detection in crowded scenes, in *2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, Fez, (2017), 1–5. <https://doi.org/10.1109/ATSIP.2017.8075584>
35. F. Solera, S. Calderara, R. Cucchiara, Socially constrained structural learning for groups detection in crowd, *IEEE Trans. Pattern Anal. Mach. Intell.*, **38** (2016), 995–1008. <https://doi.org/10.1109/TPAMI.2015.2470658>

36. S. Kim, S. Kwak, B. C. Ko, Fast pedestrian detection in surveillance video based on soft target training of shallow random forest, *IEEE Access*, **7** (2019), 12415–12426. <https://doi.org/10.1109/ACCESS.2019.2892425>
37. S. Din, A. Paul, A. Ahmad, B. B. Gupta, S. Rho, Service orchestration of optimizing continuous features in industrial surveillance using Big Data based fog-enabled Internet of Things, *IEEE Access*, **6** (2018), 21582–21591. <https://doi.org/10.1109/ACCESS.2018.2800758>
38. M. U. K. Khan, H. Park, C. Kyung, Rejecting motion outliers for efficient crowd anomaly detection, *IEEE Trans. Inf. Forensics Secur.*, **14** (2018), 541–556. <https://doi.org/10.1109/TIFS.2018.2856189>
39. A. N. Shuaibu, I. Faye, Y. Salih Ali, N. Kamel, M. N. Saad, A. S. Malik, Sparse representation for crowd attributes recognition, *IEEE Access*, **5** (2017), 10422–10433. <https://doi.org/10.1109/ACCESS.2017.2708838>
40. C. Riachy, F. Khelifi, A. Bouridane, Video-based person re-identification using unsupervised tracklet matching, *IEEE Access*, **7** (2019), 20596–20606. <https://doi.org/10.1109/ACCESS.2019.2896779>
41. H. Fradi, J. Dugelay, Sparse feature tracking for crowd change detection and event recognition, in *2014 22nd International Conference on Pattern Recognition, ICPR*, (2014), 4116–4121. <https://doi.org/10.1109/ICPR.2014.705>
42. B. Yogameena, S. Saravana Perumal, N. Packiyaraj, P. Saravanan, Ma-Th algorithm for people count in a dense crowd and their behaviour classification, in *2012 International Conference on Machine Vision and Image Processing (MVIP)*, Taipei, (2012), 17–20. <https://doi.org/10.1109/MVIP.2012.6428750>
43. R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, (2009), 935–942. <https://doi.org/10.1109/ICIP.2012.6467453>
44. X. Zhou, L. Zhang, Abnormal event detection using recurrent neural network, in *2015 International Conference on Computer Science and Applications (CSA)*, Wuhan, (2015), 222–226. <https://doi.org/10.1109/CSA.2015.64>
45. S. S. Pathan, A. Al-Hamadi, B. Michaelis, Crowd behavior detection by statistical modeling of motion patterns, in *2010 International Conference of Soft Computing and Pattern Recognition*, Paris, (2010), 81–86. <https://doi.org/10.1109/SOCPAR.2010.5686403>
46. S. Wang, Z. Miao, Anomaly detection in crowd scene using historical information, in *2010 International Symposium on Intelligent Signal Processing and Communication Systems*, Chengdu, (2010), 1–4. <https://doi.org/10.1109/ISPACS.2010.5704770>
47. P. Wang, W. Hao, Z. Sun, S. Wang, E. Tan, L. Li, et al., Regional detection of traffic congestion using in a large-scale surveillance system via deep residual TrafficNet, *IEEE Access*, **6** (2018), 68910–68919. <https://doi.org/10.1109/ACCESS.2018.2879809>
48. T. P. Nguyen, C. C. Pham, S. V. Ha, J. W. Jeon, Change detection by training a triplet network for motion feature extraction, *IEEE Trans. Circuits Syst. Video Technol.*, **29** (2019), 433–446. <https://doi.org/10.1109/TCSVT.2018.2795657>
49. C. Riachy, F. Khelifi, A. Bouridane, Video-based person re-identification using unsupervised tracklet matching, *IEEE Access*, **7** (2019), 20596–20606. <https://doi.org/10.1109/ACCESS.2019.2896779>
50. G. Olague, D. E. Hernández, E. Clemente, M. Chan-Ley, Evolving head tracking routines with brain programming, *IEEE Access*, **6** (2018), 26254–26270. <https://doi.org/10.1109/ACCESS.2018.2831633>

51. Y. Lee, S. Chen, J. Hwang, Y. Hung, An ensemble of invariant features for person reidentification, *IEEE Trans. Circuits Syst. Video Technol.*, **27** (2017), 470–483. <https://doi.org/10.1109/TCSVT.2016.2637818>
52. S. Yi, H. Li, X. Wang, Pedestrian behavior modeling from stationary crowds with applications to intelligent surveillance, *IEEE Trans. Image Process.*, **25** (2016), 4354–4368. <https://doi.org/10.1109/TIP.2016.2590322>
53. Bera, S. Kim, D. Manocha, Interactive crowd-behavior learning for surveillance and training, *IEEE Comput. Graphics Appl.*, **36** (2016), 37–45. <https://doi.org/10.1109/MCG.2016.113>
54. H. Fradi, B. Luvison, Q. C. Pham, Crowd behavior analysis using local mid-level visual descriptors, *IEEE Trans. Circuits Syst. Video Technol.*, **27** (2017), 589–602. <https://doi.org/10.1109/TCSVT.2016.2615443>
55. M. Abdar, N. Y. Yen, Design of a universal user model for dynamic crowd preference sensing and decision-making behavior analysis, *IEEE Access*, **5** (2017), 24842–24852. <https://doi.org/10.1109/ACCESS.2017.2735242>
56. X. Zhang, Q. Yu, H. Yu, Physics inspired methods for crowd video surveillance and analysis: A survey, *IEEE Access*, **6** (2018), 66816–66830. <https://doi.org/10.1109/ACCESS.2018.2878733>
57. Y. Zhang, L. Qin, R. Ji, S. Zhao, Q. Huang, J. Luo, Exploring coherent motion patterns via structured trajectory learning for crowd mood modeling, *IEEE Trans. Circuits Syst. Video Technol.*, **27** (2017), 635–648. <https://doi.org/10.1109/TCSVT.2016.2593609>
58. W. Liu, R. W. H. Lau, X. Wang, D. Manocha, Exemplar-AMMs: Recognizing Crowd movements from pedestrian trajectories, *IEEE Trans. Multimedia*, **18** (2016), 2398–2406. <https://doi.org/10.1109/TMM.2016.2598091>
59. C. Chen, Y. Shao, X. Bi, Detection of anomalous crowd behavior based on the acceleration feature, *IEEE Sensors J.*, **15** (2015), 7252–7261. <https://doi.org/10.1109/JSEN.2015.2472960>
60. V. J. Kok, C. S. Chan, GrCS: Granular computing-based crowd segmentation, *IEEE Trans. Cybern.*, **47** (2017), 1157–11680. <https://doi.org/10.1109/TCYB.2016.2538765>
61. C. Chen, Y. Shao, Crowd escape behavior detection and localization based on divergent centers, *IEEE Sensors J.*, **15** (2015), 2431–2439. <https://doi.org/10.1109/JSEN.2014.2381260>
62. Y. Zhang, L. Qin, R. Ji, H. Yao, Q. Huang, Social attribute-aware force model: Exploiting richness of interaction for abnormal crowd detection, *IEEE Trans. Circuits Syst. Video Technol.*, **25** (2015), 1231–1245. <https://doi.org/10.1109/TCSVT.2014.2355711>
63. H. Yao, A. Cavallaro, T. Bouwmans, Z. Zhang, Guest Editorial Introduction to the Special Issue on group and crowd behavior analysis for intelligent multicamera video surveillance, *IEEE Trans. Circuits Syst. Video Technol.*, **27** (2017), 405–408. <https://doi.org/10.1109/TCSVT.2017.2669658>
64. A. S. Keçeli, A. Kaya, Violent activity detection with transfer learning method, *Electron. Lett.*, **53** (2017), 1047–1048. <https://doi.org/10.1049/el.2017.0970>
65. V. Kaltsa, A. Briassouli, I. Kompatsiaris, L. J. Hadjileontiadis, M. G. Strintzis, Swarm intelligence for detecting interesting events in crowded environments, *IEEE Trans. Image Process.*, **24** (2015), 2153–2166. <https://doi.org/10.1109/TIP.2015.2409559>
66. V. H. Roldão Reis, S. J. F. Guimarães, Z. K. Gonçalves do Patrocínio, Dense crowd counting with capsule networks, in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, (2020), 267–272. <https://doi.org/10.1109/IWSSIP48289.2020.9145163>
67. P. Foggia, A. Saggese, N. Strisciuglio, M. Vento, N. Petkov, Car crashes detection by audio analysis in crowded roads, in *2015 12<sup>th</sup> IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, (2015), 1–6. <https://doi.org/10.1109/AVSS.2015.7301731>



68. S. Duan, X. Wang, X. Yu, Crowded abnormal detection based on mixture of kernel dynamic texture, in *2014 International Conference on Audio, Language and Image Processing*, (2014), 931–936. <https://doi.org/10.1109/ICALIP.2014.7009931>
69. Ç. Okuyucu, M. Sert, A. Yazici, Audio feature and classifier analysis for efficient recognition of environmental sounds, in *2013 IEEE International Symposium on Multimedia*, (2013), 125–132. <https://doi.org/10.1109/ISM.2013.29>
70. M. Baillie, J. M. Jose, An audio-based sports video segmentation and event detection algorithm, in *2004 Conference on Computer Vision and Pattern Recognition Workshop*, (2004), 110. <https://doi.org/10.1109/CVPR.2004.298>
71. M. A. Hossan, S. Memon, M. A. Gregory, A novel approach for MFCC feature extraction, in *2010 4th International Conference on Signal Processing and Communication Systems*, (2011), 1–5. <https://doi.org/10.1109/ICSPCS.2010.5709752>
72. T. D. Ganchev, *Speaker Recognition*, PhD thesis, University of Patras, 2005.
73. H. Y. Swathi, G. Shivakumar, Evolutionary computing assisted neural network for crowd behaviour classification, *Neuroquantology*, **20** (2022), 2848–2855. <https://doi.org/10.14704/nq.2022.20.9.NQ44331>
74. <http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi>
75. PETS09 dataset, Available from: <http://www.cvg.reading.ac.uk/PETS2009/a.html>.
76. M. Marsden, K. McGuinness, S. Little, N. E. O'Connor, Holistic features for real-time crowd behaviour anomaly detection, in *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, (2016), 918–922. <https://doi.org/10.1109/ICIP.2016.7532491>
77. S. Mohammadi, A. Perina, V. Murino, Violence detection in crowded scenes using substantial derivative, in *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2015. <https://doi.org/10.1109/AVSS.2015.7301787>
78. H. Mousavi, S. Mohammadi, A. Perina, R. Chellali, V. Murino, Analyzing tracklets for the detection of abnormal crowd behavior, in *2015 IEEE Winter Conference on Applications of Computer Vision*, IEEE, (2015), 148–155. <https://doi.org/10.1109/WACV.2015.27>
79. Y. Itcher, T. Hassner, O. Kliper-Gross, Violent flows: Real-time detection of violent crowd behavior, in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012. <https://doi.org/10.1109/CVPRW.2012.6239348>
80. H. Mousavi, M. Nabi, H. Kiani, A. Perina, V. Murino, Crowd motion monitoring using tracklet-based commotion measure, in *2015 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2015. <https://doi.org/10.1109/ICIP.2015.7351223>
81. L. Lu, J. He, Z. Xu, Y. Xu, C. Zhang, J. Wang, et al., Crowd behavior understanding through SIOF feature analysis, in *2017 23rd International Conference on Automation and Computing (ICAC)*, Huddersfield, (2017), 1–6. <https://doi.org/10.23919/ICAC.2017.8082086>



AIMS Press

©2023 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)