Abnormal Behaviour Detection in Massive Crowd

Dr. Hanumanthappa S[1], Yashaswini N B[2], Ullas M P[3], Gunashree H K[4], Vismaya K R[5], Dr. Rajashekar K J[6]

[1,2,3,4,5,6]*Department of Information Science and Engineering, Kalpataru Institute of Technology,Tiptur*

doi.org/10.64643/IJIRTV12I7-188195-459

Abstract—The requirement for intelligent surveillance systems that can identify anomalous activity in intricate, high-density environments is highlighted by research on crowd anomaly detection. According to previous research, detection reliability is greatly increased when spatial characteristics, motion analysis, and temporal modelling are combined using techniques including optical flow, CNN-based feature extraction, trajectory analysis, and transformer frameworks. These studies demonstrate the superiority of deep learning over conventional feature-engineered methods, particularly hybrid spatiotemporal architectures. In order to identify anomalous activity in large crowds, this work proposes a hybrid approach that combines DenseNet-201 for spatial representation, optical flow for motion interpretation, and Bi-Directional LSTM for temporal behavior modeling. The model performs best with a batch size of 32 and a learning rate of 0.0003, achieving 55% accuracy and an AUC of 0.7187 when evaluated using the UCF-Crime dataset. The outcomes show how successful the technology is in monitoring crowds in the actual world.

Index Terms—AUC, autoencoder, crowd, CNN, Dense Net, nonparametric test, transformer, VGGNet, and video anomaly detection

I. INTRODUCTION

Researchers are paying more attention to the difficult problem of identifying human activity in movies. Human action detection in videos has applications in intelligent scene modelling, video annotation and retrieval, and surveillance systems. Despite its disadvantages and privacy concerns, video monitoring is now essential to fostering a sense of security, safety, and trust Abnormal behavior detection, which can be viewed as a subset of human action identification, is essential to ensuring both indoor and outdoor safety in video surveillance. Human oversight becomes challenging and ineffectual when there are few incidents. As a result, the need for automated techniques to recognize anomalous behaviour is growing. The challenging task of recognizing human activity in films is receiving increased attention from researchers.

Intelligent scene modelling, video annotation and retrieval, and surveillance systems can all benefit from the recognition of human actions in videos. Video monitoring is now crucial to building a sense of safety, security, and trust despite its drawbacks and privacy issues. In video surveillance, abnormal behaviour detection—which can be thought of as a subset of human action identification—is crucial for guaranteeing both indoor and outdoor safety. When there are few instances, human oversight becomes difficult and ineffectual. Automated methods to identify abnormal behaviour are therefore becoming more and more important.

II. LITERATURE SURVEY

Crowd behavior analysis utilizing a variety of machine learning algorithms has been the subject of numerous studies. To find efficient methods for identifying aberrant behavior in various crowd circumstances, researchers have investigated various datasets, algorithms, and procedures. These studies also present observed findings and suggest future lines of inquiry to enhance the effectiveness and precision of crowd anomaly detection systems. This section gives an overview of current technologies and the key methods used in the field of crowd behavior analysis, which has been an active study area for many years.

In their assessment of new deep-learning techniques for crowd anomaly detection, Jiao et al. [1] looked at feature extraction techniques, datasets, and model architectures that address issues including occlusion, scene complexity, and sparse anomaly data. They came to the conclusion that while deep-learning models perform better than conventional methods, real-world problems like shifting lighting and rare identified abnormalities still lower accuracy. The authors stressed that in order to create more reliable

and broadly applicable anomaly detection systems, future advancements would depend on improved temporal modeling, multi-modal fusion, and self-supervised learning.

A description of crowd anomaly identification and estimate methods was given by Hussain [3], who concentrated on the use of mobility patterns, density analysis, and spatiotemporal signals to spot anomalous crowd activity. The study outlined the advantages and disadvantages of both modern deep-learning models and conventional handcrafted methods. The author came to the conclusion that even while contemporary deep-learning frameworks greatly increase detection accuracy, problems like complicated crowd dynamics, occlusions, and changing ambient variables still exist. The study stressed that for accurate real-world crowd analysis, future advancements will need more dependable motion modeling, better density estimation, and the application of hybrid deep architectures.

Sharif [6] combined reconstruction networks with prediction models to learn typical motion patterns in order to identify crowd irregularities. According to the study's findings, this hybrid approach increases the accuracy of anomaly identification, but performance is still impacted by issues like intricate crowd motions and actual noise, necessitating more robust temporal modeling for increased dependability.

The goal of Yoon et al. [10] was to use adaptive model pooling to create an online deep anomaly detection system for dynamic data streams. They came to the conclusion that while this method enhances real-time identification in dynamic contexts, it still has issues with noise and quick data shifts, necessitating more robust adaptive mechanisms for reliable performance.

III.METHODOLOGY

1.Overview of Methodology:

Because aberrant events are rare, occlusions are frequent, and crowd patterns are unpredictable, it is difficult to identify anomalous crowd behaviour in real-world CCTV footage. As a result, the system employs a hybrid deep learning approach that combines: CNN (DenseNet-201) → for spatial feature learning,

Learning motion patterns with Optical Flow →To comprehend crowd size and congestion, use Crowd Density

Bi-directional LSTM → for temporal sequence learning Swin Transformer® to record violence levels and worldwide spatiotemporal correlation. Accurate detection of normal/abnormal behaviour across various scenarios, crowd sizes, and video sources is ensured by this multi-stage process.

2. Video Acquisition and Frame Extraction:

2.1Video Sources (Comprehensive Details):

The first and most crucial step in developing an aberrant crowd behaviour detection system is gathering video data from several real-world sources. The system must be able to handle a variety of video sources, both offline and live, since the objective is to automatically identify abnormal or aggressive activity.

1. RTSP streams from CCTV cameras:

The system accepts data from CCTV surveillance cameras that are frequently placed in public areas, including: Streets, Mall shopping, Stations for trains, Airports, Stadiums. RTSP (Real-Time Streaming Protocol) is used by the majority of CCTV systems to stream video.

By integrating with RTSP streams, the suggested system can record: Constant real-time video, The most recent frames available,Streams with high or low FPS,Video captured by several cameras at once This is crucial because the system must always be prepared to process live video because unexpected events like violence, panic, or theft frequently happen without warning.How to Manage RTSP The system receives RTSP streams and extracts the following using the NVIDIA DeepStream SDK:The last 20 frames,Real-time buffering Only a small delay (around five seconds, depending on network and settings).This guarantees that the system can function almost instantly, which qualifies it for use in security and surveillance applications.

2. Videos Offline (MP4 Files):

The technology supports both live streaming and pre-recorded video files, which are usually saved in formats like:MP4,AVI,MKV,MOV.These offline videos could include: Video of previous crimes, Videos for training, Content on social media, Incident recordings that were manually gathered, Typical applications of offline video processing include: Machine learning model training, Algorithm testing, Assessing the accuracy of the system, Making

datasets,MP4 videos are captured by Deep Stream at their original frame rate, which means: The processing speed of a 30 FPS video is 30 frames per second. The processing speed of a 60 FPS video is 60 frames per second. This keeps motion and temporal analysis synchronized while preserving the initial time of occurrences.

2.2 Frame Extraction:

Frame extraction is crucial because deep learning models work on image sequences, treating each frame as an independent sample for spatial analysis, motion computation, and temporal modeling. For instance, a 30 FPS video creates 30 images per second. Each frame is treated like an image. From this single frame, CNNs can extract information such as: Positions of people Facial expressions Objects involved (sticks, fire, weapons) Scene structure Abnormal postures This helps classify visual patterns associated with suspicious or violent behavior.

3.Data Preprocessing:

Since raw video frames from CCTV cameras, RTSP streams, and MP4 files cannot be directly used for deep learning, data preparation is a crucial step in an anomalous behavior detection system. These unprocessed frames are converted into structured, model-ready inputs via the preprocessing pipeline, enabling precise spatial, motion, and temporal analysis. To make sure the system records every visual shift in the scene, the first step is to extract individual frames from the video at its natural frame rate. Each video frame is transformed to RGB in order to meet the input requirements of CNN models like Dense Net and ResNet because OpenCV reads video frames in BGR format. This conversion is crucial because the model's comprehension of colors and textures is disrupted by an incorrect channel order. Following color conversion, frames are scaled to a set dimension (usually 224 × 224 × 3 for CNNs) so that all inputs, regardless of the initial video resolution, retain the same shape. This lowers computing costs and helps prevent shape mismatches. Next, by scaling pixel values to a specified range—typically zero mean and unit variance—normalization is used to stabilize training.

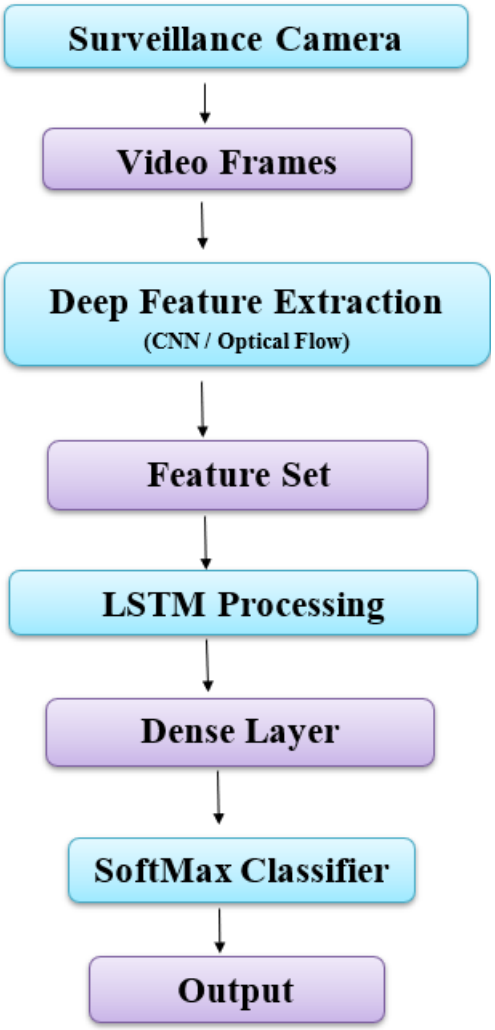

Figure1: Flow Process of Abnormal Behaviour Detection in Massive Crowd

4.Feature Extraction:

A crucial step in identifying anomalous behavior is feature extraction, which transforms video frames into useful numerical representations. Three categories of features are employed by the system: motion, crowd density, and location. DenseNet-201 extracts spatial characteristics by capturing high-level visual details like human poses, aggressive movements, fire or smoke, and suspicious objects through dense layer connections. The fundamental appearance-based comprehension of every frame is formed by these spatial encodings. Optical flow, which gauges pixel movement between successive frames, is used to obtain motion features.

Different motion patterns are produced by irregular movements, such as sprinting, fighting, or chaotic

crowd movement, and these patterns are readily seen in optical flow maps. These maps draw attention to unusual motion direction and intensity. Density maps are used to create crowd density features, which provide an estimate of the population and degree of traffic in various locations. This aids in categorizing behaviors according to crowd sizes: Fighting, Large Peaceful Gathering, Large Violent Gathering, and Normal. When combined, these characteristics offer a comprehensive grasp of population dynamics, mobility, and scene appearance.

5. Bi-Directional LSTM for Temporal Modeling:

To comprehend how human behavior changes over time, temporal modeling is crucial. The project captures temporal dependencies in both forward and backward directions using a Bi-directional Long Short-Term Memory (Bi-LSTM) network. The Bi-LSTM improves learning by concurrently examining the video timeline in reverse, in contrast to a normal LSTM that simply processes sequences from past to future. The model is extremely sensitive to temporal transitions because of its dual-direction processing, which is essential for spotting anomalous activity that changes across several frames. For instance, body motions like stepping forward, lifting arms, and abrupt acceleration are part of the sequence that precedes a fight. These changing cues are identified by the Bi-LSTM, which separates them from typical motions like walking and standing. The Bi-LSTM learns how these patterns evolve over time by successively receiving the spatial data that DenseNet-201 extracted for each frame. The system may identify activities that could seem normal in isolated frames but turn abnormal when evaluated as a sequence thanks to this temporal information. Because the Bi-LSTM can acquire long-range dependencies because to its memory cells, it can be used to identify abuse, which is characterized by repeated forceful gestures, or arson, which is characterized by a fire that steadily intensifies.

6. Tracking and Identifying Objects:

The system's capacity to locate people and examine their activity is improved by the addition of object detection and tracking. To detect and track individuals or important items in the video, the project makes use of resources like OpenCV, Haar Cascade classifiers, and DLIB. Haar Cascade is used to identify faces or upper bodies, making it possible to locate people effectively even in crowded situations. It enables real-time human detection by capturing edges and textures using basic rectangle filters. In order to guarantee that an object's movement is tracked across frames after it has been recognized, DLIB further allows tracking through correlation filters. When it comes to spotting behavioral abnormalities like chasing, abrupt running, unpredictable mobility, or people leaving or entering questionable locations, tracking is crucial.

The system learns more about people's activity habits by keeping track of their ongoing trajectories. Tracking, for instance, makes it possible to identify people approaching fire sources in arson scenarios or to detect recurrent aggressive gestures in abuse cases. By verifying that identified anomaly patterns are part of a persistent target rather than random noise, object tracking also lowers false positives.

7. CNN and LSTM Integration for Spatial-Temporal Learning:

The system's hybrid deep learning architecture, which combines CNN-based spatial analysis with LSTM-based temporal modeling, is its main strength. Each frame's intricate spatial properties, such as forms, textures, edges, postures, and contextual aspects, are captured by DenseNet-201. In the meantime, the system can identify dynamic anomalies since the Bi-LSTM learns how these visual patterns change across a video series. The system is guaranteed to comprehend both "what is happening in a frame" and "how it changes over time" thanks to this combination. For the purpose of detecting complicated aberrant actions that cannot be identified by spatial clues alone, such spatio-temporal integration is crucial.

A static posture, for instance, can seem normal at first, but if it is followed by abrupt movement, it could be a sign of conflict or assault.

Similarly, the existence of smoke or fire changes gradually, necessitating time comprehension. In order to give the LSTM a complete representation of both appearance and motion, the system combines Dense Net feature vectors with optical flow data. The detection capabilities is much improved by this fusion, which makes it possible to accurately identify abnormalities in difficult real-world surveillance scenarios such occlusions, congested areas, and erratic lighting.

8. UCF-Crime Dataset Training:

The UCF-Crime dataset, a sizable surveillance dataset with a variety of real-world anomalous behaviors like abuse, arrest, arson, assault, and fighting, is used to train the model. UCF-Crime is perfect for training reliable anomaly detection algorithms because, in contrast to limited curated datasets, it contains lengthy untrimmed movies with a variety of camera angles, lighting conditions, and crowd densities. The dataset ensures that the model learns to generalize across environments by covering a wide range of real-life circumstances. To ensure targeted, excellent training, the initiative uses just five categories.

Before being input into the DenseNet-BiLSTM pipeline, each frame of the videos is pre-processed and divided into training and testing sets. The diversity of scenes in the dataset—from street surveillance to indoor surveillance—ensures that the system is adaptable to changes in context. The model can distinguish between small behavioral cues like raised hands or abrupt movement patterns by incorporating both violent and non-violent activities. Training on this dataset guarantees that the model can accurately identify anomalous behavior in the actual world, making it dependable for real-world surveillance applications.

9.Hyperparameter tuning and model optimization:

The system is subjected to intensive hyperparameter optimization in order to obtain optimal performance. To find the optimal configuration, parameters like batch size, learning rate, and number of epochs are methodically changed. Experiments demonstrate how batch size influences the model's capacity for generalization, while varying learning rates have an impact on training stability and convergence. While higher batch sizes increase stability but need more memory, smaller batch sizes offer more frequent updates but may cause noise. The selection procedure is guided by confusion matrices, accuracy charts, and loss curves. To guarantee effective gradient propagation over the deep DenseNet-BiLSTM network, optimizers such as Adam or SGD are tested. Regularization methods and dropout layers aid in preventing overfitting.

In all anomalous categories, the final tuned model shows great performance, decreased loss, and increased accuracy. The model's practical reliability for real-time surveillance applications is ensured by hyperparameter adjustment.

10. Final Categorization and Production of Output:

Classifying the observed behavior as either normal or falling into one of the five aberrant activity categories is the system's final step. The output is sent to fully connected layers, which calculate the likelihood of each activity class, following the Bi-LSTM's processing of the temporal sequence. The final class label is determined by a SoftMax layer.

In real-world deployment circumstances, the system can initiate warnings and indicate problematic behavior. Visualization modules support human monitoring staff by highlighting important frames or motion patterns. Additionally, the system assesses performance by measuring correctness across classes using measures like accuracy and confusion matrices. Preprocessing, optical flow, Dense Net spatial learning, and LSTM temporal modeling all work together to produce significant results during this classification stage.

IV. RESULTS

The UCF-Crime dataset, which includes five anomaly classes—Arrest, Abuse, Assault, Arson, and Fighting—was used to assess the effectiveness of the suggested aberrant behavior detection method.

To find the best setup, several batch sizes and learning rates were used in each trial. In addition to visual aids such model accuracy plots, loss curves, confusion matrices, and AUC graphs, the main evaluation measures employed were Accuracy, Loss, Top-k Categorical Accuracy, and AUC (Area Under the ROC Curve).

The effectiveness of the DenseNet-201 and Bi-LSTM architectures in learning spatiotemporal patterns from video frames was assessed using these criteria. Three batch sizes—16, 32, and 64—were investigated in the initial round of tests. Batch sizes 16 and 32 produced relatively lower values, whereas batch size 64 obtained the maximum overall AUC value of 0.6875 and a consistent accuracy of 0.50. In particular, batch sizes 16 and 32 attained accuracy levels of 0.4250 and 0.3500, respectively. This showed that the number of samples that passed each training step had an impact on the model's performance. Larger batch sizes did not always increase final accuracy, but they did assist stabilize learning. The learning rate was adjusted for

each batch size in the subsequent experiment. The highest accuracy (0.6094) for batch size 16 was attained with a learning rate of 0.0003, albeit at a higher loss value. Learning rate 0.0003 once more yielded the best results for batch size 32, with an accuracy of 0.55 and the highest AUC score of 0.7187 in the entire trial. The best-performing model was found to have a batch size of 32 and a learning rate of 0.0003.

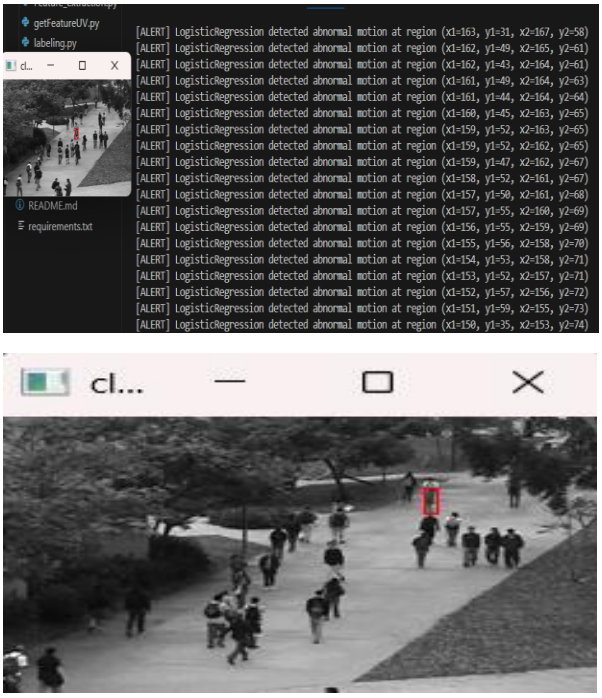

Figure2 : Results of Abnormal behaviour Detection

This combination allowed the network to converge more effectively and more correctly capture temporal correlations across LSTM layers. In a similar vein, batch size 64 outperformed batch size 32 with a learning rate of 0.0001, 50% accuracy, and an AUC value of 0.6875. A comparison with an existing method, a fully convolutional deep autoencoder, which claimed an accuracy of 50.6%, was also included in the results. With a higher accuracy of 55%, the suggested DenseNet-201 in conjunction with LSTM fared better than this earlier approach. This demonstrated the benefit of incorporating both temporal and spatial modeling for anomaly identification in videos.

Overall, the findings show that the suggested system has a good temporal detection capability and can recognize complicated aggressive or deviant behaviors in video streams with a considerable degree of accuracy. Even though the system's performance was impacted by issues like unequal class distribution and inconsistent lighting, it nevertheless produced findings that were competitive with those of conventional models. The most important discovery is that DenseNet-201 with Bi-LSTM, trained with batch size 32 and learning rate 0.0003, consistently produced the highest AUC and best accuracy, confirming this architecture as the ideal setup for the project.

V. CONCLUSION

Using DenseNet-201 for spatial analysis and Bi-Directional LSTM for temporal comprehension, the study effectively demonstrates a deep learning-based system for aberrant behavior identification in large crowds. Through the integration of optical flow-based motion extraction, spatiotemporal feature fusion, and frame-level preprocessing, the system can recognize complex violent behaviors in real-world surveillance footage, including fighting, assault, abuse, arrest, and arson. The model was trained and assessed on realistic, untrimmed, and difficult video scenarios thanks to the UCF-Crime dataset, which made it possible for the system to generalize across various settings, lighting conditions, and crowd densities.

The efficacy of the suggested strategy is confirmed by the experimental findings. The configuration with batch size 32 with a learning rate of 0.0003 proved to be the best-performing model after thorough testing of several batch sizes and learning rates; it achieved the greatest AUC value of 0.7187 and a maximum accuracy of 55%. These findings demonstrate that including both spatial and temporal modeling greatly improves anomaly detection performance, outperforming previously employed baseline models like the fully convolutional autoencoder. The system offers a solid basis for useful surveillance applications, even though its accuracy shows potential for improvement, particularly with regard to class imbalance and low-resolution surveillance footage.

Overall, the experiment shows that anomalous crowd behaviors may be successfully detected in real-time surveillance systems using deep learning in conjunction with motion analysis and temporal modeling. The results provide useful information for future improvements, including adding more sophisticated temporal architectures, growing the dataset, or implementing the model in real-time settings. As a result, the project makes a significant

contribution to the expanding field of intelligent video surveillance and creates a strong foundation for future study and advancement.

REFERENCES

[1] Jiao, L., Sharif, M. H., and Omlin, C. W. (2025). Deep crowd anomaly detection: current state, difficulties, and potential avenues for further study. A review of artificial intelligence.

[2] Alqahtani, H., Alruwais, N., Mansouri, W., Alohali, M. A., and Alshammeri, M. (2025). Enhanced crowd density monitoring for intelligent urban planning in smart cities using deep convolutional neural networks. Nature, 15. Scientific Reports

[3] A. Hussain (2024). An overview of crowd anomaly detection and estimation. Pattern Recognition Letters. ScienceDirect

[4] Patel, N., and Joshi, K. (2025). supervised deep learning techniques for identifying and detecting anomalies in crowd scenes. ELCVIA, 24.ecvia.cvc.uab.cat

[5] Nasir, R., Jalil, Z., Nasir, M., Ashraf, M.,Alsubait, T., & Saleem, S. (2025). Neural Computing & Applications: An improved framework for YOLOv8-based real-time anomalous behavior identification in dense crowds.

[6] M. H. Sharif (2023). Reconstruction and prediction networks are combined to detect deep crowd anomalies. Electronics,12(7).MDPI.

[7] Wu, Y. (2025). Convolutional neural networks are used in group behavior identification for abnormality converging scene analysis. Electrical engineering and computers. ScienceDirect

[8] M. J. Asif (2025). Crowd counting and anomaly detection are two deep learning approaches used in crowd scene analysis.arXivpreprint.

[9] Fatima, M., Kyung, C.-M., and Khan, M. U. (2020). Use expectation maximization filtering for plug-and-play anomaly detection. Ar Xivpreprint.arXiv

[10] Yoon, S., Lee, Y., Lee, J.-G., and Lee, B. S. (2022). Online deep anomaly detection from a complicated, dynamic data stream using adaptive model pooling. arXiv preprint. arXiv