SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V11I10P104

Volume 11 Issue 10, 30-46, October 2024

Original Article

Crowd Emotion and Behavior Analysis Using Lightweight CNN Model

Jignesh Vaniya[1], Safvan Vahora[2], Uttam Chauhan[3], Sudhir Vegad[4]

[1]*Gujarat Technological University, Ahmedabad, Gujarat, India.*
[2]*Information Technology Department, Government Engineering College, Modasa, Gujarat, India.*
[3]*Computer Engineering Department, Vishwakarma Government Engineering College, Gujarat, India.*
[4]*Department of Information Technology, Madhuben & Bhanubhai Patel Institute of Technology, Gujarat, India.*

[1]*Corresponding Author : jignesh.apit@gmail.com*

Received: 03 August 2024 Revised: 03 September 2024 Accepted: 04 October 2024 Published: 30 October 2024

Abstract *- Crowd behavior is a critical aspect of numerous applications such as crowd management, urban planning, and safety monitoring in the current era of the world. Convolutional Neural Networks (CNNs), one of the most recent advancements in deep learning, have demonstrated potential in the analysis of crowd behavior patterns. However, computational limitations frequently make it difficult to implement complex CNN models for crowd analysis tasks, particularly in real-time applications. The utilization of a lightweight CNN model for crowd behavior analysis on the Motion Emotion Dataset (MED) is proposed in our study. The MED dataset has diverse scenes with varying crowd emotional and behavioral aspects, making it an ideal benchmark for evaluating crowd analysis algorithms. The 2D CNN model is applied to the MED datasets to extract the features and annotations for training the lightweight CNN. The model is validated in the validation set and achieved an accuracy of 99.4% on the Emotion Dataset and 94.35% on the Behavior Dataset. The results are validated using the confusion matrix. The results indicate that the lightweight CNN model achieves competitive performance on the MED dataset while exhibiting reduced computational overhead compared to more complex models. The discoveries made aid in the advancement of effective and scalable strategies for crowd surveillance and control, with applications spanning across diverse sectors such as public safety, transportation, and event coordination.*

Keywords *- Crowd anomaly, Crowd behavior, CNN, Crowd emotional and behavioral analysis, Crowd Surveillance.*

1. Introduction

A crowd represents a gathering of individuals in a specific location, with variations depending on the circumstances. For instance, the composition of a crowd in a temple differs from that in a shopping area. The usage of the term ‘crowd’ is context-dependent, reflecting factors such as size, duration, composition, motivation, cohesion, and proximity of individuals within the group. Examining these aspects of crowd behavior is vital to proactively manage any potentially critical situations before they escalate [1].

In contemporary times, global overpopulation is giving rise to numerous crowded scenarios in many cities. These crowded situations emerge during events such as parades, entrances and exits of stations, political demonstrations, and strikes. Such situations pose an increase in security challenges [2]. Concurrently, an escalating number of cities are adopting surveillance systems utilizing video-protection cameras [3]. Initially, these surveillance systems were overseen by human agents. However, this approach proved to be ineffective, error-prone, and overwhelming over time [4]. Recent literature on incidents like mob lynching [5], protests against the CAA bill [6], the revocation of Article 370 [7], violence at multiple Indian universities [8-10], and Farmers breaching Delhi’s Red Fort in a huge tractor rally during which protest turn violence [12] underscores the relevance and necessity of automated crowd behavior analysis. Implementing crowd inspection systems enables the detection of abnormalities in public environments, serving as a valuable tool for various stakeholders in managing significant security threats to society [11].

The analysis of crowded environments has experienced a surge in popularity in recent years, driven by both academic research endeavors and the integration of advanced Artificial Intelligence (AI) technologies in various industries. The surge in popularity is primarily fueled by the escalating population growth rate and the imperative for more sophisticated and precise public monitoring systems. These systems have demonstrated their effectiveness in capturing crowd dynamics for designing public environments [13], simulating crowd behavior for game design [15], analyzing group activity [14], and monitoring crowds for visual surveillance [16]. A human expert

observer seems to be capable of monitoring the scene for unusual events in real-time and taking immediate reactions [14] accordingly. Nevertheless, psychological research demonstrates that humans have a significantly limited capacity to monitor multiple signals simultaneously [17]. In a situation like an extremely crowded scene, with multiple individuals doing different behaviors, monitoring poses a significant challenge even for a human observer. Since the last decade people have started to think about the atomization of crowd behavior. Significant research work started in early 2010 with the use of image processing. In addition, the computer vision approach added a milestone to the entire Video Anomaly Detection (VAD) System [17].

The primary challenge in this research domain revolves around uncertainty, particularly concerning the definition of a crowd and the determination of the quantity of people gathered in a given location. The context in which a crowd is defined remains ambiguous. For example, a group of thirty people gathered in a closed meeting room is the crowd, but the same number of people gathered in the park is not. The degree of uncertainty will be different in each case [18]. In such a system, designing a general-purpose model for the analysis of crowd behavior is difficult. This makes it difficult to perform an exhaustive review of published works because much research addresses different problems and is therefore not directly compared to one another. Previous studies have been concluded on the understanding of human and crowd behavior within the crowded scene [12]. As per past research studies, Crowd Statistics and behavior analysis are two main categories in the research scope. By using techniques for crowd counts, crowd statistics aim to estimate the density of the crowd. The purpose of crowd behavior analysis is to study the behavior of a crowd. Crowd behavior analysis is further subdivided into crowd tracking and activity analysis more towards behavior analysis, from which the area of interest is directed more towards the crowd emotion and behavior analysis. Current research trends are in keen interest to identify the video event in either anomaly or not only. The crowd can also be categorized based on various emotions like anger, happiness, neutrality, etc. The crowd behavior can be categorized as congestion, fighting, panic, etc.

Crowd Classification provided by Grant and Flynn in 2017 is shown below, in which counting and behavior analysis aspects are focused to identify the type of crowd. The approach to crowd counting remains consistent regardless of the scenario requiring density identification. Whether anomalies are anticipated or not, the methodology for crowd counting remains the same.

As depicted in Figure 2, Tripathi et al. proposed that assessing crowds can involve examining crowd counting, density estimation, scene analysis, and abnormality analysis.

From the previous studies and other sources that define crowd classification, a few points should be taken into consideration. Crowd counting, person tracking, density estimation, motion, etc., are the ways to identify the crowd behavior/crowd analysis in mainly two classes, which are Normal Crowds and Abnormal Crowd.

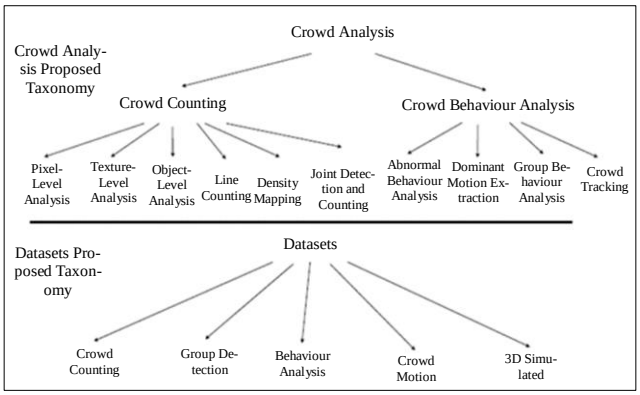

Fig. 1 Crowd classification [19]

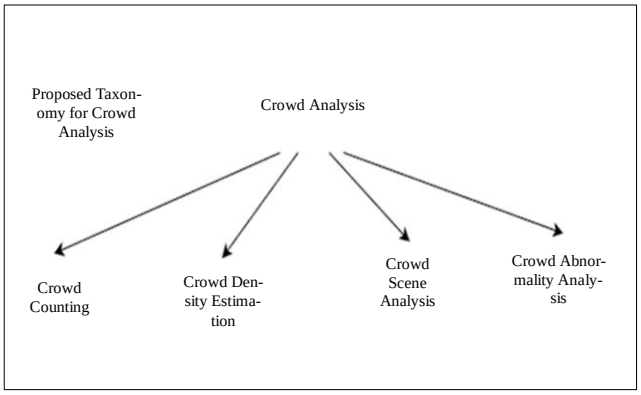

Fig. 2 Crowd taxonomy by Tripathi et al. [46]

Based on prior research and various sources that define crowd classification, several factors should be considered. These include crowd counting, person tracking, density estimation, motion analysis, etc. These methods help identify crowd behavior and analyze crowds primarily into two classes: Normal Crowd and Abnormal Crowds. In research work a novel approach and scenario aimed at discerning crowd emotions and behavior to forecast the present state of the crowd. Aligning with recent research trends, it is imperative to categorize the situation into multiple classes to bolster efficiency and accuracy. This categorization also proves beneficial for real-time prediction of crowd behavior and crowd management.

Using human emotions to depict crowd movements can aid in the comprehension of crowd behavior. To put it another way, to develop a thorough grasp of crowd behavior, human emotions can be used to bridge the semantic divide between

high-level behavior semantics and low-level motion information. A novel dataset for Crowd Emotion and Behavior has been proposed by Rabiee et al. [16].

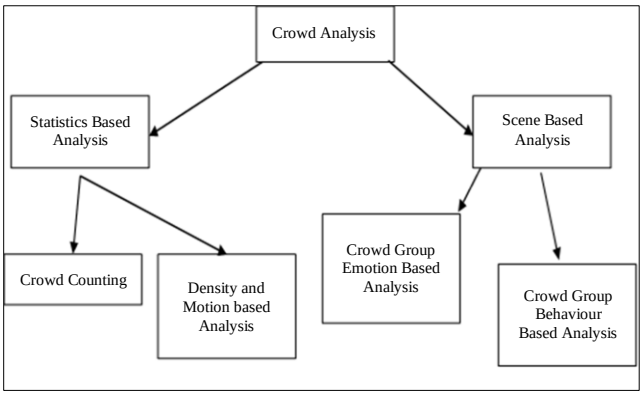

Fig. 3 Proposed taxonomy for crowd behavior analysis

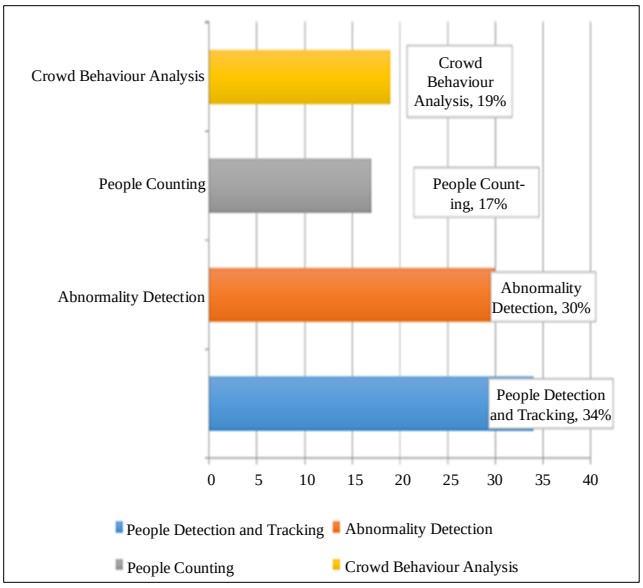

Fig. 4 Crowd analysis according to [23]

Analyzing crowd behavior presents a complex and burgeoning research domain that encompasses fields such as computer vision, soft computing, pattern recognition, machine learning, and deep learning. These fields converge to tackle various challenges, such as video surveillance, intelligent environments, crowd management, and the design of public spaces [11]. An intelligent environment can help in diverting crowds to ensure public safety, while public space design offers guidelines to assist planners in optimizing space usage.

Zitouni et al. [23] conducted a comparative analysis revealing that the number of publications on the globalized approach surpasses those on the localized-level approach by almost 50%. Density-based methods for modelling crowds are also widely favored [23].

In this paper, our aim is to identify the need for emotional and behavior analysis of the crowd by using Machine Learning. Due to the limitations of the traditional approaches, Machine Learning may play an important role. Various datasets can be used as supervised learning to identify the current state of the crowd.

In this paper, our aim is to identify the need for emotional and behavior analysis of the crowd by using Machine Learning. Due to the limitations of the traditional approaches, Machine Learning may play an important role. Various datasets can be used as supervised learning to identify the current state of the crowd. For better justification, try to classify the crowd into 7 various emotion classes and 6 different behavior classes. The remainder of the paper is structured as follows: Section 2 provides a review of previous studies concerning crowd analysis, crowd abnormal behavior analysis, and crowd anomaly detection. Section 3 discusses Machine Learning and CNN and their relevance to this study, including various methods. (Methods study) In Section 4, existing datasets are examined and compared- section 5, preprocessing of the dataset and applying the proposed methodology/architecture. Section 6 presents the prposed methodology and implementation. Section 7 evaluates the proposed solution through the confusion matrix and charts. Finally, Section 8 presents the conclusion and outlines future avenues of research.

2. Literature Reviews on the Crowd Analysis, Crowd Behavior Analysis and Crowd Anomaly Detection

The Crowd behavior analysis models involved steps like Crowd modeling, crowd counting, tracking, density estimation, behavior prediction, etc. In developing the crowd analysis model, visual sensors such as CCTV cameras are utilized to gather crowd-related data, while non-visual sensors like smartphones serve as sensing devices. The fusion of this information provides a more comprehensive understanding of crowd dynamics and behavior [21]. The subfields of AI began to gain traction in the early 2000s. Supervised learning and unsupervised learning emerged as pivotal methods for real-time prediction in response to the expanding size of datasets. From the book - concept, and real-time application of deep learning show the use of deep learning architecture with applications in natural language processing, semantic knowledge, forecasting and many more [22].

Zitouni MS et al. [23] published a review paper, “Crowd Analysis: A Survey”, in 2008. Comprehensive overview of crowd analysis methods involved in computer vision research. It emphasizes the importance of crowd analysis in various applications such as crowd management, public space design, virtual environments, visual surveillance, and intelligent environments. The authors discuss crowd models and event inference, detailing approaches like physics-inspired models, agent-based models, cellular automaton models, and nature-

based models. Additionally, the document focuses on the bridging of research in computer vision and non-vision approaches to crowd analysis, highlighting the potential of simulations and test frameworks for training and validating computer vision systems.

Ven Jyn Kok et al. [48] published a paper in 2016 that said that merging physics with biology reveals insights into the complex dynamics of collective human behavior. It explores emergent properties and self-organization within crowds, akin to phase transitions observed in physical systems. Drawing parallels with biological phenomena like swarming insects, it underscores the complexity and nonlinearity inherent in crowd dynamics. The review emphasizes interdisciplinary collaboration's practical implications, from urban planning to disaster management, in developing effective crowd control strategies. Through this synthesis, it offers a comprehensive understanding of crowd behavior, with potential applications across various real-world contexts.

Afiq et al.[49] discussed the Discussed Spatio-Temporal Technique (STT), Hidden Markov Model (HMM), Gaussian Mixture Model (GMM) and Optical Flow method to identify the abnormal behavior. They have also reviewed that CNN-based methods have been extensively adopted in recent research due to their strong hold on Object Detection and Classification.

J Zhang et al. [24] present the Collaboratively Self-supervised Video Representation (CSVR) framework aimed at action recognition. The method consists of three branches: generative pose prediction, discriminative context matching, and collaborative video generating, with extensive experiments showing state-of-the-art performance on UCF101 and HMDB51 datasets. The study demonstrates the significance of each branch, highlighting the complementary nature of static context and dynamic motion features for action recognition by the selection of dynamic motion feature, contrastive loss, video generation target, and hyperparameter investigation of σp, σc. The results show the superiority of the CSVR framework, achieving a top-1 accuracy of 90.3% and 56.5% on UCF101 and HMDB51, respectively. The CSVR framework is trained in a self-supervised manner, leveraging large amounts of unlabeled video data.

Swathi et al. [25] proposed a novel approach that combines statistical features from the Gray-Level Co-occurrence Matrix (GLCM) with deep learning features extracted using the AlexNet Convolutional Neural Network (CNN). This hybrid feature set is then used to train a Multi-feed Forward Neural Network (MFNN) model for multi-class classification of crowd behaviors. The proposed system model includes data acquisition, data preprocessing, hybrid statistical-deep feature extraction, transferable deep-learning AlexNet feature extraction, and multi-class classification using the MFNN algorithm. The researchers emphasize the significance of exploiting diverse spatio-temporal features and high-dimensional deep features to enable accurate and efficient crowd video analysis and classification. The authors compared it with other classification methods like Gaussian Mixture Model (GMM) and Support Vector Machine (SVM). The proposed model outperforms these methods in terms of accuracy, achieving 91.35%.

Ali Mollahosseini et al. [28] proposed AffectNet- A Database for Facial Expression, Valence, and Arousal Computing in the Wild. The paper addresses the limitations of existing datasets for emotion recognition. Around one million images are identified and labeled by the author. The annotation process is used to label facial expressions according to the seven basic emotions defined by Ekman, as well as dimensional valence and arousal ratings. The authors highlight the utility of AffectNet for advancing research in facial expression recognition, affective computing, and related fields. They demonstrate the effectiveness of the dataset through experiments on emotion classification tasks, achieving state-of-the-art performance compared to previous benchmarks.

Guodong Li et al. [26] proposed a model which is designed to recognize crowd behavior states and utilize an Intervention Optimization-Genetic Algorithm (IO-GA) to maximize the benefits-to-costs ratio, improving the efficiency of interventions. The model is demonstrated through simulation experiments on the Motion Emotion Dataset (MED) to analyze the changes in behavioral state, mood, and intensity of violent fighting events before and after an intervention. The comparison of original video frames, non-intervention simulations, and intervention simulations at different time frames (0, 15, 30, 45, 60, and 75) showed the impact of the intervention on crowd behavior and emotion.

Minzhong Wu at. al. [27] proposed approach focuses on recognizing the collective emotions of a crowd, which can be divided into two main categories: the recognition of individual emotions in the crowd and the recognition of group emotions of the crowd. The main categories of crowd emotions recognized in the proposed approach include anger, sadness, excitement, fear, happiness, and neutrality. The proposed approach, CS-RNN, has been compared with traditional deep learning methods, including the RNN model, CNN model, and CNN model with attention mechanism, in terms of accuracy and Kappa coefficient. The CS-RNN approach achieved an accuracy of 94.68%, which is higher than the accuracy of the other traditional deep learning methods. The proposed approach has limitations in accurately distinguishing between specific emotions, particularly in differentiating between happy and excited emotions. This challenge arises due to the reliance on motion features, which can lead to confusion in recognizing these specific emotions. Wang et al. [29] proposed a method that involves a two-step process, where a novel descriptor called multi-frame optical flow orientation (MHOFO) is first computed to capture movement information. This descriptor

is then used as input for a Cascade Deep Autoencoder (CDA) network designed to extract features of consecutive frames.

The CDA network is trained using normal samples, and abnormal samples are detected based on the reconstruction error in the testing phase. The experiments are performed on PETS2009, UMN, and UCSD datasets to evaluate the performance. On the PETS2009 dataset, the algorithm achieved high accuracy in detecting abnormal events, such as sudden changes in crowd behavior, with an area under the ROC curve (AUC) of 0.9501. The proposed methods efficiently identify the difference between walking and running in the crowd with an efficiency of 0.9752. The proposed model was validated on the UMN and UCSD datasets with considerable efficiency.

Rendón-Segador et al. [30] proposed CrimeNet model, which combines Vision Transformer (ViT) and Neural Structured Learning (NSL) with adversarial training, outperforms previous works by a significant margin, reducing false positives to practically zero and improving the state-of-the-art in violence detection on challenging datasets. CrimeNet achieves near-perfect results, with an accuracy of over 99.98% on the UCFCrime and XDViolence datasets. In the case of the NTUCCTV Fights dataset, CrimeNet achieves 100% accuracy. For future research, acknowledge the importance of cross-dataset experiments to evaluate the generalization of violence detection models.

Biao Guo et al. [31] proposed a novel method that combines spatial and temporal information using a two-stream spatial-temporal auto-encoder network with adversarial training. Additionally, the authors introduce a pseudo-abnormal dataset to address the lack of abnormal samples, and they conduct experiments on benchmark datasets to demonstrate the effectiveness of their proposed method. The quantitative assessment demonstrates that our approach outperforms current state-of-the-art methods in both frame-level and pixel-level evaluations, achieving a frame-level accuracy of 91.5% in Ped1 and 97.9% in Ped2. Additionally, our pixel-level evaluation yields scores of 82.7% on Ped1 and 95.1% on Ped2.

Manu Yadakere Murthygowda et al. [32] introduce the Integrated Multi-level Feature Fusion (IMFF) framework employs a multi-level feature fusion strategy to capture valuable insights from crowd behavior. It incorporates three tiers of feature fusion: the initial level focuses on physical attributes, the subsequent level emphasizes spatial relationships, and the final level delves into temporal characteristics. The major goal of the proposed method is to distinguish between normal and abnormal crowds. The performance evaluation of the IMFF framework using the UMN and VF datasets demonstrates its superior accuracy compared to existing methodologies, with an accuracy of 99.56%. Other aspects worth considering for future analysis include examining individual personnel mobility patterns to understand crowd behavior more comprehensively.

Ristea et al. [33] propose a lightweight masked Auto-Encoder (AE) for efficient video anomaly detection. The authors introduce an innovative approach to weight tokens based on motion gradients, enabling the model to focus on reconstructing tokens with higher motion and avoid learning to reconstruct the static background scene. A teacher and student decoder are introduced to enhance anomaly detection. The study highlights the crucial role of self-distillation, synthetic anomaly augmentation, and motion-based weights in boosting the model's accuracy. The model's efficiency is highlighted by its ability to process at unprecedented speeds, making it between 8 and 70 times faster than competing methods. Overall, the proposed lightweight masked model offers a remarkable trade-off between speed and accuracy, positioning it as a competitive approach in the field of video anomaly detection. The proposed model is lightweight compared to other state-of-the-art models.

Andra Acsintoae et al. [34] proposed supervised open-set anomaly detection in video. The paper introduces UBnormal, a new benchmark for supervised open-set anomaly detection in video. UBnormal is the first benchmark to provide a validation set, which is essential for machine learning algorithms relying on hyperparameter tuning.

Guillermo del Castillo Torres et al. proposed the use of Explainable Artificial Intelligence (XAI) techniques, specifically LIME and CEM, to interpret the results obtained by Convolutional Neural Networks (CNN) in recognizing facial expressions. The authors showcased outcomes obtained from training a basic CNN model using the UIBVFED dataset. This dataset comprises synthetic avatars demonstrating 32 facial expressions categorized into six universally recognized emotions. The CNN model achieved a global accuracy of 88% and showed high recognition rates for emotions such as joy, anger, and fear. CNN models are often considered black-box models, providing no insight into the reasoning process behind their decisions.

In the context of artificial intelligence, Explainable Artificial Intelligence (XAI) has been developed to address this issue by providing means to interpret the results obtained by machine learning models. LIME, which stands for Local Interpretable Model-agnostic Explanations, highlights the areas of the image that contribute to a classification. CEM, or the Contrastive Explanation Method, provides explanations in a way that is natural for human classification. The extensive array of features within the images hampers the efficacy of CEM, which demonstrates notably more compelling outcomes with images of reduced complexity, such as those found in the MNIST dataset.

Yu Tian et al. [36] introduce the Robust Temporal Feature Magnitude learning (RTFM) method to address bias in Multiple Instance Learning (MIL) models towards dominant negative instances, especially in cases of subtle anomalies in

weakly supervised video anomaly detection. The RTFM approach trains a feature magnitude learning function to effectively recognize positive instances, substantially improving the robustness of the MIL approach to the negative instances from abnormal videos. This method incorporates feature magnitude learning, dilated convolutions, and self-attention mechanisms to capture both long- and short-range temporal dependencies. The paper reports that the RTFM-enabled MIL model outperforms several state-of-the-art methods on benchmark datasets and achieves significantly improved subtle anomaly discriminability and sample efficiency.

The results indicate the efficacy of the feature magnitude learning approach in improving the discriminative ability of the model, especially in identifying subtle anomalies, and it has the potential to enhance sample efficiency in weakly supervised anomaly detection scenarios. The proposed model was verified with well-known datasets, which are Avenue, SanghaiTech and UCSD Ped2. The proposed method achieved an accuracy of 99.74%. For improvement, the author suggested comparing their work with existing state-of-the-art approaches and refining the training process.

KY Gan [37] et al. presented an approach to video anomaly detection utilizing weakly supervised learning and contrastive regularization. The proposed novel approach U-Net-based architecture to capture both local and global temporal dependencies effectively. With the utilization of weakly supervised contrastive regularization, the model endeavors to mitigate overfitting by acquiring more broadly applicable features. The work is supported by extensive experimentation on the UCF-Crime dataset.

Alessandro Bruno et al. [38] presented an integrated solution for crowd behavior analysis using deep learning models, emphasizing the use of computer vision techniques to detect anomalies in crowd behavior at different scales. It highlights the multidisciplinary nature of crowd behavior analysis and the application of the proposed solution in the S4AllCities H2020 project. The method also involves supervised and unsupervised learning paradigms, and experiments have been carried out on the publicly available UCSD Anomaly Detection Dataset. The precision and recall values for YOLOv5 and DeepSORT in the experimental results are as follows:

For YOLOv5: Precision: 0.98, 0.93, 0.95, 0.94, 0.92 Recall: 0.75, 0.72, 0.71, 0.78, 0.70

For DeepSORT: Precision: 0.85, 0.89, 0.83, 0.86, 0.87 Recall: 0.74, 0.72, 0.69, 0.68, 0.72

These values represent the performance of the deep learning models in detecting and tracking pedestrians from CCTV cameras' video sequences, as reported in the experimental results.

Siqi Wang et al. [39] introduce a novel approach called Visual Cloze Completion (VCC), which aims to construct Visual Cloze Tests (VCTs) by erasing patches from a Spatio-Temporal Cube (STC) and training DNNs to complete the erased patches and their optical flow, thus enabling better VAD performance. The VCC method utilizes appearance and motion cues for video event extraction and employs ensemble strategies to fully exploit the temporal context and motion information in video events. Furthermore, the document suggests employing Optical Flow for motion-driven Region of Interest (RoI) extraction to augment the accuracy of event localization. Additionally, it introduces an upgraded VCC model-level approach known as the Spatio-Temporal UNet (ST-UNet), aimed at capturing more comprehensive video semantics and temporal context information, thereby enhancing the performance of Video Anomaly Detection (VAD).

Varghese et al.'s [40] research paper focuses on analyzing crowd behavior from cognitive and psychological perspectives using computer vision techniques. The author discussed various psychological theories of crowd behavior, including group mind theory, social comparison theory, and emergent-norm theory. These theories aim to explain how crowd behavior is influenced by social and psychological interactions, emphasizing the need to consider these factors in crowd behavior analysis.

The proposed method is the fusion of cognitive computing and psychological aspects, such as the integration of psychological theories and parameters with machine learning and deep learning methods. The author also discussed important datasets like UMN, UCSD Peds1 and Peds2, PETS 2009, UCF Normal/Abnormal Web Dataset, Violent-Flows, Motion Emotion Dataset (MED) and Crowd-11, out of which MED is the only one that contains the psychological parameter emotion as an intermediate attribute.

Zhang et al. [41] focus on the challenging task of predicting crowd emotion using video surveillance data, as conventional emotion clues such as facial expressions or body gestures are not easily discernible in crowded scenes. The proposed method describes crowd emotions using four attributes: enthalpy, magnitude variance, confusion index, and crowd density, which are given as input to a fuzzy inference system to evaluate arousal and valence in crowd emotion. The enthalpy value and magnitude variance serve as inputs for the arousal fuzzy inference system, while the confusion index and crowd density serve as inputs for the valence fuzzy system.

Fuzzy rules are formulated to deduce the emotion in the crowd scene by considering the interplay between arousal, valence, and crowd features. The output of the developed fuzzy system not only classifies emotions but also assigns scores to crowd emotions based on arousal and valence. Experimental findings illustrate the effectiveness of the proposed approach in assessing arousal and valence in crowd emotion, thereby

making a significant contribution to the domains of crowd behavior analysis and emotion recognition. This new dataset is formed by selecting related image video frames from four video surveillance datasets (UMN data set, PETS2009, UCF and MED). As a future work author suggested, there is a need to explore more emotion description models and collect more crowd image data by enriching the database and also the field of crowd behavior analysis along with emotion recognition for a better understanding of the crowd and to predict the efficient current state of the crowd.

Marsden et al. [42] proposed ResnetCrowd, a deep residual architecture designed for crowd counting, violent behavior detection, and crowd density level classification. The paper proposes a multi-objective approach, highlighting the lack of a labeled multi-task dataset for crowd analysis as a significant hindrance. The ResnetCrowd model is also evaluated on additional benchmarks (UMN Dataset) to highlight its superior generalization for crowd analysis models trained for multiple objectives. The results indicate that the proposed method achieves notable accuracy in identifying crowd behavior which is approx 0.79 mAUC.

Rezaee et al. [43] emphasize the importance of incorporating real-time security monitoring based on the Web of Things (WoT) platform and machine learning algorithms to enhance the influential detection of abnormal behaviors in crowds. The study also introduces the Motion Emotion Dataset (MED) as a crucial resource for investigating the diverse conditions influencing these methods. MED facilitates the analysis of crowd and individual behaviors for security screening of abnormal events. The author discussed the accuracy of various methods for crowd anomaly detection. It mentions that previous studies have reported accuracy ranging from 68.2% to a maximum of 73% for identifying the accuracy of the crowd anomaly behavior classifier. The Hybrid approach is able to claim an efficiency of 94.13%. In future work the author suggested implementing a hybrid model and WoT platform to reduce computational time and complexity.

The author suggested a bottom-up approach based on Deep Learning for detecting group activities, emphasizing contextual factors and human-to-human interactions. Their methodology utilizes Convolutional Neural Networks (CNN) to extract action-pose features and scene-related information, while Recurrent Neural Networks (RNN) are employed to track group dynamics. They introduce two distinct approaches: one employing Long Short-Term Memory (LSTM) and another utilizing Gated Recurrent Units (GRU).

Ullah et al. [44] developed a violence detection framework using a triple-staged deep learning approach and its application to three distinct datasets. The proposed method involves detecting persons in surveillance video streams using a pre-trained MobileNet-SSD CNN model and subsequently passing the sequence of frames to a 3D CNN model for spatiotemporal features extraction and analysis. The proposed model experimented on the violent crowd, violence in movies, and hokey fight datasets to evaluate the accuracy and performance of the proposed method with an achieved accuracy of 98% for the violent crowd dataset, 99.9% for the violence in movies dataset, and 96% for the hockey fight dataset.

In summary, crowd behavior analysis involves a number of processes, such as behavior prediction, density estimates, tracking, counting, and crowd modeling. Crowd-related data is gathered through both non-visual sensors, like cell phones, and visual sensors, such as CCTV cameras. This allows for a more thorough understanding of crowd dynamics and behavior. Real-time prediction challenges have benefited greatly from the incorporation of AI techniques, especially supervised and unsupervised learning, which has made it possible to create crowd analysis models with better generalization capabilities. Numerous research studies have shown that deep learning architectures can perform well in tasks including emotion recognition, forecasting, emotion and behavior analysis, and semantic knowledge extraction.

Over the years, research on crowd analysis has changed, with thorough surveys emphasizing the value of the technique in a range of applications, including intelligent environments, public space design, crowd management, and visual surveillance. Through interdisciplinary collaboration, insights into the intricate dynamics of human collective behavior have been gained, including parallels to biological phenomena such as swarming insects. Research has highlighted the usefulness of crowd behavior research in managing disasters, planning metropolitan areas, and creating efficient crowd control techniques. The use of cutting-edge approaches, including spatiotemporal methodologies, convolutional neural networks, optical flow, hidden Markov models, and Gaussian mixture models, is one of the most recent developments in the study of crowd behavior. These techniques have been shown to be more accurate in detecting aberrant behavior and crowd emotions, opening the door for crowd analysis and management systems that are more effective. Furthermore, the creation of extensive datasets like AffectNet and the Motion Emotion Dataset (MED) has made useful resources available for the training and validation of crowd analysis algorithms. These datasets include psychological factors that improve our comprehension of crowd behavior and the identification of emotions.

Research on crowd behavior analysis will likely go in several directions in the future, such as investigating more complex emotion description models, gathering more crowd image data to enhance databases, and combining machine learning algorithms with real-time security monitoring platforms to enhance the detection of abnormal crowd behaviors. All things considered, these developments are highly promising for raising the efficacy and efficiency of crowd analysis systems in a range of practical uses.

Crowd behavior analysis involves the different stages like, Detection, Tracking, Feature Extraction, Classification and Anomaly Detection. From the mentioned stages, Detection and Stages are broadly addressed by researchers with significant accuracy. Feature Extraction has come into the keen interest of researchers after the introduction of Deep Learning algorithms, which compute the set of metrics with the dynamic features, topology structure and affective state of the crowd. From the extracted features, the Behavior classification and anomaly detection can be recognized from the video sequences.

3. Discussion of Machine Learning, CNN and its Significance in this Study

Artificial Intelligence (AI) assumes a pivotal role in prediction and automation tasks. With society transitioning towards automation and an increasing focus on ensuring people's safety, concerns arise regarding crowd management, particularly in situations where large gatherings occur. The potential for disasters in crowded environments poses a significant risk to human lives. While manual observation of CCTV cameras aids in crowd management, it is limited by human capabilities and may overlook critical observations, potentially leading to significant issues.

Machine learning techniques play a crucial role in crowd abnormality analysis, offering effective methods for detecting unusual patterns or behaviors within crowds through the utilization of various machine learning algorithms, such as supervised learning, unsupervised learning, and reinforcement learning. Supervised learning algorithms, for instance, enable the training of models on labeled datasets where abnormal crowd behaviors are explicitly identified. These models can then generalize patterns from the training data to detect anomalies in new, unseen crowd scenes.

Unsupervised learning techniques, on the other hand, allow for the detection of anomalies without labeled data, relying on the identification of patterns that deviate significantly from the norm within the crowd. The application of machine learning in crowd abnormality analysis provides a powerful tool for enhancing crowd management, public safety, and security in various settings, such as public events, transportation hubs, and urban environments.

As new deep learning models emerge, prototypes for crowd anomaly detection have transitioned from conventional 2DCNN to 3DCNN. Moreover, there has been a shift from standard deep learning methods to more sophisticated approaches such as attention-based, transformer-based, and even quantum computing-based learning techniques.

A typical Deep Learning neural network architecture in computer vision is the Convolutional Neural Network (CNN). An artificial intelligence field called computer vision makes it possible for a computer to comprehend and analyze an image or other visual data. [45] The Convolutional Neural Network (CNN) is an expanded form of Artificial Neural Network (ANN), primarily employed for extracting features from grid-like matrix datasets. It finds significant application in visual datasets such as images or videos, where discerning data patterns is crucial. Figure 5 shows the typical architecture of the CNN. CNN accomplishes multiple layers like the input layer, Convolutional layer, Pooling layer, and fully connected layers. The Convolutional layer uses filters on the input image to extract features, followed by the Pooling layer, which downsamples the image to decrease the computational load.

Finally, the fully connected layer is responsible for generating the ultimate prediction. The network refines its filters through the process of backpropagation and gradient descent to learn the most effective patterns.

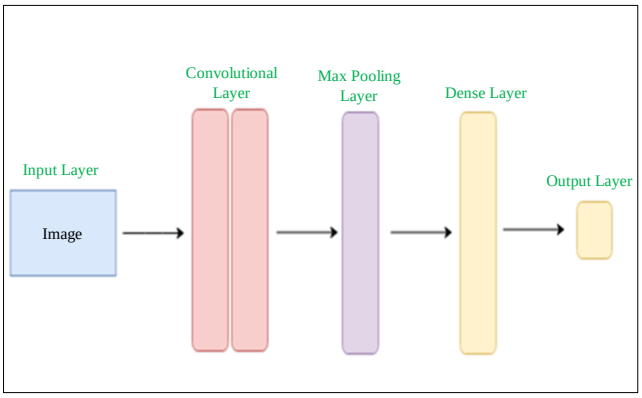

Fig. 5 Simple CNN architecture [45]

Tripathi et al. [46] presented a comprehensive survey of current convolutional neural network (CNN)-based methods for crowd behavior analysis. The author has reviewed more than 100 research papers based on the topic and concluded that CNN-based approaches are poised to lead future investigations in the realm of crowd analytics. The main aim of the survey was to achieve real-time and accurate visual surveillance of the crowded place to assist law enforcement. In this research, an attempt was made to introduce a lightweight CNN model capable of generating efficient results in minimal time. In the future, the model is planned to be embedded in a device that will detect abnormalities and notify authorities in a very short time.

4. Existing Datasets for Crowd Analysis

In this section, the discussion will cover various crowd-related video datasets. Datasets are used to evaluate the performance of crowd behavior analysis algorithms. Many prior studies have acknowledged shortcomings in their research, including inadequate datasets, insufficient image quantities, challenges in image labeling, limited classifier options (often just two or a restricted set), and a lack of specific evidence in emotion and behavior labeling.

Table 2 shows the list of various datasets available in the public and private domain with various properties. The details are collected from the [4], the comprehensive review paper which has presented an in-depth review of crowd behavior analysis and its applications. The table also includes the live streaming available for surveillance.

Datasets listed in the survey paper and other research topics, in conclusion, while there are existing datasets for crowd abnormal behavior analysis, it is evident that there are still notable gaps and challenges that need to be addressed. The current datasets vary in terms of size, diversity, and quality, leading to limitations in the robustness and generalization of models trained on them. To enhance the accuracy and efficiency, a more comprehensive and standardized dataset for the crowd's abnormal behavior is needed.

5. Adoption of MED (Motion Emotion Dataset) and Preprocessing

In the previous section, insights regarding the various crowd-related datasets are identified and studied. Researchers have applied their proposed model to these datasets to verify their work. According to previous research work, people focused on identifying the abnormality in the video sequence. So, in the end, they were able to classify two classes: normal events and abnormal events. In very little research, the emotional and behavioral characteristics of the crowd were identified. The state-of-the-art train and test datasets are mostly classified into normal and abnormal images or videos. Sad, Cry, and Happy such kind of emotions and Congested, Normal, and panic kind of behavior need to be labeled in the datasets in order to get more detail on the current crowd situation.

Our objective in doing this research is to categorize the crowd according to their emotions and behavioral traits. Therefore, the Motion Emotion Dataset (MED) Dataser has been embraced, which was released in 2016 by Haddadnia et al. [47]. The Motion Emotion Dataset (MED) is a comprehensive collection designed to capture the nuances of human emotions through motion and behavior. Comprising annotated videos portraying a wide spectrum of emotional states, MED offers researchers a rich resource for investigating affective computing and human behavior analysis. With its diverse range of human movements and expressions, the dataset aims to facilitate advancements in emotion recognition algorithms by providing ground truth labels for various emotional categories, including joy, sadness, anger, surprise, and more. MED covers a broad array of scenarios, encompassing both individual and group interactions to capture the complexities of social dynamics in emotional expression. This diversity enables researchers to develop and evaluate emotion recognition models in dynamic contexts, fostering deeper insights into human-computer interaction and emotional understanding.

Accessible and well-documented, the Motion Emotion Dataset stands as a valuable asset for the research community, driving progress in the understanding of emotions through motion. The dataset is available for download from [55], and the author published the frame annotation as well a total of 31 videos of approx. Length 1-2 min can be downloaded from the above mentioned link.

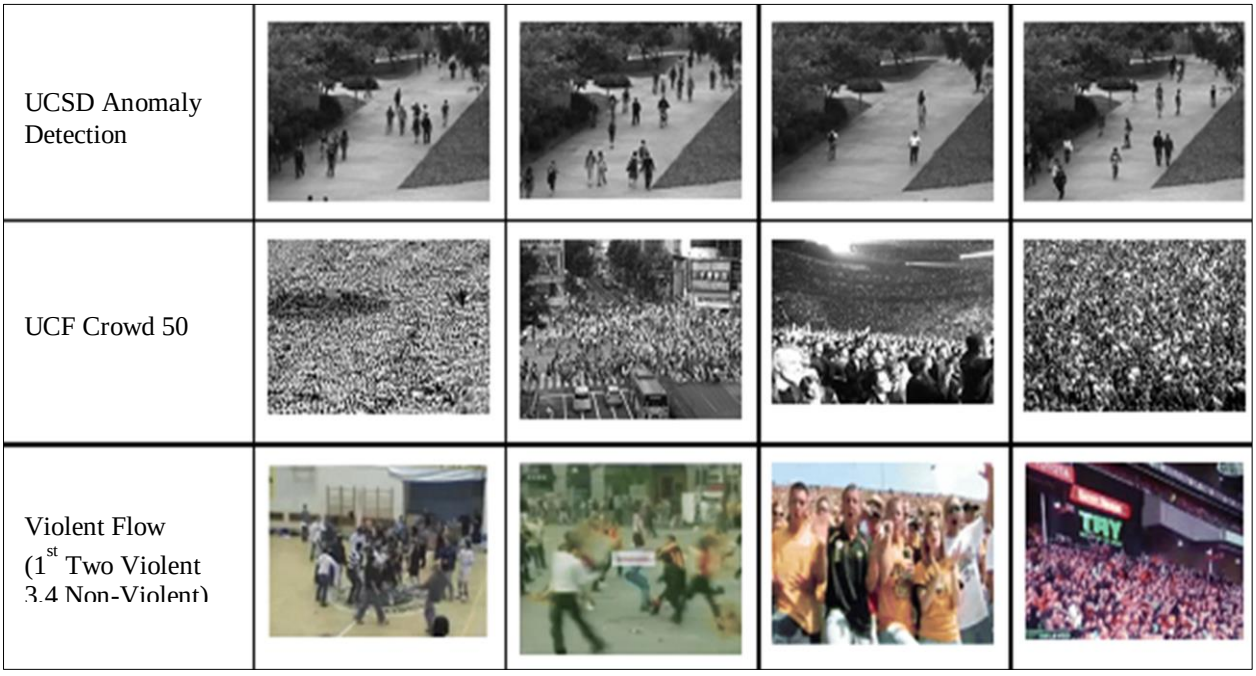

Fig. 6 Sample images from well-known datasets [4]

The author has annotated the frames of videos; for example, video01(1,1:1175)=6, which indicates that in video01, frames from 1 to 1175 belong to the number 6 emotion category that is Neutral. For behavior, 0:nothing, 1:Panic, 2:Fight, 3:Congestion, 4:Obstacle or abnormal object, 5:Neutral total 6 classes are categorized. Similar way for Emotion, 0:nothing, 1:Angry, 2:Happy, 3:Excited, 4:Scared, 5:Sad, 6:Neutral classes are categorized. Mounir Bendali-Braham et. al. [4] presented survey paper, in which the Datasets for the Crowd behavior along with crowd emotion and Crowd Anomaly detection are provided. The author has gone through the dataset starting from the year 2009 to 2019, along with the few live streaming available on the Internet.

5.1. Preprocessing of Dataset

The MED Dataset is downloaded from the above mentioned link. The folder contains the list of labels, 31 videos, and two files containing the frame no and its class name mapping. Following are some details of the dataset.

Table 1. Annotation names and numbers for each frame

Behavior	Emotion
(Total 6 Classes)	(Total 7 Classes)
0: Nothing	0: Nothing
1: Panic	1: Angry
2: Fight	2: Happy
3: Congestion	3: Excited
4: Obstacle or Abnormal Object	4: Scared
5: Neutral	5: Sad
	6: Neutral

Number of Videos in Dataset : 31 videos
Frame Size : 480 (Height) X 854 (Width)
Frame Rate : 30 fmps

Video wise frame annotation is provided along with the dataset in the following way

```
video01 = zeros(1,1680);
video01(1,1:1175)=6;
video01(1,1176:1515)=3;
video01(1,1516:1600)=4;
video01(1,1601:end)= 0;
emotionlabels{1} = video01;
```

Algorithm:

Video Vi will be divided into Fi frames
$Fi = \sum (Vi)$
Datset_Labeli = Range(X : Y) ,
X is the start of the class, and Y is the end frame number of the class
$\sum Li = \sum Fi\ (D\_Li)$

Start:
1. Divide the videos in frames and keep them in the folder as the video name
2. As per the annotation, put the frames into their class label folders to generate a training dataset and maintain 80%
3. From each class, 20% of validation frames are to be copied to the validation folder after shuffling the frames/images.
4. Steps 1-3 are to be followed for all 31 videos.

End:

After completion of the preprocessing steps, train and valid folders are generated for each class for the Behavior and the Emotion. Table 2 shows the image-wise count for each label.

6. Proposed Methodology and Implementation

After the preparedness of the dataset, the CNN model is applied to the dataset for the train and test set. In which 80% of the images were used for training and 20% for testing. Here is the proposed method in which to train the 2D CNN model on MED Dataset. This is considered as our supervised approach.

This paper presents an approach to the CNN model for feature selection of the crowd based on emotion and behavior classes, as illustrated in Figure 7. The architecture of the 2D CNN model is shown in Figure 7.

Table 2. Number of images in each labeled class after preprocessing

Emotion	Behavior
Train: Happy → 1991 Excited → 3802 Sad→1155 Scared → 2167 Neutral → 25476 Angry → 5752 Nothing → 3974 Total Images: 44317 (7 Classes)	Train: Congestion → 2379 Obstacle → 6380 Neutral → 26029 Fight → 4469 Panic → 1991 Nothing → 4040 Total Images: 45288 (6 Classes)
Test (Validation) : Happy → 400 Excited → 761 Sad → 231 Scared → 435 Neutral → 6000 Angry →1150 Nothing →800 Total Images: 9777	Test (Validation) : Congestion → 476 Obstacle → 1276 Neutral → 5206 Fight → 894 Panic → 400 Nothing → 808 Total Images: 9060

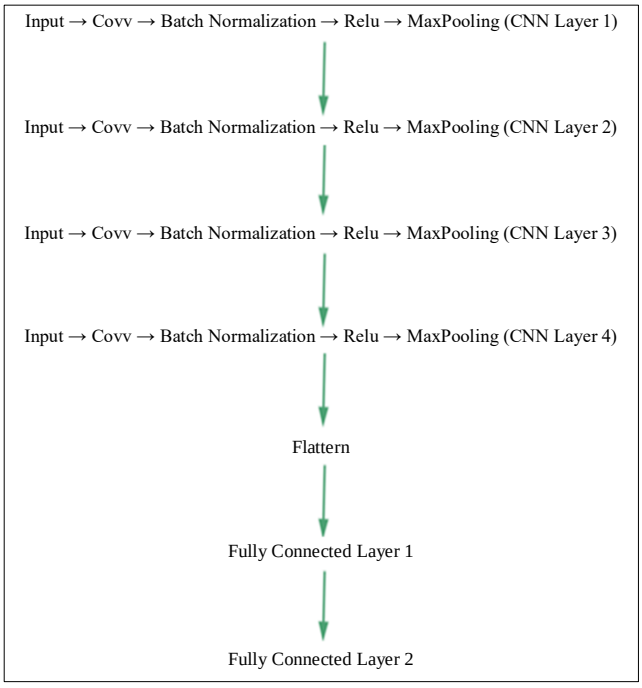

Fig. 7 Lightweight 2D CNN architecture

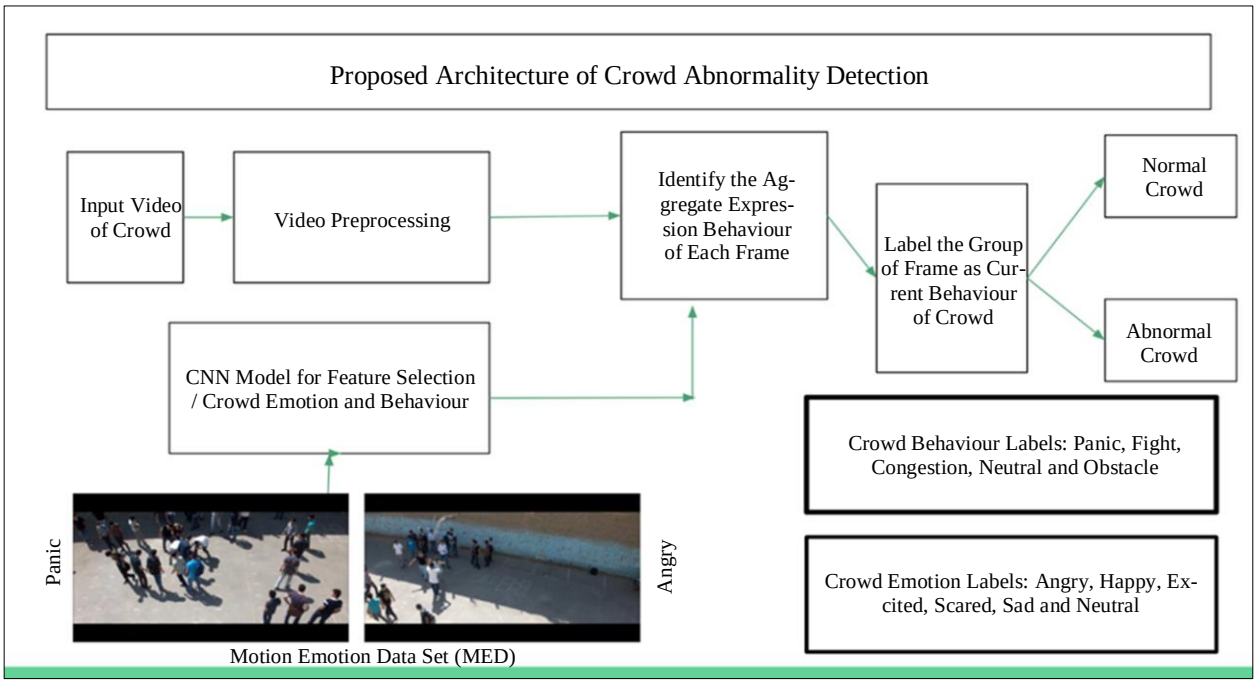

Fig. 8 Proposed architecture of crowd abnormality detection

The RELU activation function is applied to change negative values to 0, which will remove unnecessary features. A total of 4 CNN layers are used in the proposed architecture.

In the proposed lightweight 2D CNN architecture, the input images are initially applied for the convolution for feature extraction.

6.1. Implementation

Anaconda open-source package and environment management system are used to implement the proposed architecture using Python 3.11. Anaconda Navigator comes with a Jupyter Notebook that is used for batch execution and editing. Python libraries for Keras are used for various functions.

During the experiment, the size of the image was converted to 168 (Height) X 300 (Width). The experiment was executed on a Dell POWEREDGE R740 server with NVIDIA Tesla V100 32G Passive GPU, 2 Nos x Intel Xeon Gold 5118 2.3G, and 4 Nos x 32GB RDIMM 2666MT/s Dual Rank RAM.

In addition, as part of the experiments, different Hyperparameters are applied to achieve efficient results. The hyperparameter, epoch = 10, Learning Rate = 0.0001, Batch Size = 32, Kernel Size = 3 X 3 and Adam optimizer, along with early stopping, are set for both emotion and behavior datasets. The approximate training time was between 8 to 9 hours for both datasets on the High End GPU.

7. Results, Validation and Discussion

After the execution, the proposed model with the parameters mentioned above is able to achieve a Testing Loss: 0.0226 - Testing Accuracy: 99.03% - Validation Loss: 0.0143 and Validation Accuracy: of 99.18% for the emotion dataset.

For the Behavior dataset, Testing Loss: 0.0484 -Testing Accuracy: 96.95% - Validation Loss: 0.0450 - Validation Accuracy: 97.22% are achieved.

To validate the CNN model, a confusion matrix was generated using the trained model for both the emotion and behavior datasets. A confusion matrix is a performance measurement for machine learning classification problems where the output can be two or more classes. It is a table with four different combinations of predicted and actual classes, and it is often used to understand the performance of the classification algorithm.

Each cell in the confusion matrix represents the counts or percentages of instances where the predicted class matches or differs from the actual class. Figure 11 indicates the confusion matrix for the emotion dataset. For each class, almost all the validation images are truly classified. In the case of neutral, 54 images are wrongly classified as angry.

Figure 12 indicates the confusion matrix for the behavior dataset. For each class, almost all the validation images are truly classified, whereas, in the case of neutral, 182 images are wrongly classified as obstacles. The reason for the wrong classification is that those images are wrongly labeled in the MED dataset.

The results of the proposed lightweight model are compared with other state-of-the-art architectures applied to the MED dataset. Table indicates the comparison of our lightweight 2D CNN model with other state-of-the-art models. In the last row, we can see significant improvement in the accuracy.

In addition, we have also achieved 99.03% accuracy on the emotion dataset, which has not ever been experimented with. As per Table 3, Rabiee H, Haddadnia et al. [47] proposed the novel dataset MED and applied the different approaches. The models 3DCNN, V3G, C3D, Dense Trajectories, CNN, and Cognitive deep model applied to the MED dataset had High Computational complexity [47, 50, 51, 52, 54]. SVM and AlexNet models had Medium Complexity. Our proposed model gives optimum efficiency with Low Complexity. To enhance the precision of our models, the hyperparameters with both the Adam and SGD optimizers on both Emotion and Behavior datasets are applied.

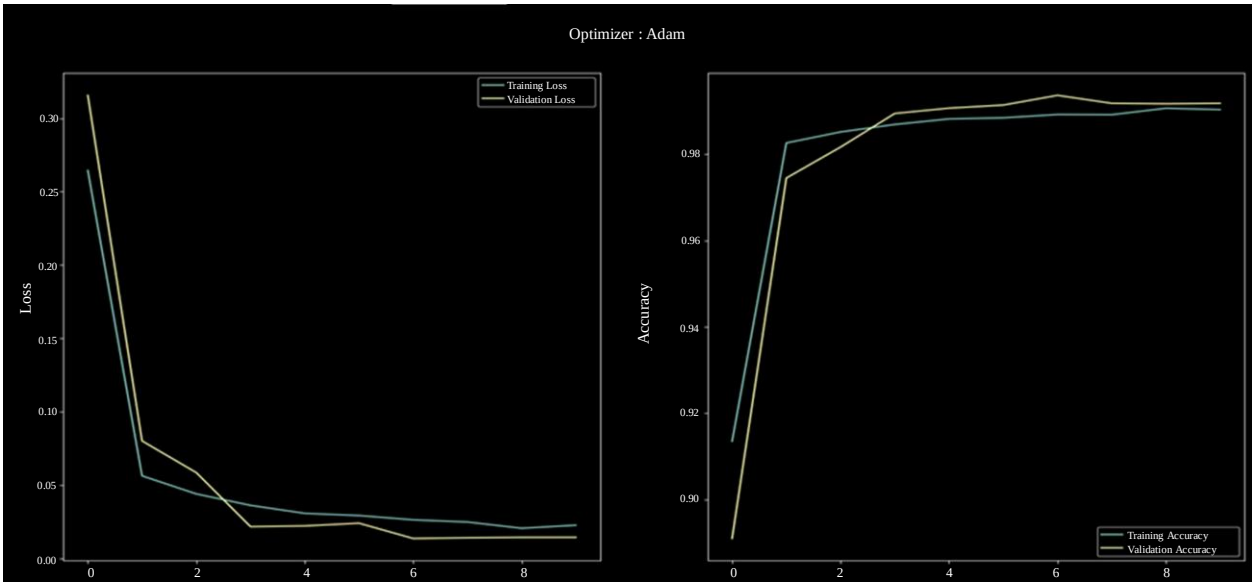

Fig. 9 Graph for emotion dataset

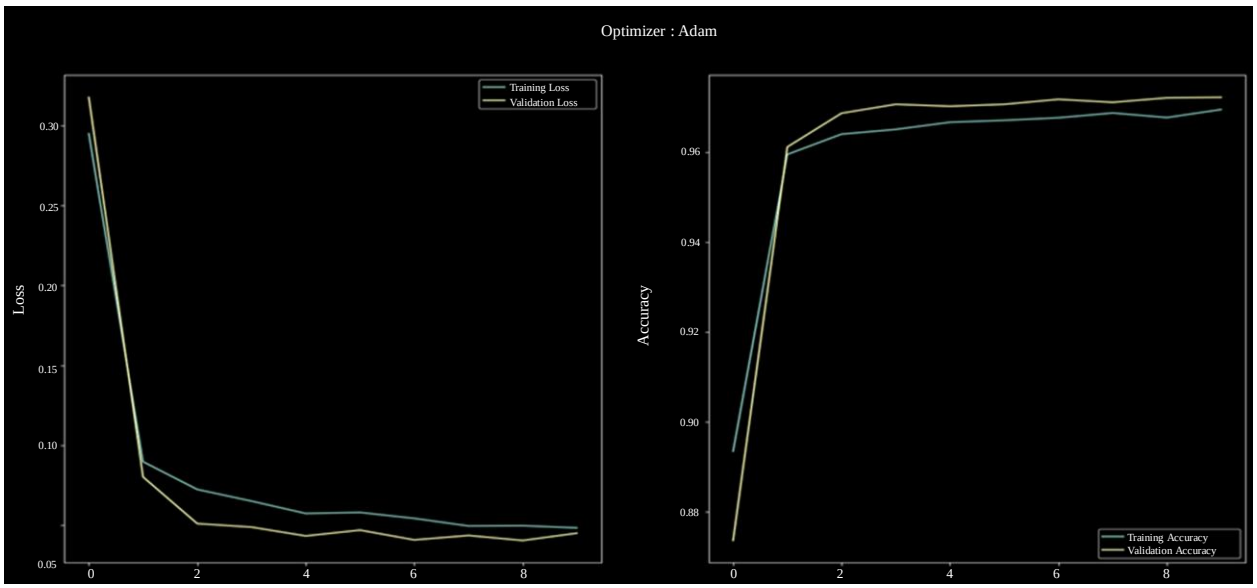

Fig. 10 Graph for behavior dataset

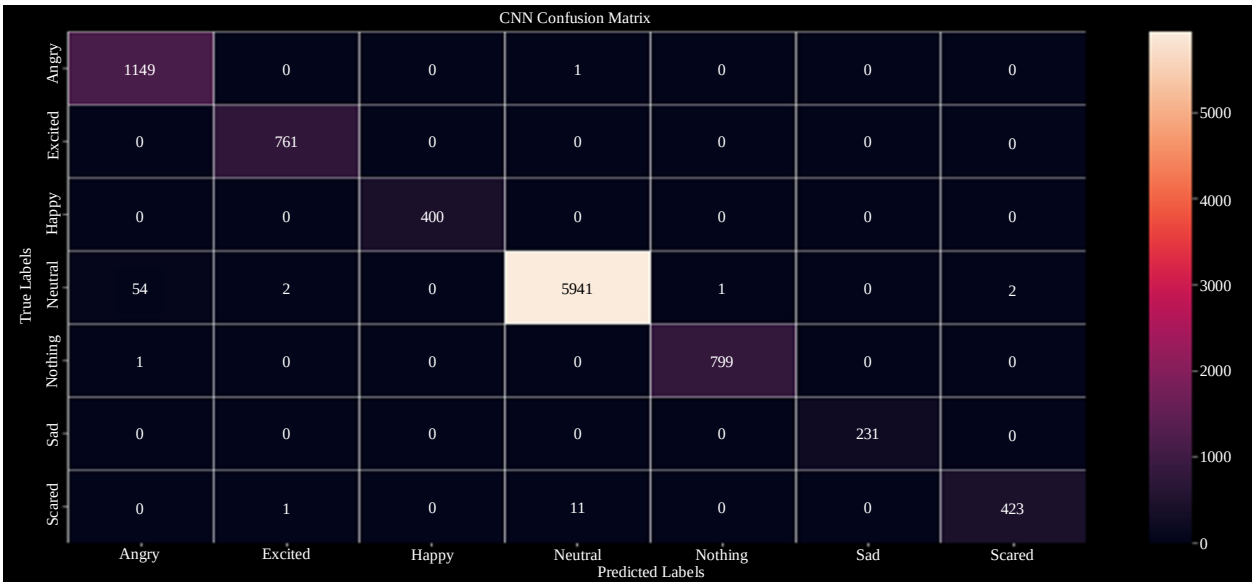

Fig. 11 Confusion matrix for emotion dataset

The systematically experiment with different combinations of learning rates, batch sizes, and numbers of epochs on our proposed architecture to thoroughly explore their impact on performance and attain comprehensive results. The early stopping and learning weight from the previously trained model is applied to gain good accuracy. For the experiment, batch sizes 16, 32 and 64 are considered with a learning rate of 0.01, 0.001 and 0.0001 and a number of epochs are taken 15.

In experiments performed on the Emotion Dataset, the best results were a testing accuracy of 99.16% and validation accuracy of 99.37%, achieved for Batch size 64 learning rate of 0.0001 and a number of epoch are 15 without early stopping. For the remaining parameters, it stopped early on about 5, 6, 9, 11, etc, epoch. The same experiment was carried out on the Behavior dataset. The best results were a testing accuracy of 97.19% and a validation accuracy of 97.29%, achieved for Batch size 64 learning rate of 0.0001 and the number of epochs are 15 without early stopping. With regards to the behavior dataset as well, the model terminated prematurely, achieving less than the highest percentage of accuracy.

The SGD optimizer was also applied with the proposed model with batch size 32, learning rate 0.001 and number of epoch 15. As a result, it stopped early after 6 epochs. Training accuracy was 94.04%, and testing accuracy was 29.69%.

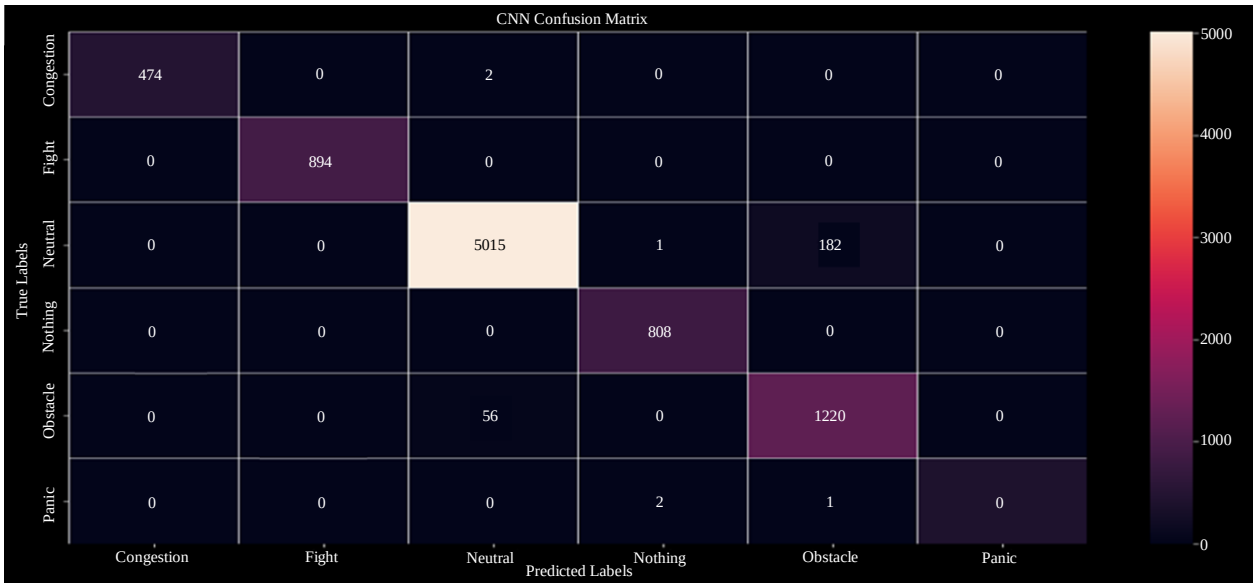

Fig. 12 Confusion matrix for behavior dataset

Table 3. Comparison of the proposed model with another state-of-the-art model

Ref.	Type	Dataset	Model Type	Accuracy (%)
[47]	Histogram of Optical Flow (HOF)	MED	SVM	37.69
[47]	Tracklet	MED	SVM and k-NN	38.17
[47]	Motion Boundary Histogram (MBH)	MED	SVM	38.8
[50]	Automated	MED	3DCNN	34.05
[50]	Automated	MED	V3G	36.99
[50]	Automated	MED	C3D	51.22
[51]	Automated	MED	Dense Trajectories	43.64
[52]	Automated	MED	CNN	71.7
[53]	Automated	MED	3DCNN	90.91
[54]	Automated	MED	Cognitive deep model	93.82
[43]	Automated	MED	Tracking and AlexNet	94.13
Proposed	Automated	**MED**	**Light Weight CNN Model**	**97.22**

8. Conclusion and Future Scope

In this study, we have explored the application of a lightweight CNN model for crowd behavior analysis on the Motion Emotion Dataset (MED). Our investigation aimed to address the challenges of computational complexity often associated with deploying deep learning models in crowd analysis tasks, particularly in real-world scenarios where computational resources are limited.

By employing a meticulously crafted lightweight CNN structure, we have showcased encouraging outcomes in tasks related to analyzing crowd behavior and emotion. The evaluation of the proposed approach showcased competitive performance on the MED dataset while maintaining reduced computational overhead compared to more complex CNN architectures in addition to the significant accuracy.

The proposed model is optimized by applying various hyper-parameter, and it concludes that, it gives 99.37% accuracy with ADAM optimizer with 64 batch size and 0.0001 learning rate. It did not perform well with the SGD optimizer. Overall, our findings underscore the significance of leveraging lightweight CNN models for crowd behavior analysis, offering

scalable and efficient solutions for real-world deployment. Future research directions may involve further optimization of the lightweight architecture, exploration of additional datasets, and integration of multi-modal data sources to enhance the robustness and applicability of crowd analysis systems in various domains.

Conflicts of Interest

A competing interest exists when a secondary interest, such as financial gain, influences professional judgment concerning the validity of research. We require that our authors reveal any possible conflict of interest in their submitted manuscripts.

References

[1] Swathi H.Y., G. Shivakumar, and H.S. Mohana, "Crowd Behavior Analysis: A Survey," *2017 International Conference on Recent Advances in Electronics and Communication Technology (ICRAECT)*, Bangalore, India, pp. 169-178, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[2] Barbara Krausz, and Christian Bauckhage, "Loveparade 2010: Automatic Video Analysis of a Crowd Disaster," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 307-319, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[3] Fatih Porikli et al., "Video Surveillance: Past, Present, and Now the Future [DSP Forum]," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 190-198, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[4] Mounir Bendali-Braham et al., "Recent Trends in Crowd Analysis: A Review," *Machine Learning with Applications*, vol. 4, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Bharath Varma Avs, List of Mob Lynching Incidents in India – 2019, Medium, 2019. [Online]. Available: https://medium.com/@bharathvarmaavs/list-of-mob-lynching-incidents-in-india-2019-5b97773f677f

[6] India's Citizenship Protests - How over Three Months of Protests have Unfolded, Reuters Graphics, 2020. [Online]. Available: https://www.reuters.com/graphics/INDIA-CITIZENSHIP/PROTESTS/jxlbpgqlpqd/index.html

[7] Spriha Srivastava, India Revokes Special Status for Kashmir. Here's what it Means, CNBC, 2019. [Online]. Available: https://www.cnbc.com/2019/08/05/article-370-what-is-happening-in-kashmir-india-revokes-special-status.html

[8] Prabhash K Dutta, Beyond JNU Violence: From Renaissance to Bloodshed, A Campus Story, India Today, 2020. [Online]. Available: https://www.indiatoday.in/news-analysis/story/beyond-jnu-when-university-campuses-different-jamia-amu-1634384-2020-01-06

[9] Amrit Dhillon, Students Protest across India after Attack at Top Delhi University, The Guardian, 2020. [Online]. Available: https://www.theguardian.com/world/2020/jan/06/students-injured-in-india-after-masked-attackers-raid-top-university

[10] Morgot Cohen, A History of Violence at Indian Universities, The National, 2010. [Online]. Available: https://www.thenationalnews.com/world/asia/a-history-of-violence-at-indian-universities-1.557030

[11] Bhawana Tyagi, Swati Nigam, and Rajiv Singh, "A Review of Deep Learning Techniques for Crowd Behavior Analysis," *Archives of Computational Methods in Engineering*, vol. 29, no. 7, pp. 5427-5455, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[12] India Protest: Farmers Breach Delhi's Red Fort in Huge Tractor Rally, BBC, 2021. [Online]. Available: https://www.bbc.com/news/uk-55793731

[13] Mehdi Moussaïd, Dirk Helbing, and Guy Theraulaz, "How Simple Rules Determine Pedestrian Behavior and Crowd Disasters," *Applied Physical Sciences*, vol. 108, no. 17, pp. 6884-6888, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[14] Moin Nabi, Alessio Del Bue, and Vittorio Murino, "Temporal Poselets for Collective Activity Detection and Recognition," *2013 IEEE International Conference on Computer Vision Workshops*, Sydney, NSW, Australia, pp. 500-507, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[15] Barry G. Silverman et al., "Human Behavior Models for Agents in Simulators and Games: Part II: Gamebot Engineering with PMFserv," *Presence: Teleoperators and Virtual Environments*, vol. 15, no. 2, pp. 163-185, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[16] Hamidreza Rabiee et al., "Emotion-Based Crowd Representation for Abnormality Detection," *arXiv*, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[17] Noah Sulman et al., "How Effective is Human Video Surveillance Performance?," *2008 19th International Conference on Pattern Recognition*, Tampa, FL, USA, pp. 1-3, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[18] Gurjit Singh Walia, and Rajiv Kapoor, "Recent Advances on Multicue Object Tracking: A Survey," *Artificial Intelligence Review*, vol. 46, no. 1, pp. 1-39, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[19] Jason M. Grant, and Patrick J. Flynn, "Crowd Scene Understanding from Video: A Survey," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 2, pp. 1-23, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[20] Ovgu Ozturk, Toshihiko Yamasaki, and Kiyoharu Aizawa, "Detecting Dominant Motion Flows in Unstructured / Structured Crowd Scenes," *2010 20th International Conference on Pattern Recognition*, Istanbul, Turkey, pp. 3533-3536, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[21] Adriana Draghici, and Maarten Van Steen, "A Survey of Techniques for Automatically Sensing the Behavior of a Crowd," *ACM Computing Surveys*, vol. 51, no. 1, pp. 1-40, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[22] Smriti Srivastava et al., *Concepts and Real-Time Applications of Deep Learning*, 1st ed., Springer Cham, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[23] M. Sami Zitouni et al., "Advances and Trends in Visual Crowd Analysis: A Systematic Survey and Evaluation of Crowd Modelling Techniques," *Neurocomputing*, vol. 186, pp. 139-159, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[24] Jie Zhang et al., "Collaboratively Self-Supervised Video Representation Learning for Action Recognition," *arXiv*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[25] H.Y. Swathi, and G. Shivakumar, "Hybrid Feature-Assisted Neural Model for Crowd Behavior Analysis," *SN Computer Science*, vol. 2, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[26] Guodong Li, Lei Wang, and Minzhong Wu, "Crowd Behavior Intervention Based on Emotional Contagion," *Proceedings of the 2022 6th International Conference on Computer Science and Artificial Intelligence*, pp. 1-6, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[27] Minzhong Wu, Lei Wang, and Guodong Li, "Crowd Emotion Recognition Based on Causal Spatiotemporal Structure," *Proceedings of the 8th International Conference on Computing and Artificial Intelligence (ICCAI '22)*, pp. 368-374, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[28] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor, "Affectnet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18-31, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[29] Tian Wang et al., "Abnormal Event Detection via the Analysis of Multi-Frame Optical Flow Information," *Frontiers of Computer Science*, vol. 14, pp. 304-313, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[30] Fernando J. Rendón-Segador et al., "Crimenet: Neural Structured Learning Using Vision Transformer for Violence Detection," *Neural Networks*, vol. 161, pp. 318-329, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[31] Biao Guo et al., "Two-Stream Spatial-Temporal Auto-Encoder with Adversarial Training for Video Anomaly Detection," *IEEE Access*, vol. 12, pp. 125881-125889, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[32] Manu Yadakere Murthygowda, Ravikumar Guralamata Krishnegowda, and Shashikala Salekoppalu Venkataramu, "An Integrated Multi-Level Feature Fusion Framework for Crowd Behaviour Prediction and Analysis," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 30, no. 3, pp. 1369-1380, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[33] Nicolae-Catalin Ristea et al., "Self-Distilled Masked Auto-Encoders are Efficient Video Anomaly Detectors," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15984-15995, 2024. [Google Scholar] [Publisher Link]

[34] Andra Acsintoae et al., "Ubnormal: New Benchmark for Supervised Open-Set Video Anomaly Detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20143-20153, 2022. [Google Scholar] [Publisher Link]

[35] Guillermo del Castillo Torres et al., "Understanding How CNNs Recognize Facial Expressions: A Case Study with LIME and CEM," *Sensors*, vol. 23, no. 1, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[36] Yu Tian et al., "Weakly-Supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4975-4986, 2021. [Google Scholar] [Publisher Link]

[37] Kian Yu Gan et al., "Contrastive-Regularized U-Net for Video Anomaly Detection," *IEEE Access*, vol. 11, pp. 36658-36671, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[38] Alessandro Bruno et al., "High-Level Feature Extraction for Crowd Behaviour Analysis: A Computer Vision Approach," *Image Analysis and Processing. ICIAP 2022 Workshops*, pp. 59-70, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[39] Siqi Wang et al., "Video Abnormal Event Detection by Learning to Complete Visual Cloze Tests," *arXiv Preprint*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[40] Elizabeth B. Varghese, and Sabu M. Thampi, "Towards the Cognitive and Psychological Perspectives of Crowd Behaviour: A Vision-Based Analysis," *Connection Science*, vol. 33, no. 2, pp. 380-405, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[41] Xuguang Zhang et al., "Crowd Emotion Evaluation Based on Fuzzy Inference of Arousal and Valence," *Neurocomputing*, vol. 445, pp. 194-205, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[42] Mark Marsden et al., "Resnetcrowd: A Residual Deep Learning Architecture for Crowd Counting, Violent Behaviour Detection and Crowd Density Level Classification," *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Lecce, Italy, pp. 1-7, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[43] Khosro Rezaee et al., "A Survey on Deep Learning-Based Real-Time Crowd Anomaly Detection for Secure Distributed Video Surveillance," *Personal and Ubiquitous Computing*, vol. 28, pp. 135-151, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[44] Fath U Min Ullah et al., "Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network," *Sensors*, vol. 19, no. 11, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[45] Introduction to Convolution Neural Network. [Online]. Available: https://www.geeksforgeeks.org/introduction-convolution-neural-network/

[46] Gaurav Tripathi, Kuldeep Singh, and Dinesh Kumar Vishwakarma, "Convolutional Neural Networks for Crowd Behaviour Analysis: A Survey," *The Visual Computer*, vol. 35, pp. 753-776, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[47] Hamidreza Rabiee et al., "Novel Dataset for Fine-Grained Abnormal Behavior Understanding in Crowd," *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 95-101, Colorado Springs, CO, USA, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[48] Ven Jyn Kok, Mei Kuan Lim, and Chee Seng Chan, "Crowd Behavior Analysis: A Review Where Physics Meets Biology," *Neurocomputing*, vol. 177, pp. 342-362, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[49] A.A. Afiq et al., "A Review on Classifying Abnormal Behavior in Crowd Scene," *Journal of Visual Communication and Image Representation*, vol. 58, pp. 285-303, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[50] Camille Dupont, Luis Tobías, and Bertrand Luvison, "Crowd-11: A Dataset for Fine Grained Crowd Behaviour Analysis," *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA, pp. 9-16, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[51] Hamidreza Rabiee, Javad Haddadnia, and Hossein Mousavi, "Crowd Behavior Representation: An Attribute-Based Approach," *SpringerPlus*, vol. 5, no. 1, pp. 1-7, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[52] Lazaros Lazaridis, Anastasios Dimou, and Petros Daras, "Abnormal Behavior Detection in Crowded Scenes Using Density Heatmaps and Optical Flow," *2018 26th European Signal Processing Conference (EUSIPCO)*, Rome, Italy, pp. 2060-2064, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[53] Elizabeth B. Varghese, and Sabu M. Thampi, "A Deep Learning Approach to Predict Crowd Behavior Based on Emotion," *Smart Multimedia*, pp. 296-307, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[54] Elizabeth B. Varghese, Sabu M. Thampi, and Stefano Berretti, "A Psychologically Inspired Fuzzy Cognitive Deep Learning Framework to Predict Crowd Behavior," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 1005-1022, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[55] Motion Emotion Dataset (MED), Github. [Online]. Available: https://github.com/hosseinm/med