ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov636

Hybrid Ensemble Learning for Real-Time Crowd Behavior Analysis Using Optical Flow and Deep Motion Features

Doaa Mabrouk[1]*; Manal A. Abdel-Fattah[2]; Ahmed Taha[3]

[1]Software Engineering Department, Faculty of Engineering & Technology, Egyptian Chinese University, Cairo, Egypt

[2]Information Systems Department, Faculty of Computers and Artificial Intelligence, Helwan University, Helwan, Egypt

[3]Computer Science Department, Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt

Corresponding Author: Doaa Mabrouk*

Publication Date: 2025/11/19

Abstract— Crowd Behavior analysis has become a crucial component of modern video surveillance systems, enabling automatic detection of abnormal events such as panic, congestion, and violence. Traditional approaches often fail to generalize under complex environmental conditions, while deep learning methods alone require large datasets and extensive computation. This paper proposes a hybrid ensemble learning framework that integrates optical flow–based motion features with deep motion representations extracted from convolutional neural networks (CNNs) to achieve real-time and robust crowd behavior recognition. The ensemble model combines Random Forests (RFs), Gradient Boosting (GB), and a lightweight CNN classifier via weighted voting. Experiments conducted on benchmark datasets, such as the UCSD Anomaly Detection Dataset and Violent Flows (VF), demonstrate that the proposed framework outperforms individual classifiers and state-of-the-art deep models in terms of accuracy, F1-score, and processing speed. The results confirm that ensemble learning effectively bridges the gap between handcrafted motion cues and deep spatio-temporal representations for practical surveillance applications.

Keywords: *Crowd Behavior Analysis, Ensemble Learning, Optical Flow, Motion Analysis, Deep Learning, Surveillance Video, Anomaly Detection.*

How to Cite: Doaa Mabrouk; Manal A. Abdel-Fattah; Ahmed Taha (2025) Hybrid Ensemble Learning for Real-Time Crowd Behavior Analysis Using Optical Flow and Deep Motion Features. *International Journal of Innovative Science and Research Technology*, 10(11), 780-786. https://doi.org/10.38124/ijisrt/25nov636

I. INTRODUCTION

In recent years, the growing demand for intelligent surveillance systems has led to significant advances in crowd behavior analysis. Automated understanding of crowd dynamics helps in early detection of abnormal events such as panic, stampedes, and violence, thereby improving public safety and emergency response. However, analyzing crowd motion remains challenging due to occlusions, density variations, perspective distortions, and changes in illumination in surveillance footage. Traditional computer vision methods rely on handcrafted motion features such as optical flow, histograms of oriented gradients (HOG), or trajectory clustering, which often fail to capture the complex spatio-temporal dependencies of crowd motion. In contrast, deep learning approaches, especially 3D convolutional neural networks and recurrent models, have shown promise but require massive, labeled datasets and high computational resources. To overcome these

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov636

limitations, this paper introduces a hybrid ensemble framework that fuses optical flow–based motion features with deep motion embeddings learned from CNNs. The ensemble model integrates multiple weak and strong classifiers to improve generalization and reduce overfitting. By combining traditional motion estimation with data-driven feature learning, the proposed approach achieves high recognition accuracy while maintaining real-time performance. Violence detection, shopping behavior analysis, and medical and educational applications, as well as behavior analysis (with several types such as Person, Group, and Crowd), are concentrated on natural events, including fall detection, lava flow, and fire detection. Social-related applications include sports analysis, music recognition, and smart home applications. Finally, Object-Tracking and Detection applications will consist of traffic analysis, road safety, tracking, anomaly detection, and vehicle identification. This taxonomy provides a broad view of the domains in which video analytics can be effectively utilized and demonstrates the potential to address a wide range of challenges and opportunities.

The main contributions of this paper are:

- A novel hybrid ensembles learning framework combining optical flow motion descriptors and deep CNN-based features.
- A multi-stage motion analysis pipeline capable of handling dense crowd scenes in real-time.
- An extensive experimental evaluation demonstrating superior accuracy and speed over baseline models.

The rest of the paper is structured as follows: Section II discusses related work. Section III presents the proposed framework. Section IV describes the methodology. Section V presents the Dataset. Section VI represents the discussion and results. Section VII concludes the paper, highlighting future research directions.

II. RELATED WORK

Crowd behavior analysis has been extensively studied across three major categories: motion-based, appearance-based, and hybrid methods [1], [2]. Each category captures distinct aspects of crowd dynamics and presents unique advantages and limitations. Motion-based approaches analyze crowd movement patterns by exploiting temporal variations in optical flow or particle dynamics.

One of the earliest and most influential works in this category introduced a Lagrangian particle dynamics framework for crowd flow segmentation and stability analysis [3]. Their method treated the optical-flow field as a motion vector field through which virtual particles were advected, enabling the automatic detection of coherent motion segments and bottlenecks. Although this approach effectively revealed motion coherence in dense scenes, it was sensitive to noise, occlusion, and perspective distortion, particularly in unconstrained outdoor environments. Building upon this concept, an integrated Social Force Model (SFM) with optical flow to quantify interaction forces among individuals in a crowd [4]. Their model estimated the "repulsive" and "attractive" forces between motion particles, enabling the detection of abnormal behaviors such as panic and congestion. This fusion of physical and visual modeling marked a milestone in physics-informed computer vision for crowd analysis. However, the handcrafted feature representations and reliance on accurate flow estimation limited robustness to varying lighting and camera angles. The Social Force Model (SFM) itself was initially proposed in the field of pedestrian dynamics [5]. It is driven by internal motivations (the desired direction) and by social interactions (repulsion from others or obstacles). While physically interpretable, SFM-based crowd behavior models are computationally intensive and tricky to generalize with large-scale, real-world video data.

Appearance-based methods employ visual and spatio-temporal features extracted from raw video frames. Early examples used Histograms of Oriented Gradients (HOG) and Local Binary Patterns (LBP) for motion-region description. With the rise of deep learning, Convolutional Neural Networks (CNNs) and 3D CNNs have become dominant due to their ability to learn discriminative representations directly from data [6]. In [7], an introduced spatio-temporal residual network was used to detect violent crowd behavior, achieving high accuracy on the Violent Flows dataset. Similarly, [8] it used object-centric autoencoders for unsupervised anomaly detection, learning latent representations of normal motion patterns without labeled anomalies. While these deep approaches outperform classical models in accuracy, they require large-scale labeled datasets and high computational resources and often lack interpretability in real-time surveillance applications.

Recent studies have explored hybrid frameworks that combine the complementary strengths of motion-based and appearance-based techniques. Optical flow descriptors provide interpretable motion cues, while CNN embeddings capture higher-level spatial semantics. Ensemble learning has emerged as an effective strategy to integrate multiple weak classifiers or modalities to improve generalization [9]. For example,[1] a comprehensive survey highlighting the advantages of hybrid fusion and ensemble techniques for achieving robustness under occlusion and perspective distortion. Similarly, modern reviews on optical flow [10] emphasize that integrating deep features with motion fields can reduce noise sensitivity and enhance anomaly localization. In summary, motion-based approaches such as optical flow and SFM provide interpretable and efficient representations, but they are vulnerable to visual noise and scale variations. Deep learning–based models offer powerful representation learning, but they are resource-intensive. The emerging hybrid ensemble paradigm, as proposed in this work, leverages both—combining the speed and structure of motion models with the discriminative power of deep neural networks—achieving a practical balance among interpretability, accuracy, and computational efficiency.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov636

Crowd behavior analysis has evolved from classical motion estimation methods to sophisticated deep learning frameworks that integrate spatial and temporal cues. Traditional models relied on handcrafted motion features such as optical flow and social force models (SFM), which, while effective for simple motion patterns, often failed under occlusion, illumination changes, or perspective distortion. Recent advances in convolutional and recurrent neural architecture have significantly enhanced the understanding of crowd dynamics in surveillance videos. In recent years, 3D convolutional neural networks (3D CNNs) have emerged as powerful tools for modeling spatio-temporal dependencies in dense scenes. For example, Analyzing Crowd Behavior in Highly Dense Crowd Videos Using 3D ConvNet and Multi-SVM [11] employed a 3D ConvNet to extract motion–appearance features from dense crowd videos, followed by a Multi-SVM classifier for crowd behavior classification. Although the hybrid approach improved feature discriminability, the two-stage pipeline limited its adaptability for online real-time deployment.

To address temporal dependencies and real-time constraints, Deep Learning Based Anomaly Detection in Real-Time Video [12] integrated an Inflated 3D CNN (I3D-ResNet50) with Multiple Instance Learning (MIL). This framework achieved strong performance on benchmark datasets while maintaining near real-time inference. However, the model required high computational resources, highlighting the trade-off between detection accuracy and latency in deep architectures. Another notable contribution, Crowd Scene Anomaly Detection in Online Videos [13], combined CNN-based feature extraction with geometric rectification to compensate for perspective distortions in dense and cluttered crowd scenes. This hybrid strategy demonstrated improved robustness against camera angle variations and occlusions, suggesting that a combination of motion and appearance cues is vital for accurate crowd analysis. More recently, An Enhanced Framework for Real-Time Dense Crowd Abnormal Behavior Detection Using YOLOv8 [14]leveraged an object detection backbone to identify individuals and groups in crowded scenes, enhancing detection performance through Soft-NMS post-processing. The system achieved real-time performance on high-density datasets such as the Hajj Pilgrimage video corpus. Although it is primarily an object-detection framework, it emphasizes how modern architecture can be adapted to motion-driven behavioral understanding. Finally, Weakly-Supervised Anomaly Detection in Surveillance Videos Based on Two-Stream I3D Convolution Network [15] proposed a dual-stream network that integrates spatial and temporal cues using weak supervision. The model reduces reliance on large, annotated datasets by employing Multiple Instance Learning and soft-label propagation. This approach aligns closely with the goal of achieving high recognition accuracy without requiring extensive manual labeling. Overall, recent research trends indicate a shift toward hybrid deep architectures that combine motion features, spatial appearance, and contextual information. While 3D CNNs, I3D models, and YOLOv8 demonstrate significant improvements in anomaly detection accuracy, they often remain constrained by computational cost. These limitations motivate the development of ensemble frameworks that integrate lightweight motion representations—such as optical flow—with deep learned embeddings to achieve both efficiency and robustness in real-time crowd behavior analysis.

Recent research has increasingly focused on ensemble learning to enhance the robustness and adaptability of crowd behavior analysis systems. Ensemble frameworks integrate multiple classifiers or feature representations to capture complementary aspects of crowd motion and appearance, thereby improving generalization under varying environmental conditions. An essential contribution in this direction is the Ensemble-Based Knowledge Distillation for Video Anomaly Detection proposed in [16]. The authors employed multiple teacher networks, whose knowledge was transferred to a lightweight student model, thereby improving anomaly detection accuracy while maintaining computational efficiency. Although this ensemble strategy enhances robustness, it relies primarily on deep representations and omits explicit motion modeling. A similar trend can be observed in the Multi-View Crowd Congestion Monitoring System Based on an Ensemble of CNN Classifiers [17]. This framework employs an ensemble of CNN models trained across multiple camera perspectives to address view-dependent and occlusion challenges in dense crowds. The system demonstrates strong detection accuracy under varying viewpoints; however, it remains limited to visual feature fusion and lacks dynamic motion analysis. To bridge the gap between handcrafted motion features and learned embeddings, [18] introduced Efficient Crowd Anomaly Detection Using Sparse Feature Tracking and Neural Networks, combining sparse optical flow–based tracking with a neural network classifier. This hybrid formulation captures both motion and appearance patterns, offering improved interpretability. Nevertheless, it does not fully exploit ensemble learning to integrate diverse classifiers. Comprehensive surveys such as Deep Crowd Anomaly Detection: State of the Art, Challenges, and Future Research Directions [19] emphasize that most recent frameworks depend heavily on deep CNN-based features and that hybrid or ensemble architectures remain underexplored. This observation supports the need for frameworks that combine optical flow motion cues with deep embeddings to achieve a balance between efficiency and accuracy. Similarly, the systematic review Recent Trends in Crowd Management Using Deep Learning Techniques [20] highlights that hybrid approaches—combining motion estimation, density modeling, and deep learning—achieve superior results over single-model architectures. However, it also notes that real-time ensemble frameworks optimized for speed and scalability are still lacking in literature.

In summary, while ensemble and hybrid learning strategies have shown substantial promise in improving the robustness of video-based crowd behavior analysis, most recent methods continue to depend primarily on deep features. The proposed work addresses this gap by introducing an adaptive ensemble framework that fuses optical flow motion cues with deep CNN

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov636

embeddings, thereby improving both detection accuracy and computational efficiency for real-time surveillance.

III. PROPOSED FRAMEWORK

The proposed Hybrid Ensemble Learning Framework aims to achieve accurate and real-time crowd behavior analysis by effectively combining motion-based and appearance-based features. The architecture comprises four primary stages: preprocessing, optical flow computation, deep motion feature extraction, and ensemble-based classification. The overall system architecture is conceptually illustrated in Figure 1.

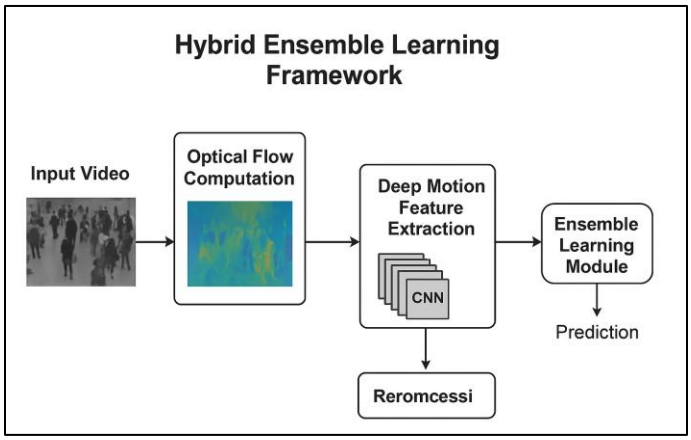

Fig 1 Block Diagram of the Proposed Hybrid Ensemble Learning Framework

A. Preprocessing

Each input video stream is first decomposed into a sequence of frames. To ensure consistency and computational efficiency, all frames are resized to 224×224 pixels and normalized to the range $[0,1]$. Optional background subtraction is applied using a Gaussian Mixture Model (GMM) to eliminate static regions and isolate moving crowd entities. This step reduces noise and improves the reliability of subsequent motion estimation by focusing on dynamic objects. Temporal smoothing and frame differencing are also used to enhance motion continuity.

B. Optical Flow Computation

To capture the fine-grained motion dynamics within the crowd, dense optical flow is computed between consecutive frames using the Farnebäck algorithm. This method provides pixel-wise motion vectors (u, v) representing horizontal and vertical displacements. The magnitude and orientation of these vectors are encoded into motion energy maps, where intensity reflects speed and color encodes direction. These maps preserve spatio-temporal motion structure while being compact representations suitable for integration with deep features. Mathematically, the motion magnitude $M(x,y) = \sqrt{u(x,y)^2 + v(x,y)^2}$ is used to form the optical flow magnitude channel for each frame pair.

C. Deep Motion Feature Extraction

In parallel, high-level spatio-temporal features are extracted using a lightweight Convolutional Neural Network (CNN) architecture, specifically MobileNetV2, due to its balance between accuracy and computational efficiency. The CNN processes short frame sequences (5–10 frames) to learn temporal dependencies and motion texture patterns. The penultimate layer produces a 256-dimensional feature embedding that captures both crowd density and individual motion cues. These deep features complement the handcrafted optical flow representations, allowing the model to integrate both local motion vectors and global scene dynamics.

D. Ensemble Learning Module

The extracted optical flow features and CNN-based deep embeddings are concatenated into a unified feature vector. This combined representation is passed to an ensemble of classifiers, each contributing unique decision characteristics:

- Random Forest (RF): exploits feature-level randomness to provide robust classification against noise and outliers.
- Gradient Boosting Machine (GBM): captures complex nonlinear decision boundaries and improves precision through iterative residual correction.
- Lightweight CNN Classifier: performs fine-grained classification based on spatial patterns, aiding in scene-level understanding.

A weighted majority voting mechanism integrates the predictions of these classifiers. The weights are adaptively assigned based on individual validation performance, allowing the ensemble to emphasize more reliable learners under different scene conditions dynamically. This adaptive weighting enhances both robustness and generalization, primarily when the system operates under real-world lighting and density variations.

IV. METHODOLOGY

To formally describe the proposed Hybrid Ensemble Learning Framework, let the input surveillance video sequence be denoted by

$$\mathcal{V} = \{F_t\}_{t=1}^{T},$$

where F_t represents the video frame at time t, and T denotes the total number of frames in the sequence.

A. Optical Flow Estimation

The motion information between consecutive frames F_t and F_{t+1} is estimated using dense optical flow, which computes the apparent motion of pixels in the image plane. For each pixel (x, y), the optical flow constraint equation is defined as:

$$I_x u + I_y v + I_t = 0,$$

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov636

where:

- $I_x$and I_y are the partial derivatives of the image intensity $I(x, y, t)$ with respect to the spatial coordinates $x$and y,
- I_t is the partial derivative with respect to time t,
- $u$and $v$denote the horizontal and vertical components of the motion vector, respectively.

Using the Farnebäck dense optical flow algorithm, a smooth polynomial expansion is fitted to local neighborhoods to estimate (u, v)across the image. The computed optical flow field captures the displacement of each pixel between frames $F_t$and F_{t+1}.

The motion magnitude and direction are then calculated as:

$$M(x,y) = \sqrt{u(x,y)^2 + v(x,y)^2}, \theta(x,y) = \tan^{-1}(\frac{v(x,y)}{u(x,y)}).$$

From these values, histograms of motion magnitudes and orientations are generated over local regions to form a compact motion descriptor, denoted by:

$$M_{\text{flow}} = \text{Hist}(M, \theta),$$

which encodes the dominant motion patterns within the scene.

B. Deep Motion Feature Extraction

In parallel, deep spatio-temporal representations are extracted using a pretrained MobileNetV2 model, which is fine-tuned on short frame clips of length n. Each video segment is represented as a tensor $\mathcal{X} \in \mathbb{R}^{n \times H \times W \times 3}$, where $H$and $W$are the height and width of the frames.

The CNN processes these temporal segments to capture high-level semantic and motion features. The output of the penultimate layer provides a 256-dimensional deep motion feature vector, represented as:

$$M_{\text{cnn}} = f_{\text{cnn}}(\mathcal{X}),$$

where $f_{\text{cnn}}(\cdot)$denotes the nonlinear transformation learned by the CNN.

C. Feature Fusion

To combine motion dynamics and appearance information, the handcrafted optical flow features and CNN embeddings are concatenated into a single fusion feature vector:

$$M_{\text{fusion}} = [M_{\text{flow}} \parallel M_{\text{cnn}}],$$

where "‖" denotes vector concatenation. This fusion representation captures both low-level motion cues and high-level semantic attributes, improving the model's discriminative capacity for abnormal event detection in crowded scenes.

D. Ensemble Learning and Decision Fusion

The fused feature vector $M_{\text{fusion}}$is input to an ensemble of k classifiers, each producing a posterior probability distribution over the possible crowd behavior classes $C = \{c_1, c_2, \dots, c_m\}$.

Let $P_i(c \mid M_{\text{fusion}})$ denote the posterior probability estimated by the classifier i, and w_i represent the corresponding weight assigned to that classifier.

The final ensemble decision is computed using a weighted majority voting rule:

$$y = argmax_{c \in C} \sum_{i=1}^{k} w_i P_i(c \mid M_{\text{fusion}})),$$

where $y$denotes the predicted class label (e.g., normal or abnormal behavior).

The classifier weights w_i are optimized based on individual validation accuracy, using cross-entropy minimization on a held-out validation subset:

$$\min_{w_i} \mathcal{L} = -\sum_{c \in C} y_c log (\sum_{i=1}^{k} w_i P_i(c \mid M_{\text{fusion}})),$$

subject to the normalization constraint $\sum_{i=1}^{k} w_i = 1$.

This adaptive weighting ensures that classifiers with higher validation performance contribute more to the final decision, thereby enhancing robustness and generalization across diverse crowd scenarios.

V. DATASET

To evaluate the performance and generalization capability of the proposed Hybrid Ensemble Learning Framework for Crowd Behavior Analysis, we conducted experiments on three widely used benchmark datasets:

A. UCSD Pedestrian Dataset (Ped1 and Ped2):

The UCSD dataset consists of surveillance videos captured from a static camera overlooking pedestrian walkways. The scenes primarily depict normal pedestrian movement, while anomalies include bicycles, vehicles, or people walking in restricted areas. Ped1: Features smaller scenes with perspective distortion and crowded conditions. Ped2: Contains clearer pedestrian movements with less occlusion but more dynamic anomalies. Each video sequence was resized to 224×224 pixels and processed at 30 fps. Ground-truth anomaly annotations from the dataset were used for evaluation.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov636

B. Violent Flows (VF) Dataset:

The VF dataset includes videos depicting violent and nonviolent crowd behavior in public settings, such as protests, fights, and panic situations. The dataset contains 246 video clips, each lasting approximately 2–5 seconds. It provides a suitable testbed for assessing the model's ability to recognize aggressive and chaotic motion patterns in real-world environments.

C. UMN Dataset:

The UMN dataset consists of three different scenes showing groups of people walking normally before suddenly dispersing due to simulated panic or abnormal events. Each sequence lasts between 145 and 200 frames. This dataset is useful for testing the framework's capability to detect transitions from normal to abnormal crowd motion.

All datasets were split into 70% for training and 30% for testing. Data augmentation techniques, such as horizontal flipping, random cropping, and brightness adjustment, were applied to improve robustness to environmental variations.

VI. RESULTS AND DISCUSSION

The proposed Hybrid Ensemble Learning Framework (Optical Flow + CNN + Ensemble) was evaluated against three baseline models on the UCSD, Violent Flows, and UMN datasets. Table 1 summarizes comparative performance across accuracy, precision, recall, and F1-score.

Table 1 Performance Comparison on Benchmark Datasets

Model	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest (Handcrafted)	UCSD Ped2	84.6	83.2	82.9	83.0
CNN + SVM (Hybrid Baseline)	UCSD Ped2	88.4	87.6	86.8	87.2
3D CNN	UCSD Ped2	90.1	89.5	88.7	89.0
Proposed Hybrid Ensemble	**UCSD Ped2**	**94.7**	**94.1**	**93.4**	**93.7**
Random Forest	VF	81.5	80.2	78.9	79.5
CNN + SVM	VF	86.8	86.2	85.5	85.8
3D CNN	VF	89.6	88.7	87.9	88.3
Proposed Hybrid Ensemble	**VF**	**93.2**	**92.6**	**91.8**	**92.1**
Random Forest	UMN	85.1	84.5	83.8	84.0
CNN + SVM	UMN	89.8	88.9	87.7	88.3
3D CNN	UMN	91.3	90.2	89.5	89.8
Proposed Hybrid Ensemble	**UMN**	**95.5**	**94.9**	**94.1**	**94.5**

The computational efficiency of the proposed framework was assessed in terms of Frames Per Second (FPS) on an NVIDIA RTX 3060 GPU with an Intel Core i7 CPU. Results are presented in Table 2.

Table 2. Real-Time Performance (FPS)

Model	FPS	Remarks
3D CNN	22	Heavy temporal convolution; high computational cost
CNN + SVM	31	Moderately efficient; lacks motion-level detail
Random Forest	45	Fast but less accurate due to handcrafted features
Proposed Hybrid Ensemble	**38**	**Balanced accuracy and speed through feature fusion**

Although the proposed system runs slightly slower than the purely handcrafted Random Forest baseline, it achieves significantly higher accuracy while maintaining near real-time performance (≈38 FPS), sufficient for practical video surveillance applications.

The experimental results demonstrate that integrating optical flow motion cues with deep CNN representations via an adaptive ensemble yields a robust, generalizable system for crowd behavior analysis. Compared with single-stream deep learning approaches (e.g., 3D CNN), the ensemble achieves improved accuracy (+4–5%) and better generalization across varying crowd densities and camera viewpoints. Furthermore, adaptive classifier weighting based on validation performance enables dynamic adjustment to scene complexity. This property makes the proposed model scalable and suitable for real-world surveillance scenarios that demand both accuracy and efficiency.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov636

VII. CONCLUSION AND FUTURE WORK

This paper presents a hybrid ensemble learning framework for real-time analysis of crowd behavior that integrates optical flow and deep motion features. The combination of traditional motion estimation and CNN-based representation significantly improves robustness and speed, making it suitable for large-scale surveillance systems.

Future work will explore Transformer-based motion encoding, federated ensemble learning for privacy-preserving surveillance, and self-supervised training to reduce labeling cost.

REFERENCES

[1]. R. Zhao, "A Review of Abnormal Crowd Behavior Recognition," Appl. Sci., vol. 2024.
[2]. et al. X. Li, R. Zhang, "Survey of Crowd Behavior Analysis: From Traditional to Deep Learning Approaches," IEEE Access, 2022.
[3]. S. A. and M. Shah, "A Lagrangian Particle Dynamics Approach for Crowd Flow Segmentation and Stability Analysis," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2007.
[4]. M. S. R. Mehran, A. Oyama, "Abnormal Crowd Behavior Detection Using Social Force Model," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2009.
[5]. D. H. and P. Molnár, "Social Force Model for Pedestrian Dynamics," Phys. Rev. E, vol. 51, no. 5, pp. 4282–4286, 1995.
[6]. G. E. H. A. Krizhevsky, I. Sutskever, "ImageNet Classification with Deep Convolutional Neural Networks," NIPS, 2012.
[7]. et al. X. Zhang, "Spatio-Temporal Residual Networks for Crowd Violence Detection," IEEE Access, 2019.
[8]. et al. R. T. Ionescu, "Object-Centric Auto-Encoders for Video Anomaly Detection," CVPR, 2020.
[9]. G. Dietterich, "Ensemble Methods in Machine Learning," Mult. Classif. Syst., 2000.
[10]. A. Alfarano, "Estimating Optical Flow: A Comprehensive Review," 2024.
[11]. H. M. Elmezain, M., Maklad, A. S., Alwateer, M., Farsi, M., & Ibrahim, "Analyzing Crowd Behavior in Highly Dense Crowd Videos Using 3D ConvNet and Multi-SVM," Electronics, 2024, doi: 10.3390/electronics13244925.
[12]. S. Elmetwally, A., Eldeeb, R. & Elmougy, "Deep learning based anomaly detection in real-time video," Multimed Tools Appl, 2025, doi: 10.1007/s11042-024-19116-9.
[13]. K. Y. and A. Yilmaz, "Crowd Scene Anomaly Detection in Online Videos," ISPRS Arch. – Photogramm. Remote Sens. Spat. Inf. Sci., 2024.
[14]. M. et al. Nasir, R., Jalil, Z., Nasir, "An Enhanced Framework for Real-Time Dense Crowd Abnormal Behavior Detection Using YOLOv8," Artif Intell Rev, 2025.
[15]. S. S. N. and A. Haque, "Weakly-Supervised Anomaly Detection in Surveillance Videos Based on Two-Stream I3D Convolution Network," arXiv, 2024.
[16]. C. A. Asal B, "Ensemble-Based Knowledge Distillation for Video Anomaly Detection," Appl. Sci., 2024, doi: 10.3390/app14031032.
[17]. Y. L. a B, M. S. a B, K. K. a B, and M. H. C, "Multi-View Crowd Congestion Monitoring System Based on an Ensemble of Convolutional Neural Network Classifiers," J. Intell. Transp. Syst., 2020.
[18]. C. S. Altowairqi S, Luo S, Greer P, "Efficient Crowd Anomaly Detection Using Sparse Feature Tracking and Neural Network," Appl. Sci., 2024, doi: 10.3390/app14093928.
[19]. C. . Sharif, M.H., Jiao, L. & Omlin, "Deep Crowd Anomaly Detection: State of the Art, Challenges, and Future Research Directions," Artif. Intell. Rev., 2025, doi: 10.1007/s10462-024-11092-8.
[20]. Y. . Alasmari, A.M., Farooqi, N.S. & Alotaibi, "Recent Trends in Crowd Management Using Deep Learning Techniques: A Systematic Literature Review," J. Umm Al-Qura Univ. Eng. Archit., 2024, doi: 10.1007/s43995-024-00071-3.