
A Real-Time Smart Surveillance Framework Using Edge-IoT for Suspicious Activity Detection and Facial Emotion Analysis

Nidhi P. Nandankar, Shraddha R. Khonde

Department of Computer Engineering, Modern Education Society's Wadia College of Engineering, Pune- 411001, nidhinandankar25@gmail.com, shraddha.khonde@mescoepune.org

Abstract.

The growing need for intelligent, proactive surveillance has exposed the limitations of traditional CCTV systems, which rely heavily on continuous human monitoring and post-incident analysis. To address these challenges, this work presents an Edge-IoT-based real-time smart surveillance framework that integrates human activity recognition with facial emotion analysis using deep learning. The proposed system performs on-device processing on a Raspberry Pi, enabling local video analysis with low latency while preserving user privacy. Facial regions are detected using a lightweight Haar Cascade approach, followed by emotion classification with a CNN trained on the FER-2013 dataset. In parallel, human activities such as walking, running, and aggressive movements are analyzed using motion-based features. A decision-level fusion strategy integrates activity patterns with emotional cues such as fear and anger to accurately distinguish normal from suspicious behavior, thereby minimizing false alarms. Experimental results demonstrate reliable real-time performance on edge hardware, confirming the system's suitability for smart surveillance applications in public and institutional environments.

Keywords: Smart Surveillance; Facial Emotion Recognition; Suspicious Activity Detection; Convolutional Neural Network; IoT

1. INTRODUCTION

Traditional CCTV requires continuous human monitoring and has limited real-time intelligence [1]. Deep learning improves activity detection but often requires high computation or cloud processing, causing latency and privacy concerns [2–4]. Motion-based methods may produce false alarms, while emotion recognition is rarely integrated with activity detection [5–7]. This paper proposes a Raspberry Pi-based edge-IoT system combining activity and emotion recognition for accurate, private, and low-latency surveillance.

2. LITERATURE REVIEW

Existing methods provide good detection accuracy but face challenges such as high computational cost, false alarms, limited contextual awareness, and cloud dependency [10–18]. The key research gap is the lack of an integrated, real-time, privacy-preserving edge system combining activity and emotion recognition.

3. PROPOSED SYSTEM ARCHITECTURE AND METHODOLOGY

The suggested framework adheres to an Edge-IoT-enabled intelligent surveillance paradigm capable of detecting human behaviors through joint analysis of both human actions and human facial emotions. The suggested approach differs significantly from traditional CCTV camera systems, which depend heavily on nonstop video recording and subsequent human supervision of the recorded videos

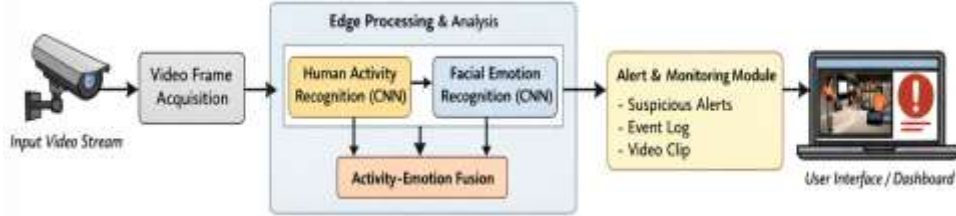


Figure 3.1. Architecture of the Proposed Edge-IoT system

Overall Framework

As shown in Fig.3.1 the architecture is organized into four functional layers:

- Video Acquisition Layer,
- Edge Processing and Feature Extraction Layer,
- Activity–Emotion Fusion and Decision Layer, and
- Alert and Monitoring Layer.

A Raspberry Pi is used as the main edge device, connected to the surveillance camera that will record video streams and conduct inference in real time.

Methodology

The concept of intelligent surveillance systems by leveraging Raspberry Pi with edge computing is introduced, employing the decision-level fusion mechanism of activity and facial emotion recognition using CNN-based architecture. This will enable real-time and efficient analysis without relying on cloud connectivity. The overall workflow of the proposed system consists of the following stages:

1. Video frame acquisition and preprocessing on Raspberry Pi
2. CNN-based human activity feature extraction
3. CNN-based facial emotion recognition
4. Decision-level activity–emotion fusion
5. Final suspicious activity decision and alert generation

3.2.1 Video Preprocessing and Frame Sequencing

Let the continuous video stream captured by the Raspberry Pi camera module or USB camera be represented as: $V = \{F_1, F_2, \dots, F_n\}$ (1) where F_i denotes the i^{th} video frame. All captured video frames are resized and normalized to a size compatible with "lightweight" CNAs; that is, each individual frame will be processed in real-time by a Raspberry-Pi operating system when resized to a single height and width measurement (e.g., 1280x720). The final sliding window is then constructed by "sliding" that window over those k consecutive video frames:

$$S_t = \{F_t, F_{(t+1)}, \dots, F_{(t+k)}\} \quad (2)$$

Using sliding windows helps the model detect motion patterns over shorter periods of time, which allows for greater accuracy when detecting abnormal motion patterns based on prior observations alone. Such simplified temporal modeling is computationally efficient and well-suited for edge devices, while still offering meaningful activity context compared to more complex temporal learning frameworks

3.2.2 CNN-Based Human Activity Recognition

The human activity recognition unit employs an efficient CNN architecture for extracting motion information of spatial nature from the frame sequence S_t by means of the convolutional and pooling layers.

$$X_l = f(W_l * X_{l-1} + b_l). \quad (3)$$

(3) where W_l and b_l stand for the convolution and bias parameters of the l -th layer, $*$ is a convolution operator, and $f(\cdot)$ is an activation function. The extracted features are used by the softmax classifier to assign an activity label from the list including, e.g., walking, running, fighting, and abrupt motion:

$$P(A_i | S_t) = \frac{e^{z_i}}{\sum_{j=1}^C e^{z_j}} \quad (4)$$

where C refers to the total number of activity classes. Activities such as running or fast walking are considered ambiguous rather than suspicious in order to avoid false positives

3.2.3 CNN-Based Facial Emotion Recognition

Facial emotion recognition operates in parallel with activity detection. Face regions are first detected from video frames using a Haar Cascade or lightweight CNN-based face detector suitable for Raspberry Pi deployment. Let the extracted facial image be denoted as I_f .

A CNN-based emotion classifier predicts emotional states:

$$E = \{\text{Neutral}, \text{Happy}, \text{Anger}, \text{Fear}\} \quad (5)$$

Emotion probabilities are computed using the softmax function:

$$P(E_k | I_f) = \frac{e^{y_k}}{\sum_{m=1}^M e^{y_m}} \quad (6)$$

where M represents the number of emotion classes. Facial emotion recognition provides semantic cues related to **human intent**, which cannot be reliably inferred from motion patterns alone.

3.2.4 Activity–Emotion Decision-Level Fusion Strategy

To reduce false alarms and improve reliability, a rule-based decision-level fusion approach combines activity ((A)) and emotion ((E)) predictions:

- Normal activity + Neutral/Happy: Non-Suspicious
- Ambiguous activity + Anger/Fear: Suspicious
- Abnormal activity: Suspicious

This approach uses emotional context to interpret ambiguous activities, improving detection reliability and reducing false positives.

4. RESULTS

a) GUI and Edge Device: Tkinter is used as our GUI in the Raspberry Pi such that we can do live monitoring

control of the camera as well as receive real-time notifications using the buzzer.

b) Facial Emotion Recognition: Haar Cascade is the face detector tool used by our system, followed by a special computer software called CNN that identifies emotions on the face. It can identify seven emotions and represents them through different color representations.

c) Human Activity Detection: MediaPipe Pose generates landmark features for the classification of human

activities as either normal or suspicious with the labeled outputs.

d) Activity–Emotion Fusion & Alerts: There are specific rules set for combining activity and emotion to generate notifications as follows:

- Running + neutral/happy = Normal
- Running + fear/anger = Suspicious
- Fighting + all emotions = Suspicious

Detection of suspicious behavior initiates buzz alerts accompanied by red labels. Screen captures may also be taken to log any incidents.

e) Advantages of the Edge: All computation is done locally on the Raspberry Pi without the need for cloud services and hence gives real-time results at ~6-8 fps. This also minimizes false positives.

Output Visualization



Figure 4.1. GUI Home Screen (with Detection). Figure 4.2 Real-time Emotion Detection Output.

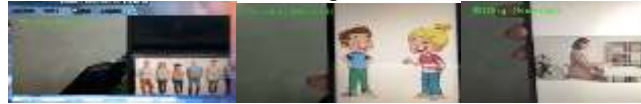


Figure 4.2. Normal Activity Detection (Sitting/Standing)

5. CONCLUSION

This study presents an Edge-IoT surveillance system for real-time human activity, pose, and facial emotion recognition. Using a Raspberry Pi, the system enables fast, privacy-preserving, and independent edge-based decision-making. The proposed framework achieves 96% accuracy, reduces false alarms, and supports efficient surveillance in environments such as schools and public spaces.

6. REFERENCES

- [1] M. M. Zaidi et al., "Suspicious human activity recognition from surveillance videos using deep learning," *IEEE Access*, vol. 12, pp. 105497–105510, 2024, doi: 10.1109/ACCESS.2024.3436653.
- [2] M. Biswas et al., "State-of-the-art violence detection techniques: A review," *Asian Journal of Computer Science and Information Technology*, vol. 13, pp. 29–42, 2022, doi: 10.9734/AJRCOS/2022/v13i130305.
- [3] P. Kuppusamy and V. C. Bharathi, "Human abnormal behaviour detection using CNNs in crowded and uncrowded surveillance – A survey," *Measurement: Sensors*, vol. 24, p. 100510, 2022, doi: 10.1016/j.measen.2022.100510.
- [4] M. Khanam, M. Sharif, X. Cheng and S. Kadry, "Suspicious action recognition in surveillance based on handcrafted and deep learning methods: A survey of the state of the art," *Computers & Electrical Engineering*, vol. 120, p. 109811, 2024.
- [5] K. U. Duja, I. A. Khan and M. Alsuhailani, "Video surveillance anomaly detection: A review on deep learning benchmarks," *IEEE Access*, vol. 12, pp. 164811–164842, 2024, doi: 10.1109/ACCESS.2024.3491868.
- [6] T. Saba et al., "Suspicious activity recognition using proposed deep L4-branched-actionnet with entropy coded ant colony system optimization," *IEEE Access*, vol. 9, pp. 89181–89197, 2021, doi: 10.1109/ACCESS.2021.3091081.
- [7] T. Rehman et al., "Internet-of-things-based suspicious activity recognition using multimodalities of computer vision for smart city security," *Security and Communication Networks*, vol. 2022, Art. no. 8383461, 2022, doi: 10.1155/2022/8383461.
- [8] O. Saidani et al., "An efficient human activity recognition using hybrid features and transformer model," *IEEE Access*, vol. 11, pp. 101373–101386, 2023, doi: 10.1109/ACCESS.2023.3314492.

- [9] N. Dwivedi, D. Singh and D. Kushwaha, "A novel approach for suspicious activity detection with deep learning," *Multimedia Tools and Applications*, vol. 82, pp. 1–24, 2023, doi: 10.1007/s11042-023-14445-7.
- [10] W. Ullah et al., "CNN features with bidirectional LSTM for real-time anomaly detection in surveillance networks," *Multimedia Tools and Applications*, vol. 80, no. 11, pp. 16979–16995, 2021.
- [11] R. Vrskova et al., "A new approach for abnormal human activities recognition based on ConvLSTM architecture," *Sensors*, vol. 22, no. 8, p. 2946, 2022, doi: 10.3390/s22082946.
- [12] M. Q. Gandapur and E. Verdú, "ConvGRU-CNN: Spatiotemporal deep learning for real-world anomaly detection in video surveillance system," *Int. J. Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 4, p. 88, 2023.
- [13] Q. U. A. Arshad et al., "Anomalous situations recognition in surveillance images using deep learning," *Computers, Materials & Continua*, vol. 76, no. 1, pp. 1103–1125, 2023.
- [14] R. Rajeswari, "Anomalous human activity recognition from video sequences using BRISK features and convolutional neural networks," *Galaxy International Interdisciplinary Research Journal*, vol. 10, no. 2, pp. 269–280, 2022. U. Khan, S. Afzal and J.
- [15] W. Lee, "Human activity recognition via hybrid deep learning-based model," *Sensors*, vol. 22, no. 1, p. 323, 2022, doi: 10.3390/s22010323.
- [16] R. R. Subramanian and V. Vasudevan, "A deep genetic algorithm for human activity recognition leveraging fog computing frameworks," *Journal of Visual Communication and Image Representation*, vol. 77, p. 103132, 2021, doi: 10.1016/j.jvcir.2021.103132.
- [17] Z. Xiao et al., "A federated learning system with enhanced feature extraction for human activity recognition," *Knowledge-Based Systems*, vol. 229, p. 107338, 2021, doi: 10.1016/j.knosys.2021.107338.
- [18] Jia et al., "Multi-frequency and multi-domain human activity recognition based on SFCW radar using deep learning," *Neurocomputing*, vol. 444, pp. 274–287, 2021, doi: 10.1016/j.neucom.2020.07.136.

Biographies



Nidhi P. Nandankar received the bachelor's degree in information technology from Savitribai Phule Pune University in 2022, and the master's degree in computer engineering from Savitribai Phule Pune University in 2026. She is currently working as an Assistant Professor in the Department of Computer Engineering at Modern Education Society's Wadia College of Engineering, Pune- 411001. Her research interests include Internet of Things , Machine Learning.



Dr. S.R. Khonde is an Associate Professor in Computer Engineering at M.E.S. Wadia College of Engineering, Pune, with over 20 years of teaching experience. She holds a Ph.D. in Computer Science and Engineering and has published 45+ journal and 60+ conference papers, along with two textbooks. Her research interests include Data Mining, Cybersecurity, Network Security, and Blockchain Technologies.