International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 3 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

Automatic Recognition of Suspicious Human Activities from Surveillance Videos Using Deep Learning

Dr. Abdul Khadeer, [1]Associate Professor, Department of Computer Science and Engineering,

Deccan College of Engineering and Technology, Hyderabad, India

Sumayyah Sadat,[2]Mtech Student, Department of Computer Science and Engineering,

Deccan College of Engineering and Technology, Hyderabad, India

ABSTRACT -- The increasing deployment of surveillance cameras in public and security-sensitive environments has resulted in the continuous generation of video data that is difficult to examine manually at scale. Conventional surveillance largely depends on human operators to observe camera feeds and identify events requiring attention, which can become inefficient because of prolonged monitoring, multiple video sources, and the complexity of human activities. This paper presents an automated deep-learning-based framework for recognizing human activities from surveillance videos. The proposed system uses OpenCV for video acquisition, frame extraction, and preprocessing, followed by a hybrid Convolutional Neural Network–Long Short-Term Memory (CNN-LSTM) model for activity classification. The CNN component extracts spatial characteristics from individual video frames, while the LSTM component learns temporal relationships across successive frames. The implemented system considers six activity categories: Normal, Fighting, Fire, Burglary, Accident, and Shooting. Following classification, the predicted activity and associated confidence score are presented to the user, and an email notification can be generated when the detected activity requires an alert. The experimental evaluation reported for the developed framework achieves an accuracy of 96.50%, precision of 96.04%, recall of 97.00%, and F1-score of 96.52%. These results indicate that the CNN-LSTM approach can effectively learn spatial and temporal patterns from the considered surveillance-video data. The proposed framework provides an automated approach to surveillance activity analysis and can reduce dependence on continuous manual observation while supporting faster identification of potentially suspicious events.

Index Terms— Suspicious Activity Recognition, Intelligent Surveillance, Deep Learning, Computer Vision, Convolutional Neural Network, Long Short-Term Memory, CNN-LSTM, Video Classification, Human Activity Recognition, Surveillance Video Analysis, OpenCV, Automated Alert System.

1. INTRODUCTION

1.1 Background

Closed-Circuit Television (CCTV) and other video-surveillance systems are widely used for security monitoring in locations such as public areas, commercial facilities, educational institutions, transportation environments, hospitals, banks, and industrial premises. These systems continuously record visual information that can be used for monitoring and subsequent investigation. However, the increasing number of surveillance cameras also increases the volume of video that security personnel must examine. Continuous observation of multiple feeds can become difficult over extended periods and may result in delayed recognition of important events or failure to notice activities requiring immediate attention.

Traditional surveillance approaches generally depend on human observation or simple motion-

based techniques. Although such methods can provide basic monitoring functionality, they have limited capability to distinguish between different human activities from video sequences. Movement in a surveillance scene does not necessarily indicate a security threat, and determining the meaning of an action often requires analysis of both the visual

characteristics of individual frames and the changes occurring over time.

Recent developments in artificial intelligence and deep learning have created opportunities for automated video analysis. Deep-learning models can learn representations directly from visual data and have been applied to image classification, object recognition, and human activity recognition. For surveillance applications, however, analyzing a single frame is insufficient for many activities because the identity of an action is often determined by the sequence and progression of movements across multiple frames.

The proposed project, **"Automatic Recognition of Suspicious Human Activities from Surveillance Videos Using Deep Learning,"** addresses this problem through a CNN-LSTM-based activity recognition framework. OpenCV is employed to read surveillance videos, extract frames, and perform preprocessing operations. The processed frames are supplied to a CNN for spatial feature extraction, after which an LSTM analyzes the resulting sequence to capture temporal relationships. The combined model classifies the surveillance input into six predefined categories: Normal, Fighting, Fire, Burglary, Accident, and Shooting.

The system additionally displays the predicted activity with a confidence score and supports email notification when an identified activity requires an alert. This combination of automated classification and notification provides a practical mechanism for reducing reliance on continuous manual surveillance.

1.2 Research Gap

Existing surveillance systems provide continuous video recording, but recording footage alone does not provide automated understanding of the activities occurring within a scene. Conventional monitoring approaches generally require security personnel to inspect camera feeds and identify suspicious events manually. When several cameras operate simultaneously, maintaining continuous attention across all feeds becomes difficult. Basic motion-detection techniques also provide limited information about the type or context of human activity.

Research in deep-learning-based video analysis has demonstrated that automated activity recognition can improve surveillance monitoring. CNN-based approaches are effective for learning visual characteristics from individual frames, while recurrent architectures such as LSTM can model information distributed across temporal sequences. Your literature survey identifies CNN-LSTM approaches for surveillance activity classification, autoencoder-based anomaly detection, violence detection, pose-based recognition, and deep-neural-network-based intelligent surveillance as relevant existing directions.

However, several challenges remain. Deep-learning surveillance systems generally require suitable labelled video data and sufficient computational resources. Activity recognition can also become difficult when scenes contain multiple people, varying backgrounds, changes in illumination, or complex movements. Pose-based methods can be affected by occlusion and complicated backgrounds, while anomaly-detection approaches may produce incorrect detections when environmental conditions differ from the training data.

Another issue is the distinction between **spatial and temporal information**. A single video frame can provide information about objects, shapes, and visual patterns, but it may not adequately represent how an activity develops over time. A surveillance activity such as fighting or an accident may require several consecutive frames to distinguish it from an ordinary scene. Therefore, a framework capable of jointly representing frame-level visual information and temporal sequence information is appropriate for this problem.

This research addresses these challenges through a CNN-LSTM-based surveillance framework. The proposed system combines OpenCV-based video processing, CNN spatial feature extraction, LSTM temporal modeling, activity classification, confidence estimation, and automated email notification within a single processing workflow. The implementation focuses on six predefined surveillance activity categories and evaluates the resulting classification performance using standard classification measures.

1.3 Research Objectives

The major objectives of the proposed research are:

1. **To develop an automated surveillance-video analysis system** capable of

recognizing predefined human activities without requiring continuous manual examination of video footage.

2. **To implement OpenCV-based video processing** for reading surveillance videos, extracting frames, and preparing the visual data for deep-learning analysis.
3. **To develop a CNN-based spatial feature extraction mechanism** capable of learning relevant visual characteristics from individual surveillance frames.
4. **To incorporate an LSTM-based temporal modeling mechanism** for learning relationships between consecutive video frames.
5. **To integrate CNN and LSTM models** into a unified architecture for surveillance activity classification.
6. **To classify surveillance videos into six predefined activity categories:** Normal, Fighting, Fire, Burglary, Accident, and Shooting.
7. **To provide the predicted activity together with a confidence score** so that the classification result can be interpreted by the user.
8. **To implement an automated email notification mechanism** for activities identified as requiring suspicious-activity alerts.
9. **To evaluate the developed CNN-LSTM framework** using accuracy, precision, recall, and F1-score.
10. **To investigate the suitability of the integrated framework for intelligent surveillance applications** by examining its ability to recognize different activity patterns from the considered surveillance-video dataset.

1.4 Limitations of the Study

The proposed system has several limitations that should be considered when interpreting its results.

First, classification performance is dependent on the quality and characteristics of the surveillance-video dataset used during development and evaluation. The implemented dataset is described as a custom surveillance-video dataset containing six predefined activity classes, with an 80% training and 20% testing division. Consequently, the reported performance should not automatically be generalized to surveillance environments containing activity categories, camera conditions, or populations that are substantially different from those represented in the dataset.

The system is also affected by the quality of the input video. Surveillance footage may contain changes in illumination, image quality, background conditions, motion blur, and other environmental variations. Such factors can influence the quality of extracted visual features and consequently affect classification performance. The project documentation identifies variations in illumination, background conditions, image quality, and human movement as relevant challenges for surveillance activity recognition.

Another limitation is the predefined nature of the classification problem. The implemented framework recognizes six specified categories: Normal, Fighting, Fire, Burglary, Accident, and Shooting. An activity that does not belong to these categories cannot necessarily be interpreted correctly by the classifier as a new or previously unknown type of suspicious behavior.

The temporal model also depends on the quality and organization of the frame sequences supplied to it. Since the LSTM learns relationships among consecutive frames, unsuitable frame extraction or insufficient temporal information may reduce its ability to distinguish activities with similar visual characteristics.

Furthermore, the current implementation primarily demonstrates automated classification of uploaded surveillance videos together with result notification. Although the broader system design discusses possible extensions involving multiple cameras, cloud computing, IoT, edge computing, and larger surveillance environments, these should not be represented as fully implemented components of the current experimental system unless separately validated.

Finally, the reported evaluation metrics represent performance on the considered experimental data. Wider deployment would require additional validation across different locations, camera

viewpoints, lighting conditions, crowd densities, activity variations, and surveillance environments.

1.5 Rationale of the Study

The motivation for this research arises from the increasing difficulty of manually examining large quantities of surveillance footage. A conventional CCTV system can continuously record events, but the usefulness of that footage for immediate security response depends heavily on the ability of personnel to identify relevant activities in time. Continuous observation of several video streams is demanding and can increase the possibility of delayed or missed recognition of important events.

An automated activity-recognition mechanism can assist by examining video sequences and identifying predefined activity patterns without requiring an operator to inspect every frame manually. However, effective surveillance analysis requires more than frame-level image classification. Human activities develop over time, meaning that both the visual characteristics of individual frames and the temporal progression of those characteristics contribute to the interpretation of an event.

The rationale for combining CNN and LSTM is therefore based on their complementary roles. CNN processing provides a mechanism for learning spatial representations from individual frames, while LSTM processing enables the system to model temporal dependencies across the sequence. The project documentation specifically adopts this combination so that spatial and temporal information can be considered together during activity recognition.

The study is also motivated by the need for timely notification after an activity has been recognized. Rather than limiting the system output to a classification label, the implemented framework provides a confidence score and can generate an email notification when suspicious activity is identified. This adds an alert-oriented component to the recognition pipeline and can help reduce the delay between automated detection and user awareness.

Therefore, the proposed research focuses on developing a practical deep-learning-based surveillance framework that combines video preprocessing, spatial feature extraction, temporal sequence analysis, classification, and notification. The reported experimental results—96.50% accuracy, 96.04% precision, 97.00% recall, and 96.52% F1-score—provide evidence of the effectiveness of the implemented framework on the evaluated data.

2. LITERATURE REVIEW

The increasing use of Artificial Intelligence (AI), Machine Learning (ML), and computer vision has created new possibilities for automated surveillance and human activity recognition. Conventional surveillance systems generally depend on continuous observation of CCTV footage, while automated approaches attempt to identify relevant activities from visual data without requiring every frame to be manually inspected. Suspicious human activity recognition is particularly challenging because activities are represented not only by the appearance of individual video frames but also by the sequence and temporal progression of human movements. This literature review examines traditional surveillance approaches, deep-learning-based activity recognition, CNN-based spatial feature extraction, LSTM-based temporal analysis, anomaly detection, pose-based recognition, and intelligent video-surveillance approaches that provide the foundation for the proposed CNN-LSTM-based suspicious human activity recognition system.

2.1 Traditional Surveillance and Activity Monitoring

Traditional surveillance systems primarily depend on CCTV cameras and continuous human observation. Security personnel monitor one or more video feeds and manually identify events that may require attention. Basic motion-detection mechanisms can also be employed to identify movement within a monitored scene.

However, traditional surveillance approaches have several limitations:

a) Continuous dependence on human supervision.

b) Difficulty in monitoring multiple surveillance feeds simultaneously.

c) Limited capability to understand complex human activities.

d) Increased possibility of missed events because of prolonged monitoring.

e) Delayed identification of suspicious activities.

f) Limited ability to distinguish between normal movement and specific suspicious activities.

Key Insight: Traditional surveillance provides continuous visual monitoring but places a significant burden on human operators. The limitations of manual observation and basic motion detection create a need for automated systems capable of interpreting activities directly from surveillance videos.

2.2 Deep Learning-Based Human Activity Recognition

Deep Learning has become an important approach for automated video analysis because deep neural networks can learn meaningful representations directly from visual data. Human activity recognition uses these learned representations to identify patterns associated with different actions or behaviors.

Research in suspicious human activity recognition has explored deep-learning models for automatically classifying activities from surveillance footage. CNN-based approaches are effective for extracting spatial characteristics from individual frames, while sequential learning models can be used to analyze how these characteristics change over time.

Deep-learning-based surveillance systems primarily focus on:

a) Automatic extraction of visual features from video frames.

b) Recognition of activity-specific visual patterns.

c) Analysis of temporal relationships within video sequences.

d) Automated classification of predefined human activities.

e) Reduction of dependence on continuous manual monitoring.

Despite these advantages, deep-learning approaches face several challenges:

1. Model performance depends on the quality and diversity of training data.
2. Large labelled datasets may be required for reliable activity recognition.
3. Training deep-learning models can require considerable computational resources.
4. Recognition performance may decrease under substantially different surveillance conditions.
5. Complex scenes involving multiple individuals may increase classification difficulty.

Key Insight: Deep learning provides an effective foundation for automated human activity recognition, but reliable surveillance analysis requires models capable of representing both visual characteristics and temporal activity patterns.

2.3 CNN-Based Spatial Feature Extraction

Convolutional Neural Networks (CNNs) are widely used in computer-vision applications because they can automatically learn spatial features from images and video frames. In surveillance activity recognition, CNNs can identify visual characteristics such as shapes, textures, object arrangements, and movement-related patterns present within individual frames.

The CNN component of a video-recognition system generally performs the following functions:

a) Receives preprocessed video frames.

b) Learns spatial representations from individual frames.

c) Extracts important visual characteristics.

d) Converts frame-level visual information into feature representations.

e) Provides extracted features for subsequent activity analysis.

CNN-based recognition offers several advantages because feature extraction is learned automatically rather than being completely dependent on manually designed descriptors. However, frame-level spatial information alone may not be sufficient to determine the complete nature of a human activity.

For example, individual frames from different activities may contain visually similar arrangements of people and objects. The distinction between such activities may become apparent only when the sequence of movements is considered.

Key Insight: CNNs provide an effective mechanism for learning spatial information from surveillance frames, but spatial representation alone does not fully capture the temporal progression of human activities. This creates a requirement for complementary sequence-based modeling.

2.4 LSTM-Based Temporal Activity Analysis

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network designed for processing sequential information. In video activity recognition, LSTM networks can analyze feature sequences obtained from consecutive frames and learn relationships between earlier and later observations.

LSTM-based temporal analysis can consider:

a) Sequential frame information.

b) Relationships between consecutive visual features.

c) Temporal progression of human movements.

d) Changes in activity patterns over a video sequence.

e) Long-term dependencies within sequential data.

The use of LSTM is particularly relevant to surveillance activity recognition because many human activities cannot be reliably identified from a single frame. The sequence of movements provides additional information that helps distinguish one activity from another.

However, LSTM-based approaches also have limitations:

1. Their effectiveness depends on the quality of the input feature sequences.
2. Inadequate frame extraction may reduce useful temporal information.
3. Longer sequences can increase computational requirements.
4. Performance can depend on the temporal characteristics represented in the training data.
5. Activities that are poorly represented in the training dataset may remain difficult to classify.

Key Insight: LSTM provides temporal modeling capabilities that complement CNN-based spatial feature extraction. Combining both approaches can provide a more complete representation of activities occurring across surveillance-video sequences.

2.5 CNN-LSTM-Based Activity Recognition

The combination of CNN and LSTM provides a framework in which spatial and temporal characteristics of video data can be analyzed together. The CNN extracts relevant features from individual frames, while the LSTM processes the resulting sequence of features to learn temporal dependencies.

The general CNN-LSTM processing procedure consists of:

a) Acquisition of surveillance video.

b) Extraction and preprocessing of video frames.

c) Spatial feature extraction using CNN.

d) Formation of a temporal sequence from extracted features.

e) Temporal sequence analysis using LSTM.

f) Classification of the resulting representation.

g) Generation of the final activity prediction.

The proposed system follows this general approach. OpenCV is used to process the input surveillance videos and extract the required frames. The CNN extracts spatial information from the processed frames, and the LSTM analyzes the temporal relationships among the extracted features. The resulting representation is passed to the classification stage to identify the activity category. The project documentation specifically describes the CNN-LSTM combination as a mechanism for learning both spatial features and temporal patterns from surveillance videos.

The implemented system considers six activity categories:

1. Normal
2. Fighting
3. Fire
4. Burglary
5. Accident

6. Shooting

The final classification is presented to the user together with a confidence score.

Key Insight: CNN-LSTM architecture combines complementary spatial and temporal learning capabilities, making it suitable for surveillance-video activity classification where both individual-frame information and sequential movement patterns contribute to activity recognition.

2.6 Anomaly Detection in Surveillance Videos

Anomaly detection represents another important research direction in intelligent surveillance. Unlike predefined activity-classification systems, anomaly-detection approaches attempt to identify observations that differ from learned normal patterns. Autoencoder-based methods have been investigated for detecting unusual events without requiring explicit labels for every possible abnormal activity.

Anomaly-detection approaches can provide the following capabilities:

a) Identification of deviations from normal activity patterns.

b) Detection of previously unseen abnormal behavior.

c) Reduction of dependence on predefined activity categories.

d) Automated analysis of large surveillance-video collections.

However, anomaly-detection approaches may face challenges when environmental conditions vary significantly. Changes in illumination, background appearance, camera position, crowd density, and other scene characteristics may be interpreted incorrectly as abnormal behavior.

The literature survey included in the project identifies autoencoder-based anomaly detection as an existing direction while noting the possibility of false positives under significant environmental variation.

Key Insight: Anomaly detection can identify deviations from learned patterns, but predefined activity classification provides a more direct approach when the target activity categories are known. The proposed system therefore focuses on classification of six specified surveillance activities rather than unrestricted anomaly discovery.

2.7 Pose-Based Human Activity Recognition

Human pose information provides another method for analyzing activity by representing body configuration and movement through keypoints. Pose-based approaches can extract information about human body positions and use these representations for activity classification.

Pose-based activity recognition can support:

a) Analysis of human body configurations.

b) Representation of movement patterns through keypoints.

c) Recognition of behavioral patterns.

d) Reduction of dependence on complete image appearance.

However, pose-based approaches can be sensitive to occlusion, complex backgrounds, crowded scenes, and difficulties in accurately identifying body keypoints. These limitations can affect recognition performance in surveillance environments where individuals may partially overlap or where visual conditions are uncontrolled.

The project's literature survey identifies OpenPose-based human activity recognition as an existing approach and notes its sensitivity to occlusions and complex backgrounds.

Key Insight: Pose-based recognition provides useful information about human movement, but surveillance environments can introduce occlusion and background-related challenges. A CNN-LSTM approach using video-frame information provides an alternative mechanism for activity recognition without making pose estimation the primary processing stage.

2.8 Intelligent Video Surveillance Using Deep Neural Networks

Deep neural networks have also been applied to intelligent video-surveillance systems for automated object detection, tracking, and activity recognition. Such approaches attempt to reduce manual surveillance requirements by using machine-learning models to analyze visual information automatically.

Intelligent surveillance frameworks can support:

a) Automated video analysis.

b) Detection of relevant visual patterns.

c) Recognition of human activities.

d) Reduction of manual monitoring requirements.

e) Improved surveillance management.

Despite these advantages, deep-neural-network-based surveillance systems may require significant computational resources. Their performance can also depend on the complexity of the surveillance environment and the quality of the available training data.

The project literature survey identifies intelligent video-surveillance approaches using deep neural networks as an important research direction while highlighting high hardware and computational requirements as a limitation.

Key Insight: Deep neural networks provide strong capabilities for automated surveillance analysis, but practical implementation requires an appropriate balance between recognition capability, computational requirements, dataset availability, and system complexity.

2.9 Limitations of Existing Systems

Despite significant developments in automated surveillance and human activity recognition, existing approaches continue to exhibit several limitations:

i. Dependence on labelled datasets for predefined activity-classification tasks.

ii. Difficulty in recognizing activities when environmental conditions differ significantly from the training data.

iii. Increased computational requirements associated with complex deep-learning models.

iv. Limited ability of frame-only approaches to represent temporal activity information.

v. Potential false positives in anomaly-detection systems under changing environmental conditions.

vi. Sensitivity of pose-based approaches to occlusion and complex backgrounds.

vii. Difficulty in handling multiple activities or individuals within complicated surveillance scenes.

viii. Continued requirement for human supervision in many conventional surveillance environments.

ix. Limited integration between automated activity recognition and immediate user notification.

x. Difficulty in monitoring large numbers of surveillance feeds through manual observation.

Key Insight: The literature indicates a need for an automated surveillance framework that can combine spatial feature extraction, temporal sequence analysis, activity classification, and notification capabilities while maintaining a practical processing pipeline.

3. RESEARCH METHODOLOGY

A systematic research methodology is required to develop an effective surveillance framework capable of recognizing human activities from video sequences. The methodology adopted in this study focuses on the development of an Automatic Recognition of Suspicious Human Activities from Surveillance Videos Using Deep Learning. The proposed approach combines surveillance-video acquisition, frame extraction, preprocessing, spatial feature extraction, temporal sequence analysis, activity classification, confidence estimation, and automated notification within a unified framework. The methodology is designed to analyze visual and temporal characteristics of surveillance footage and identify predefined activities through a CNN-LSTM-based deep-learning model.

3.1 Research Design

This research follows an applied and experimental research design aimed at developing a practical deep-learning-based solution for automated surveillance activity recognition. The study follows a systematic development process in which surveillance videos are collected, frames are extracted and prepared, deep-learning models are applied, activity predictions are generated, and the resulting classification performance is evaluated using quantitative measures.

The research design includes the following components:

1. **Video-Based Analysis:** Surveillance videos are analyzed to identify visual and temporal patterns associated with different human activities. The video data are processed as sequences of frames rather than being treated only as individual images.
2. **Image and Sequence Processing:** OpenCV is used for video reading, frame extraction, and preprocessing. The resulting frames provide the input for spatial feature extraction and subsequent temporal analysis.
3. **Experimental Model Development:** A CNN-LSTM architecture is developed to combine spatial feature extraction and temporal sequence learning. The CNN processes individual frames, while the LSTM analyzes the sequence of extracted features.
4. **Quantitative Evaluation:** The developed classification framework is evaluated using accuracy, precision, recall, and F1-score to determine its recognition performance.
5. **Automated Activity Recognition:** The trained framework classifies surveillance videos into predefined activity categories and generates a confidence score for the predicted class.
6. **Automated Notification:** The system provides an email notification mechanism for activities identified as requiring a suspicious-activity alert.

3.2 System Modules

The proposed intelligent surveillance system consists of the following major modules:

• Video Input Module

The Video Input Module provides the surveillance video required for activity recognition. The system accepts surveillance-video data and passes the input to the video-processing stage for further analysis. The input video represents a sequence of visual observations from which the system extracts the information required for activity classification.

• Video Processing Module

The Video Processing Module is responsible for reading the input video and extracting individual frames. OpenCV is used to process the video data and prepare the frames for subsequent deep-learning analysis.

The processing stage converts the continuous video input into a sequence of frames that can be analyzed by the recognition model.

• Frame Preprocessing Module

The Frame Preprocessing Module prepares extracted video frames for deep-learning processing. The preprocessing stage includes operations required to transform the raw frames into an appropriate representation for the CNN.

The preprocessing process includes:

• Frame extraction
• Image resizing
• Pixel-value normalization
• Input formatting
• Preparation of frame sequences

These operations provide consistent input to the feature-extraction stage and support effective model processing.

• CNN Feature Extraction Module

The CNN Feature Extraction Module forms the spatial-learning component of the proposed framework. The CNN processes individual video frames and extracts visual characteristics relevant to activity recognition.

The module learns spatial representations from the processed frames and converts the visual information into feature representations that can subsequently be analyzed as a sequence.

• LSTM Temporal Analysis Module

The LSTM Temporal Analysis Module processes the sequence of features obtained from the CNN. Its purpose is to learn temporal relationships among successive frames and identify patterns associated with the progression of human activities.

The module considers:

• Sequential frame information
• Temporal relationships
• Changes in visual features
• Activity progression
• Long-term sequence dependencies

The combination of CNN-based spatial features and LSTM-based temporal analysis enables the system to consider both the appearance of individual frames and the evolution of activities over time.

• Activity Classification Module

The Activity Classification Module receives the temporally processed representation and determines the activity category associated with the input surveillance video.

The implemented framework considers six activity categories:

• Normal
• Fighting
• Fire
• Burglary
• Accident
• Shooting

The classification stage produces the predicted activity together with the associated confidence score.

• Confidence Estimation Module

The Confidence Estimation Module provides an indication of the model's confidence in the generated classification. The confidence value is presented together with the predicted activity so that users can interpret the classification output more effectively.

• Alert and Notification Module

The Alert and Notification Module is responsible for notifying the user when a suspicious activity requiring an alert is identified. The implemented framework supports email notification, allowing the detected event to be communicated to the registered user.

• Result Display Module

The Result Display Module presents the recognition output to the user. The displayed information includes the predicted activity and its corresponding confidence score.

• Model Training Module

The Model Training Module forms the learning component of the proposed system. It is responsible for training the CNN-LSTM framework using the prepared surveillance-video data.

The training process includes:

• Dataset preparation
• Frame extraction
• Data preprocessing
• Feature learning
• Temporal sequence formation
• Model training
• Classification evaluation

• Prediction Generation Module

After model training, the Prediction Generation Module applies the learned CNN-LSTM model to surveillance-video input. It generates the predicted activity category and associated confidence score.

The outputs include:

• Predicted activity
• Activity confidence
• Suspicious-activity notification when applicable

3.3 Tools and Technologies Used

The proposed system is developed using the following technologies and computational components:

1. **Python** – Used as the primary programming language for implementing the video-processing and deep-learning components of the surveillance system.
2. **OpenCV** – Used for reading surveillance videos, extracting frames, and performing the required image and video preprocessing operations.
3. **Convolutional Neural Network (CNN)** – Used for extracting spatial features from individual surveillance-video frames.
4. **Long Short-Term Memory (LSTM)** – Used for analyzing sequences of extracted features and learning temporal relationships between successive video frames.

5. **CNN-LSTM Architecture** – Used as the primary deep-learning framework for combining spatial and temporal information during activity recognition.
6. **Softmax Classification** – Used at the classification stage to generate activity-class predictions.
7. **Custom Surveillance Video Dataset** – Used for training and evaluating the activity-recognition framework. The dataset contains six activity categories corresponding to the implemented classification task.
8. **Email Notification Mechanism** – Used to communicate suspicious-activity detection results to the registered user.

3.4 System Architecture Description

The proposed Automatic Recognition of Suspicious Human Activities from Surveillance Videos Using Deep Learning operates as a sequential video-processing and activity-classification pipeline. The architecture integrates video input, frame extraction, preprocessing, spatial feature extraction, temporal modeling, activity classification, confidence estimation, and notification within a unified surveillance framework.

Initially, surveillance video is provided as the input to the system. OpenCV reads the video and extracts individual frames from the input sequence. The extracted frames are subsequently prepared through the required preprocessing operations so that they can be supplied to the deep-learning model.

After preprocessing, the frames are passed to the CNN component. The CNN learns spatial characteristics from individual frames and converts the visual information into feature representations. These representations preserve relevant information about the visual content of the surveillance scene.

The extracted feature representations are then arranged as a temporal sequence and provided to the LSTM component. The LSTM analyzes the sequence and learns relationships between successive frame-level features. This temporal processing allows the framework to consider how visual patterns develop across the video rather than relying only on isolated frames.

The resulting representation is passed to the classification stage. The Softmax-based classifier determines the most likely activity category among the six predefined classes: Normal, Fighting, Fire, Burglary, Accident, and Shooting.

The final prediction and confidence score are presented to the user. When the recognized activity satisfies the suspicious-activity alert condition, the notification component can generate an email alert for the registered user.

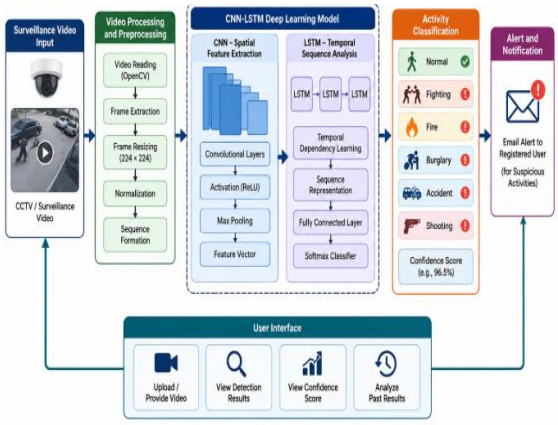

Fig. System Architecture

Step-wise Architecture

Step 1: Surveillance Video Input Layer: A surveillance video is supplied to the system as the primary input for activity recognition.

Step 2: Video Processing: OpenCV reads the input video and extracts individual frames from the surveillance sequence.

Step 3: Frame Preprocessing: The extracted frames are prepared for model processing through resizing, normalization, and appropriate input formatting.

Step 4: CNN Spatial Feature Extraction: The preprocessed frames are supplied to the CNN, which learns relevant spatial characteristics from individual frames.

Step 5: Temporal Sequence Formation: The extracted frame-level features are organized into sequential representations corresponding to the temporal structure of the video.

Step 6: LSTM Temporal Modeling: The LSTM processes the feature sequence and learns temporal relationships between successive observations.

Step 7: Activity Classification: The learned representation is supplied to the classification stage, which determines the most probable activity among the predefined categories.

Step 8: Confidence and Notification: The predicted activity and confidence score are presented to the user. When a suspicious activity requiring notification is identified, an email alert is generated.

3.5 Model Implementation and Validation

The proposed system implements a deep-learning-based video classification framework rather than relying exclusively on manual surveillance. The implementation combines video preprocessing, CNN-based spatial feature extraction, LSTM-based temporal modeling, classification, and automated notification.

Implementation Steps

1. Define the surveillance activity categories to be recognized by the system.
2. Collect and organize the surveillance-video dataset according to the defined activity classes.
3. Read the surveillance videos using OpenCV.
4. Extract frames from the input videos.
5. Preprocess the extracted frames by applying the required resizing, normalization, and input-formatting operations.
6. Provide the processed frames to the CNN for spatial feature extraction.
7. Organize the extracted CNN features into temporal sequences.
8. Provide the feature sequences to the LSTM for temporal relationship learning.
9. Train the CNN-LSTM framework using the prepared surveillance-video data.
10. Apply the trained model to input surveillance videos and generate the predicted activity.
11. Calculate and display the confidence associated with the classification result.
12. Generate an email notification when a suspicious activity requiring an alert is detected.

Validation

The developed system is validated using quantitative classification performance measures and observed prediction results.

The validation process considers:

- Classification accuracy
- Precision
- Recall
- F1-score
- Activity classification correctness
- Confidence of predictions
- Notification functionality
- Recognition performance across the defined activity categories

The reported evaluation results for the developed CNN-LSTM activity-classification framework are:

- Accuracy: **96.50%**
- Precision: **96.04%**
- Recall: **97.00%**
- F1-score: **96.52%**

These results indicate the classification performance obtained by the developed framework on the evaluated surveillance-video data.

3.6 Ethical and Practical Considerations

The proposed surveillance activity-recognition system is intended to support security monitoring and should be operated responsibly.

• Responsible Decision Support

The system provides automated activity predictions to assist surveillance personnel. The generated classification should therefore be treated as decision-support information rather than an unquestionable determination of an event.

• Dataset Quality

Recognition performance depends on the quality and diversity of the surveillance videos used for model development. Appropriate dataset preparation is therefore important for reducing errors and improving model reliability.

• False Detection Consideration

Automated classification can produce incorrect predictions under difficult visual or environmental conditions. Suspicious-activity alerts should therefore be interpreted together with appropriate human verification where required.

• Privacy Consideration

Surveillance systems process visual information involving people. Deployment should therefore consider appropriate privacy, access-control, data-management, and organizational requirements when handling surveillance footage.

• Responsible Notification

Automated email notifications should provide useful information to authorized users without being interpreted as definitive evidence of an incident without further verification.

• Model Reliability

The model should be evaluated using representative surveillance data before deployment in environments that differ substantially from the experimental conditions.

3.7 Evaluation Criteria

The performance of the proposed surveillance activity-recognition system is evaluated using the following criteria:

1. **Classification Accuracy**

Measures the proportion of surveillance-video instances correctly classified by the model among the total evaluated instances.

2. **Precision**

Measures the proportion of correct positive predictions among the instances classified as a particular activity. Precision is important for assessing the reliability of generated activity classifications and reducing incorrect suspicious-activity interpretations.

3. **Recall**

Measures the ability of the model to correctly identify instances belonging to the actual activity classes. Recall is particularly important in surveillance applications where failure to identify relevant activities can reduce the usefulness of automated monitoring.

4. **F1-Score**

Provides a combined measure of precision and recall and is used to assess the overall balance between correct activity identification and incorrect classification.

5. **Activity Recognition Performance**

Evaluates the ability of the CNN-LSTM framework to distinguish among the six implemented activity categories: Normal, Fighting, Fire, Burglary, Accident, and Shooting.

6. **Confidence Estimation**

Evaluates whether the system provides a meaningful confidence value together with the predicted activity, allowing users to interpret the classification output.

7. **Notification Performance**

Evaluates whether the system correctly initiates the email notification mechanism when a suspicious activity requiring an alert is identified.

8. **Temporal Recognition Capability**

Evaluates the ability of the LSTM component to use sequential frame information for distinguishing activities that cannot be adequately characterized from individual frames alone.

9. **Processing Reliability**

Considers the ability of the complete processing pipeline to accept surveillance videos, extract frames, perform model inference, generate classifications, and present the resulting output without processing failure.

10. **Practical Surveillance Support**

Considers the ability of the proposed framework to reduce dependence on continuous manual observation by automatically analyzing surveillance videos and notifying users about detected suspicious activities.

4. DATA ANALYSIS

4.1 Dataset Analysis

The proposed Automatic Recognition of Suspicious Human Activities from Surveillance Videos Using Deep Learning utilizes a custom surveillance-video dataset containing video samples organized according to predefined human activity categories. The dataset provides the basis for training and evaluating the deep-learning model developed for automated activity recognition.

The activity categories considered by the system are Normal, Fighting, Fire, Burglary, Accident, and Shooting. The surveillance videos are processed using OpenCV, through which individual frames are extracted and prepared for subsequent deep-learning analysis. The project documentation specifies an 80% training and 20% testing division for the activity-recognition dataset.

The extracted frames provide the visual information required by the CNN component, while sequences of extracted features are subsequently processed by the LSTM component to capture temporal relationships between successive frames.

Analysis

Before applying the CNN-LSTM model, the surveillance videos undergo preprocessing to convert the raw video information into suitable frame-based input. The preprocessing stage includes video reading, frame extraction, resizing, normalization, and preparation of frame sequences.

The prepared dataset supports the following analytical process:

- **Normal Activity Recognition:** Identifies surveillance sequences corresponding to normal activity.
- **Fighting Activity Recognition:** Identifies video sequences representing fighting-related activity.
- **Fire Activity Recognition:** Identifies surveillance sequences containing fire-related activity.
- **Burglary Activity Recognition:** Identifies activity associated with burglary scenarios.
- **Accident Activity Recognition:** Identifies video sequences corresponding to accident-related events.
- **Shooting Activity Recognition:** Identifies surveillance sequences associated with shooting activity.

The CNN extracts spatial characteristics from the individual frames, while the LSTM analyzes the sequence of extracted features to learn temporal activity patterns. The final classification stage generates the predicted activity and its associated confidence score.

The dataset therefore provides both spatial and temporal information required for the CNN-LSTM-based activity-recognition process.

4.2 System Performance Analysis

The performance of the proposed system is analyzed based on its ability to correctly recognize predefined human activities from surveillance-video sequences. The framework combines video preprocessing, CNN-based spatial feature extraction, LSTM-based temporal analysis, and activity classification to generate the final prediction.

The performance analysis considers the following observations:

- The system successfully processes surveillance videos and converts them into frame sequences suitable for deep-learning analysis.
- The CNN component extracts spatial characteristics from the processed video frames.
- The LSTM component analyzes sequential feature information and captures temporal relationships between successive frames.
- The classification component identifies the activity category from the learned spatial-temporal representation.
- The system provides a confidence score together with the predicted activity.
- The notification mechanism can generate an email alert when a suspicious activity requiring notification is identified.

The documented experimental results for the developed CNN-LSTM activity-classification framework report an accuracy of **96.50%**, precision of **96.04%**, recall of **97.00%**, and F1-score of **96.52%**.

These results indicate strong classification performance on the evaluated surveillance-video data. However, the reported values should be interpreted in relation to the dataset, activity categories, preprocessing procedure, and experimental conditions used in the project. They should not be considered universal performance guarantees for surveillance environments that differ substantially from the evaluated data.

4.3 Performance Metrics Evaluation

The proposed surveillance activity-recognition system is evaluated using quantitative classification measures appropriate for assessing multi-class video classification performance.

• **Accuracy:** Accuracy measures the proportion of correctly classified surveillance-video instances among the total number of evaluated instances. The documented CNN-LSTM experiment reports an accuracy of **96.50%**.

• **Precision:** Precision measures the proportion of correct predictions among the instances classified as a particular activity. Higher precision indicates that fewer incorrect activity classifications are produced. The reported precision is **96.04%**.

• **Recall:** Recall measures the ability of the model to correctly identify instances belonging to the actual activity categories. It is particularly relevant to surveillance applications because failure to recognize an activity may result in a missed event. The reported recall is **97.00%**.

• **F1-Score:** The F1-score provides a combined measure of precision and recall and is useful for evaluating the balance between correct recognition and incorrect classification. The reported F1-score is **96.52%**.

• **Activity Classification Performance:** The classification performance is assessed across the six predefined activity categories: Normal, Fighting, Fire, Burglary, Accident, and Shooting.

• **Confidence Estimation:** The system provides a confidence score with the predicted activity, allowing the user to assess the model's confidence in the generated classification.

• **Notification Performance:** The notification functionality is evaluated based on the system's ability to initiate an email alert when a suspicious activity requiring notification is detected.

• **Temporal Recognition Performance:** The CNN-LSTM architecture is assessed based on its ability to utilize sequential frame information in addition to individual-frame visual characteristics.

4.4 User Evaluation and Observations

The proposed system provides an automated surveillance-analysis interface through which users can provide surveillance-video input and obtain activity-recognition results. The system is designed to identify six predefined activity categories and present the predicted activity together with its confidence score.

The analysis of the developed system provides the following observations:

• **Automated Activity Recognition:** The system analyzes surveillance videos automatically and classifies the observed activity without requiring every frame to be manually examined.

• **Spatial-Temporal Analysis:** The combination of CNN and LSTM enables the system to consider both spatial characteristics of individual frames and temporal relationships between successive frames.

• **Multi-Class Activity Classification:** The system recognizes six predefined categories: Normal, Fighting, Fire, Burglary, Accident, and Shooting.

• **Confidence-Based Output:** The predicted activity is accompanied by a confidence score, providing additional information about the classification result.

• **Automated Notification:** The system supports email notification when a suspicious activity requiring an alert is detected.

• **Reduced Manual Monitoring:** Automated video analysis can reduce dependence on continuous observation of surveillance footage and assist users in identifying potentially important events.

• **Modular Architecture:** The video-processing, feature-extraction, temporal-analysis, classification, and notification components are organized as separate functional stages, allowing individual components to be modified or extended without redesigning the complete recognition workflow.

• **Practical Surveillance Support:** The developed framework provides an automated approach for

analyzing surveillance videos and identifying predefined human activities, supporting the use of deep learning in intelligent surveillance applications.

5. IMPLEMENTATION AND EXPERIMENTS AND RESULTS ANALYSIS

5.1 System Implementation

The proposed Automatic Recognition of Suspicious Human Activities from Surveillance Videos Using Deep Learning is implemented as a modular video-analysis system that combines surveillance-video processing, frame extraction, deep-learning-based feature extraction, temporal sequence analysis, activity classification, confidence estimation, and automated notification. The implementation is intended to provide an automated mechanism for analyzing surveillance videos and identifying predefined human activities.

The system uses Python as the primary programming environment for implementing the video-processing and deep-learning operations. OpenCV is used for reading surveillance videos, extracting frames, and performing the required preprocessing operations. The core recognition architecture combines a Convolutional Neural Network (CNN) with a Long Short-Term Memory (LSTM) network. The CNN is responsible for learning spatial characteristics from individual frames, while the LSTM processes the resulting sequence of features to capture temporal relationships within the video.

The implementation is organized into several functional components. The Video Processing Module receives surveillance-video input and extracts frames using OpenCV. The Frame Preprocessing Module prepares the extracted frames for model processing. The CNN Feature Extraction Module learns spatial representations from the individual frames, while the LSTM Temporal Analysis Module processes the extracted feature sequences.

The Activity Classification Module determines the predicted activity from the learned spatial-temporal representation. The implemented system recognizes six activity categories: Normal, Fighting, Fire, Burglary, Accident, and Shooting. The classification stage generates the predicted activity together with a confidence score.

The Alert and Notification Module provides an additional mechanism for communicating detected suspicious activities to the registered user through email notification. This allows the system to extend beyond classification by providing an automated response mechanism for relevant detected events.

Before model training, the surveillance-video data are processed to generate appropriate model inputs. The processing includes video reading, frame extraction, resizing, normalization, and temporal sequence preparation. The resulting data are then used for training and evaluating the CNN-LSTM architecture.

5.2 Experimental Setup

The experimental setup is designed to evaluate the effectiveness of the developed CNN-LSTM framework for recognizing human activities from surveillance videos. The experiment follows a structured sequence consisting of dataset preparation, video processing, frame extraction, preprocessing, model training, testing, and performance evaluation.

The surveillance-video dataset contains samples representing six activity categories: Normal, Fighting, Fire, Burglary, Accident, and Shooting. The project documentation specifies an 80% training and 20% testing division. The training portion is used for learning the spatial and temporal characteristics associated with the activity classes, while the testing portion is used to evaluate the trained model using previously unseen video samples.

The experimental procedure consists of the following stages:

1. **Dataset Preparation:** Surveillance videos are collected and organized according to the predefined activity categories.
2. **Video Processing:** The collected videos are processed using OpenCV to enable systematic frame extraction.
3. **Frame Extraction:** Individual frames are extracted from the surveillance-video sequences.

4. **Frame Preprocessing:** Extracted frames are prepared through resizing, normalization, and appropriate input formatting.
5. **Spatial Feature Extraction:** The CNN processes individual frames and learns relevant spatial representations.
6. **Temporal Sequence Formation:** CNN-derived feature representations are arranged according to their temporal order within the corresponding video sequences.
7. **Temporal Model Training:** The LSTM processes the feature sequences and learns temporal relationships associated with the activity categories.
8. **Activity Classification:** The learned spatial-temporal representation is supplied to the classification stage to determine the predicted activity.
9. **Model Testing:** The trained CNN-LSTM model is applied to the testing data to determine its recognition performance.
10. **Performance Evaluation:** The model is evaluated using accuracy, precision, recall, and F1-score.
11. **Confidence Generation:** The system provides a confidence score together with the predicted activity.
12. **Notification:** When a suspicious activity requiring notification is detected, the email notification mechanism communicates the result to the registered user.

5.3 Experimental Results

The experimental results demonstrate the performance of the developed CNN-LSTM framework for the evaluated surveillance-video activity-classification task. The documented experimental evaluation reports an accuracy of **96.50%**, indicating that the model correctly classified a high proportion of the evaluated surveillance-video instances.

The detailed performance values reported for the developed system are:

Metric	CNN-LSTM Model
Accuracy	96.50%
Precision	96.04%
Recall	97.00%
F1-Score	96.52%

The reported accuracy of 96.50% indicates strong overall classification performance on the evaluated data. The precision value of 96.04% indicates that a high proportion of the activities predicted by the model corresponded to the appropriate activity classes. The recall value of 97.00% demonstrates that the model successfully identified a high proportion of the relevant activity instances. The F1-score of 96.52% indicates a strong balance between precision and recall.

The experimental results provide evidence supporting the use of a combined CNN-LSTM architecture for surveillance-video activity recognition. The CNN contributes spatial feature extraction, while the LSTM provides temporal sequence modeling, allowing the system to analyze both frame-level visual characteristics and activity progression.

The reported values represent the documented experimental setup and dataset. Therefore, they should not be interpreted as universal performance guarantees for surveillance environments with substantially different camera conditions, activity distributions, video quality, backgrounds, or environmental characteristics.

5.4 Graphical Analysis

Graphical analysis provides a compact representation of the experimentally obtained classification results. The performance graph can represent the four documented evaluation metrics of the CNN-LSTM model: accuracy, precision, recall, and F1-score.

The corresponding values are:

- Accuracy: **96.50%**
- Precision: **96.04%**
- Recall: **97.00%**
- F1-Score: **96.52%**

The graphical comparison indicates that all four reported performance measures remain above 96%. Recall represents the highest reported metric at 97.00%, while precision is reported at 96.04%. The F1-score of 96.52% indicates that the model maintains a strong balance between precision and recall.

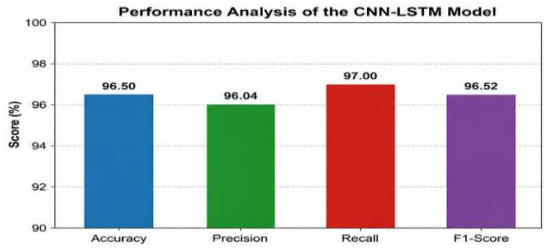

Fig. 5.1. Performance Analysis of the CNN-LSTM Model

The performance graph can be used to visually compare the classification measures and demonstrate the overall consistency of the model across the selected evaluation criteria.

Activity Classification Analysis

The developed system performs multi-class recognition across six predefined activity categories:

• Normal

• Fighting

• Fire

• Burglary

• Accident

• Shooting

The classification result for each surveillance-video input is generated by the trained CNN-LSTM model. The output contains the predicted activity and its corresponding confidence score.

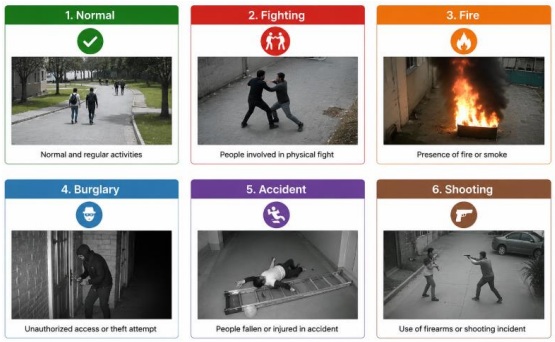

Fig. 5.2. Activity Classification Categories

Observation: The documented numerical results indicate strong classification performance for the evaluated CNN-LSTM model. The results support the feasibility of combining spatial and temporal deep-learning representations for automated recognition of predefined activities in surveillance videos.

5.5 Results Analysis

The experimental evaluation indicates that the proposed surveillance framework can effectively apply deep learning to automated human activity recognition. The developed CNN-LSTM model achieved an overall accuracy of **96.50%**, demonstrating strong classification performance on the evaluated surveillance-video dataset.

The detailed evaluation further reports **96.04% precision, 97.00% recall, and 96.52% F1-score**. The combination of these values indicates that the model provides a balanced classification performance for the evaluated activity-recognition task.

From a surveillance perspective, the results demonstrate the usefulness of processing video information through both spatial and temporal learning. The CNN extracts visual characteristics from individual frames, while the LSTM analyzes how these characteristics change across a sequence. This combination allows the system to consider information that may not be adequately represented by a single frame.

The system recognizes six predefined activity categories: Normal, Fighting, Fire, Burglary, Accident, and Shooting. This multi-class structure allows the framework to distinguish between normal activity and several predefined events that may require security attention.

The confidence-estimation component provides additional information together with the predicted activity. This enables the user to observe not only the classification result but also the model's confidence associated with that prediction.

The notification component further extends the functionality of the recognition framework. When a suspicious activity requiring notification is identified, the system can initiate an email alert to the registered user. This provides a mechanism for communicating potentially important events without

requiring continuous manual observation of the video output.

The experimental findings therefore support the feasibility of the proposed **Automatic Recognition of Suspicious Human Activities from Surveillance Videos Using Deep Learning** as an automated surveillance-analysis framework. The system integrates video processing, CNN-based spatial feature extraction, LSTM-based temporal analysis, multi-class activity recognition, confidence estimation, and email notification within a common architecture.

At the same time, the reported results should be interpreted within the boundaries of the available dataset and experimental configuration. Further evaluation using larger and more diverse surveillance datasets, different camera viewpoints, illumination conditions, backgrounds, crowd densities, and real-world surveillance environments would be required to establish broader generalization of the model.

6. CONCLUSION AND FUTURE WORK

The proposed **Automatic Recognition of Suspicious Human Activities from Surveillance Videos Using Deep Learning** presents an automated framework for recognizing predefined human activities from surveillance-video sequences. The system integrates video processing, frame extraction, CNN-based spatial feature extraction, LSTM-based temporal analysis, activity classification, confidence estimation, and email notification within a unified surveillance framework. This approach reduces dependence on continuous manual observation and provides an automated mechanism for analyzing surveillance footage.

The CNN component extracts relevant spatial characteristics from individual video frames, while the LSTM component analyzes temporal relationships among successive feature representations. The combined CNN-LSTM architecture enables the system to consider both visual information and the progression of activities across video sequences. The developed system recognizes six predefined activity categories: Normal, Fighting, Fire, Burglary, Accident, and Shooting.

The experimental evaluation reports an accuracy of **96.50%**, precision of **96.04%**, recall of **97.00%**, and F1-score of **96.52%**. These results demonstrate strong classification performance for the evaluated surveillance-video dataset and indicate the effectiveness of combining spatial and temporal deep-learning techniques for automated activity recognition. The confidence-estimation component provides additional information regarding the generated prediction, while the email notification mechanism enables detected suspicious activities to be communicated to the registered user.

However, the obtained results are dependent on the dataset, activity categories, preprocessing procedures, and experimental conditions used during development and evaluation. Variations in camera viewpoints, illumination, background conditions, occlusion, video quality, and activity characteristics may affect recognition performance. Therefore, the reported results should be interpreted within the scope of the evaluated experimental environment.

6.1 FUTURE WORK

Future development of the proposed system can focus on improving its applicability to more complex and diverse surveillance environments. The activity dataset can be expanded by incorporating additional normal and suspicious activities and by including greater variation in camera viewpoints, illumination conditions, backgrounds, and human movement patterns. Such expansion can improve the ability of the model to generalize beyond the conditions represented in the current dataset.

The system can also be extended toward real-time surveillance-stream analysis instead of processing previously recorded video sequences. Further research may investigate more advanced CNN architectures, temporal-learning techniques, attention mechanisms, and other deep-learning approaches to improve recognition performance and computational efficiency.

Additional work can investigate multi-person activity recognition, simultaneous event detection, improved confidence calibration, and more comprehensive alert-management mechanisms. Integration with real-world CCTV infrastructure,

edge-computing platforms, and larger surveillance datasets can further evaluate the practical performance of the proposed framework.

These extensions can strengthen the robustness, scalability, and practical applicability of the proposed deep-learning-based surveillance system while maintaining its primary objective of automated human-activity recognition.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," *Advances in Neural Information Processing Systems*, vol. 27, pp. 568–576, 2014.

[2] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[3] L. Sun, K. Jia, K. Chen, D.-Y. Yeung, B. E. Shi, and S. Savarese, "Lattice Long Short-Term Memory for Human Action Recognition," *Proc. IEEE International Conference on Computer Vision*, 2017, pp. 2147–2156.

[4] B. Mahasseni and S. Todorovic, "Regularizing Long Short-Term Memory With 3D Human-Skeleton Sequences for Action Recognition," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3054–3062.

[5] T. Shu, S. Todorovic, and S.-C. Zhu, "CERN: Confidence-Energy Recurrent Network for Group Activity Recognition," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5523–5531.

[6] K. Vignesh, G. Yadav, and A. Sethi, "Abnormal Event Detection on BMTT-PETS 2017 Surveillance Challenge," *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 36–43.

[7] Y. Ma, L. Yao, and others, "Learning Activity Progression in LSTMs for Activity Detection and Early Detection," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[8] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, "A Multi-Stream Bi-Directional Recurrent Neural Network for Fine-Grained Action Detection," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[9] S. Sudhakaran and O. Lanz, "Convolutional Long Short-Term Memory Networks for Recognizing First Person Interactions," *Proc. IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2339–2346.

[10] A. Ramaswamy, K. Seemakurthy, J. Gubbi, and B. Purushothaman, "Spatio-Temporal Action Detection and Localization Using a Hierarchical LSTM," *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 764–765.

[11] R. Thomanek, T. Rolletschke, B. Platte, C. Hosel, C. Roschke, R. Manthey, M. Heinzig, R. Vogel, F. Zimmer, M. Vodel, M. Eibl, and M. Ritter, "Real-Time Activity Detection of Human Movement in Videos via Smartphone Based on Synthetic Training Data," *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 2020, pp. 160–164.

[12] F. Zhou, L. Wang, Z. Li, W. Zuo, and H. Tan, "Unsupervised Learning Approach for Abnormal Event Detection in Surveillance Video by Hybrid Autoencoder," *Neural Processing Letters*, vol. 52, pp. 961–975, 2020, doi: 10.1007/S11063-019-10113-W.

[13] OpenCV, "OpenCV Documentation: Video I/O," OpenCV Documentation.

[14] OpenCV, "OpenCV Documentation: Getting Started with Videos," OpenCV Documentation.

[15] Z. Lan, Y. Zhu, A. G. Hauptmann, and S. Newsam, "Deep Local Video Feature for Action Recognition," *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.