WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Security Camera Analysis for Threat Detection

Srija Reddy[1], Likhith Choudari[2], Himneesh Kambampally[3], Premkumar Chithaluru[4], J Himabindu [5], R Vijaya Lakshmi [6]

[1]Department of IT, MGIT(A), Gandipet, Hyderabad, 500075, Telangana, India.
[2]Department of IT, MGIT(A), Gandipet, Hyderabad, 500075, Telangana, India.
[3]Department of IT, MGIT(A), Gandipet, Hyderabad, 500075, Telangana, India.
Email: asrija-csb213203@mgit.ac.in; clikhith-csb213216@mgit.ac.in; khimneesh-csb213227@mgit.ac.in; chpremkumar-it@mgit.ac.in; jhimabindu-it@mgit.ac.in; rvijayalakshmi-it@mgit.ac.in;

ABSTRACT

The AI-driven security camera system aims to enhance real-time threat detection and response through advanced machine learning techniques. The system utilizes a live camera feed to monitor and analyze the environment for potential secu- rity threats, including the presence of weapons, violent behavior, accidents, or unattended suspicious objects. By employing object detection, behavior analysis, and anomaly detection algorithms, the camera automatically identifies and flags potential risks. In the event of a detected threat, the system triggers an alarm or sound notification and visually highlights the threat on the screen, enabling immediate attention and response from security personnel or automated systems. This innovative solution improves public safety and security by providing proac- tive threat detection with minimal human intervention, making it suitable for use in public spaces, commercial establishments, transportation hubs, and other high-security environments.

Keywords: AI-driven security camera, Real-time threat detection, Machine learning techniques,Environment monitoring.

1. Introduction

Real-time violence detection using surveillance systems has emerged as a critical area of research aimed at enhancing public safety by identifying irregular and violent behav- ior in live video streams. The primary goal of such systems is to classify normal and abnormal activities accurately, enabling timely interventions in environments like pub- lic spaces, organizations, and transportation hubs. However, the implementation of such systems faces several challenges, including the dynamic nature of violent actions, the unavailability of sufficient labeled datasets, and the computational demands of processing video data in real time.

Manually annotating videos for training models is a labor-intensive and costly process, leading to limitations in dataset availability. Additionally, defining violent actions precisely is difficult due to their variability in context and execution. Existing systems often struggle to balance high accuracy with low computational overhead, particularly in complex surveillance environments.

Recent advancements in convolutional neural networks (CNNs) have provided promising solutions for addressing these challenges. Various CNN architectures, includ- ing AlexNet, VGG-16, and GoogleNet, have been explored for violence detection tasks. However, the MobileNet model, renowned for its lightweight design and computational efficiency, has shown superior performance in terms of accuracy, loss, and processing speed[1].

The shift to modern internet protocol (IP) cameras, which support 4K resolu- tion—27 times larger than traditional analogue cameras—has further improved the quality of surveillance. However, the increased resolution has also raised concerns regarding bandwidth usage. Each camera can require up to 10 Mbps for streaming, depending on factors such as resolution, frame rate, and compression. Continuous data streaming from multiple cameras in a surveillance network exacerbates band- width issues, often leading to higher operational costs and connectivity challenges. Furthermore, the integration of cloud computing in surveillance amplifies network vulnerabilities, emphasizing the need for robust network infrastructure.

Video surveillance systems are widely used in public spaces, enterprises, financial institutions, and retail sectors, serving as a deterrent against crime and a tool for solving criminal activities. Despite their benefits, the growing use of these systems raises privacy concerns. With modern image processing technologies enabling detailed analysis, there is apprehension about misuse of recorded data. Privacy protection laws like the General Data Protection Regulation (GDPR) in Europe and the Federal Information Security Management Act (FISMA) in the USA require organizations to handle surveillance data responsibly, ensuring compliance with strict data protection standards[2].

As the volume of video footage increases exponentially, it becomes increasingly difficult for human analysts to process and identify relevant events promptly. Analysts often have to wait hours before abnormal activities are captured and reported, which delays response times and diminishes the effectiveness of surveillance systems. Video anomaly detection is further complicated by its inherent nature as a one-class problem. In practical scenarios, models are typically trained on normal footage, and any unusual patterns are flagged as anomalies. However, due to the impossibility of capturing every possible normal action in a dataset, the system may misclassify regular activities as anomalies, resulting in false alarms[3].

Fig. 1 sample of a violence

In video surveillance, detecting and classifying abandoned objects is a critical task for maintaining security and safety in public and private spaces. The objective is to identify unattended items in real-time, which could pose security risks or operational hazards. This research focuses on comparing two feature extraction techniques—Scale- Invariant Image Transform (SIFT) keypoints and geometric primitive features—to determine their effectiveness in accurately classifying objects in a video surveillance context.Feature extraction methods influence classification accuracy and false positive rates across various classification schemes. The experimental results highlight that the classifier based on statistics of geometric primitives' features delivers superior recogni- tion accuracy while maintaining a low false alarm rate. Additionally, the performance of this classification approach remains consistent across different learning algorithms, showcasing its adaptability and robustness[4].

Detecting violence and weaponized activities in closed-circuit television (CCTV) footage is a crucial component of modern surveillance systems, aimed at ensuring pub- lic safety and security in urban environments. This research addresses the challenges of identifying both weaponized and non-weaponized violence by introducing a special- ized dataset—the Smart-City CCTV Violence Detection (SCVD) dataset—designed to capture and analyze weapon distribution in surveillance videos.

To address the complexities of working with 3D video data, we propose the SSIVD- Net (Salient-Super-Image for Violence Detection) method. SSIVD-Net effectively reduces the dimensionality and complexity of 3D video data, minimizing informa- tion loss while enhancing inference speed, performance, and interpretability through the use of Salient-Super-Image representations. This approach ensures that the sys- tem remains scalable and efficient, catering to the requirements of smart cities where real-time video analysis must handle large volumes of data seamlessly.

Additionally introduces the Salient-Classifier, a novel architecture that combines a kernelized approach with a residual learning strategy, further boosting detection accuracy and robustness across different scenarios,rigorously evaluate our proposed SSIVD-Net and Salient-Classifier models on the SCVD dataset and compare their performance against existing state-of-the-art (SOTA) models in violence detection. The results demonstrate significant improvements in detecting both weaponized and non-weaponized instances, showcasing the system's adaptability and effectiveness[5].

2. Literature review

a hybrid model combining CNNs for spatial feature extraction and LSTMs for tempo- ral sequence learning, achieving 98 percent accuracy with 131 frames per second. The system addressed key challenges in violence detection, such as varying conditions and real-time processing needs[1].A U-Net-like architectures and LSTMs to integrate spa- tial and temporal features for violence detection. Their efficient model used MobileNet V2 as an encoder for computational optimization[2].

The field of violence detection in video surveillance. The study categorized method- ologies into three primary approaches: conventional methods, machine learning-based techniques, and deep learning-based models. It analyzed public datasets, feature extraction techniques, and classification methods, highlighting the challenges associ- ated with dataset quality, object identification, and false positive rates. The review identified open issues and emerging trends, emphasizing the importance of balancing accuracy and computational efficiency[6].

A framework for violence detection in crowded surveillance scenes. They utilized the Gaussian Model of Optical Flow (GMOF) for adaptive crowd behavior modeling and introduced the Orientation Histogram of Optical Flow (OHOF) as a descriptor for distinguishing violent from non-violent events. The approach showed robustness and real-time capabilities with high detection precision[7]. The CCTV-Fights dataset, comprising real-world fight

scenarios to address limitations in existing datasets. They evaluated methods like Two-Stream CNNs and 3D CNNs and emphasized the sig- nificance of motion information, particularly Optical Flows, in improving detection performance for real-world applications[8].

3D Convolutional Networks for violence detection, capturing both spatial and temporal features. Their work demonstrated the importance of motion patterns in accurately recognizing violent activities and eliminated the reliance on handcrafted fea- tures, making the method more generalized[9].Detecting school violence, introducing a Decision Tree-SVM (DT-SVM) classifier that achieved high accuracy and precision. The approach used algorithms like Relief-F and Wrapper for feature dimensionality reduction and addressed misclassifications in complex scenes[10].

A spatio-temporal model integrating spatial and temporal attention mechanisms with 2D CNNs. The model effectively handled lighting and occlusion challenges, offering robust performance in real-time violence detection[11].Enhanced surveillance systems by integrating cloud computing and machine learning. Their framework pro- cessed large-scale video data for real-time criminal detection, employing adaptive machine learning models for dynamic environments[12].A deep feature fusion frame- work for violence detection in diverse scenarios. By combining multiple deep learning models, the system demonstrated scalability and superior performance compared to traditional methods[13]. a novel model designed specifically for detecting abnormal events in surveillance camera footage. In recent years, the proliferation of surveillance cameras has become commonplace in both public and private spaces, serving as a criti- cal tool for enhanc- ing security measures. However, the reliance on human surveillance can often lead to oversight due to human error, rendering the monitoring process inef- ficient and prone to missed anomalous events. To address this challenge, researchers have sought to automate the detection of abnormal activities in surveillance camera feeds. Our approach leverages a fusion of well-established deep learning architectures, namely ResNet50 for feature extraction and ConvLSTM for temporal analysis, to iden- tify and classify anomalous behavior within our time-series dataset.[15].

2.1 Challenges

- Difficulty in obtaining large, annotated datasets that comprehensively cover all real- world monitoring scenarios, limiting the training and evaluation of models.
- Challenges in object identification accuracy, especially in environments with low

image quality, occlusions, or complex backgrounds.

- Issues in selecting effective feature extraction methods that remain consistent across varying environments, lighting, and camera angles.
- Scalability problems in processing large volumes of video footage in real-time while

maintaining system responsiveness.

- Difficulty in detecting violent interactions accurately in crowded scenes due to complex interactions and multiple overlapping actions.
- Computational efficiency challenges, as real-time detection requires significant

processing speed and system resources.

- Variability in background complexity, such as crowd behavior and environmental changes, making motion detection and classification less reliable.
- Difficulty in distinguishing subtle non-violent interactions from violent actions due

to the nuances in human movement and behavior.

This paper addresses these challenges with effective analysis.

3. System Model

3.1 Security Camera Analysis

CCTV systems in modern urban landscapes has resulted in an immense volume of video data generated daily. Manual monitoring of this footage is impractical and inefficient, hindering the timely identification and prevention of violent incidents. Tra- ditional video analysis methods often produce inaccurate results, with high rates of false positives and missed events. There is a critical need for the development and implementation of robust machine learning (ML) and deep learning (DL) algorithms specifically tailored for violence detection in CCTV footage. These systems must over- come challenges such as diverse lighting conditions, occlusions, and variations in hu- man behavior, while ensuring real-time processing capabilities. The goal is to create intelligent systems that can reliably identify and respond to violent activities captured in CCTV footage, thereby enhancing public safety and aiding effective law enforce- ment.

3.2 Problem description

Security camera analysis for violence and anomaly detection is a critical challenge in modern surveillance systems, aiming to ensure public safety and protect prop- erty. Despite advances in machine learning and video processing, several issues persist. Obtaining comprehensive, annotated datasets remains difficult due to limi- tations in capturing real-world interactions, variations in lighting conditions, camera angles, occlusions, and background complexity. Feature extraction methods, whether traditional or deep learning-based, struggle to maintain robustness across diverse envi- ronments, often requiring extensive computational resources. Scalability and real-time processing present additional obstacles, as large-scale surveillance networks in smart cities demand high bandwidth, fast inference speeds, and substantial memory. False positives and classification errors arise from difficulties in distinguishing subtle inter- actions, such as differentiating non-violent actions from violent events, particularly in crowded or noisy scenes. The integration of cloud-based surveillance adds to the challenges by introducing issues of network instability, transmission costs, and data privacy concerns. Regulations like GDPR and FISMA further complicate the deploy- ment of these technologies, requiring strict adherence to data protection and privacy laws. In summary, developing an efficient and accurate security camera analysis system must address dataset limitations, feature extraction efficiency, scalability, real-time adaptability, background variability, privacy concerns, and computational constraints, all while ensuring quick and reliable detection across complex, dynamic surveillance environments.

4. Proposed Solution

4.1 Detailed Description

This part is explaining how Security camera analysis is working. Its technical descrip- tion is given step by step by means of algorithm and its flowcharts in subsections ahead section 4.1.1

4.1.1 Algorithms and flowcharts

MobileNetV2, renowned for its efficiency and lightweight design, serves as a potent feature descriptor in our violence detection system. With its depthwise separable con- volutions and inverted residuals, MobileNetV2 efficiently captures intricate features from surveillance video frames. Its flexibility in trade-offs between model size, latency, and accuracy makes it an ideal choice for processing large volumes of CCTV footage, enabling swift and accurate identification of violent events amidst diverse environmen- tal conditions.

MobileNetV2 is a highly efficient and scalable deep learning architecture that has proven valuable in the context of security camera analysis. Designed with a focus on optimizing computational efficiency without sacrificing performance, MobileNetV2

is particularly suited for real-time applications in resource-constrained environments, such as large-scale surveillance networks, smart cities, and edge devices. This makes it an ideal choice for deploying security camera analysis systems that need to handle continuous video streams with low latency and limited hardware resources.

MobileNetV2 leverages a unique depthwise separable convolution technique, which significantly reduces the number of computational operations and memory usage while maintaining high detection accuracy. Its use of depthwise convolutions separates spa- tial and channel-wise operations, enabling more efficient feature extraction with fewer parameters. This allows the architecture to perform well on devices with constrained computational power, such as embedded systems and IoT devices, ensuring smooth integration into real-world surveillance installations.

In the context of violence detection and anomaly detection, MobileNetV2 excels at real-time object detection, human action recognition, and event classification due to its fast inference speed and low latency. It can process high-resolution video frames quickly, making it possible to detect suspicious interactions, abandoned objects, or abnormal activities almost instantaneously. Furthermore, MobileNetV2 maintains robust performance across various lighting conditions, camera angles, and crowded environments, addressing challenges like occlusions and background noise.

The integration of MobileNetV2 in security camera systems also offers scalability, as its compact architecture facilitates deployment across large surveillance networks with thousands of cameras. It minimizes the need for high bandwidth and data transmission costs by enabling on-device processing, thus reducing reliance on cloud infrastructure. This makes it possible to analyze data locally while ensuring faster response times and better privacy control.

In our proposed system, we harness the power of MobileNetV2, a lightweight and effi- cient convolutional neural network (CNN) architecture, to serve as a fea- ture descriptor for processing our surveillance videos. By leveraging MobileNetV2, we extract rich and discriminative features from the video frames, capturing intricate pat- terns and nuances crucial for violence detection. These features serve as the foundation for our violence.

The use of the Gaussian Model of Optical Flow (GMOF) to extract candidate violence regions in a video. These candidate regions are adaptively modeled based on deviations from the normal behavior of the crowd observed in the scene. Implement adaptive modeling within GMOF to identify deviations in crowd behavior, allowing for the dynamic extraction of candidate violence regions.

YOLOv3, short for You Only Look Once version 3, is a state-of-the-art object detec- tion algorithm known for its speed and accuracy. Unlike traditional object detection methods that require multiple passes over an image, YOLOv3 performs detection in a single step, making it incredibly fast. It divides

the input image into a grid and predicts bounding boxes and class probabilities for each grid cell, enabling efficient detection of multiple objects in real-time.

OpenCV (Open Source Computer Vision Library) plays a crucial role in secu- rity camera analysis by providing real-time video processing, object detection, and threat identification capabilities. It facilitates the efficient capture, manipulation, and analysis of video streams from surveillance cameras. OpenCV's key features, such as

frame extraction, image preprocessing, motion detection, and object tracking, allow the implementation of advanced security applications.

For security camera analysis, OpenCV enables real-time video streaming by access- ing video feeds using its cv2.VideoCapture() function. Individual frames can be processed to detect objects, track movements, or identify anomalies like weapons, violent behavior, or suspicious activity. Preprocessing techniques like resizing, gray- scaling, and noise removal ensure optimized input for further analysis. OpenCV also supports integration with machine learning and deep learning models, such as MobileNet, YOLO (You Only Look Once), and SSD (Single Shot Multibox Detector), which are used for detecting people, weapons, or other threats in the video frames.

OpenCV includes Haar cascades and HOG (Histogram of Oriented Gradients) with SVM for detecting faces and pedestrians. More advanced deep learning mod- els like MobileNet-SSD can be applied for detecting humans in crowded or complex environments.

Algorithm 1 Security Camera Analysis using MobileNetV2

```
Initialize video stream
video_stream ← cv2.VideoCapture('camera_url')
Load MobileNetV2 model: mobile_net ← tf.keras.applications.MobileNetV2(weights = 'imagenet')
while True do
    Capture a frame: (ret, frame) ← video_stream.read()
    if ret is False then
        break
    end if
    Resize the frame: frame_resized ← cv2.resize(frame, (224, 224))
    Normalize the frame: frame_normalized ← frame_resized/255.0
    Create a batch: frame_batch ← tf.convert_to_tensor([frame_normalized])
    Predict objects: predictions ← mobile_net.predict(frame_batch)
    Decode predictions: decoded_predictions ← tf.keras.applications.mobilenet_v2.decode_predictions(predictions, top = 5)[0]
    for (label, description, confidence) in decoded_predictions do
        Overlay text on the frame: text ← f"label : confidence : .2f"
        cv2.putText(frame, text, (10, 30+i×30), cv2.FONT_HERSHEY_SIMPLEX, 1, (0, 255, 0), 2)
    end for
    Display the frame: cv2.imshow('SecurityCameraAnalysis', frame)
    if cv2.waitKey(1)&0xFF == ord('q') then
        break
    end if
end while
Release video resources: video_stream.release()
Destroy all OpenCV windows: cv2.destroyAllWindows()
```

The provided **Algorithm 1** describes a system for real-time security camera anal- ysis by integrating OpenCV and the MobileNetV2 deep learning model. The system starts by initializing the video stream from a security camera URL using OpenCV's cv2.VideoCapture() method and loading the MobileNetV2 model pre-trained on the ImageNet dataset. It then enters an infinite loop to continuously process video frames. For each iteration, the system captures a frame from the video stream and checks if the capture was successful. If it fails, the loop terminates. Each frame is then prepro- cessed by resizing it to 224x224 pixels (the input size expected by the MobileNetV2 model) and normalizing the pixel values to a range of [0, 1]. The preprocessed frame is passed through the MobileNetV2 model, which performs object detection and returns predictions containing detected objects and their confidence scores. These predic- tions are decoded into meaningful labels and confidence percentages. The system then overlays this detection information onto the original video frame using OpenCV's cv2.putText() function. The annotated video frames are displayed in real-time in an OpenCV window. The loop continues

until the user presses the 'q' key, at which point it gracefully terminates by releasing the video stream resources and closing all OpenCV windows. This ensures that resources are properly cleaned up, and the system shuts down efficiently. The integration of MobileNetV2 and OpenCV enables robust real-time detection and annotation of objects in video streams, making it a powerful solution for security monitoring and camera-based object detection.

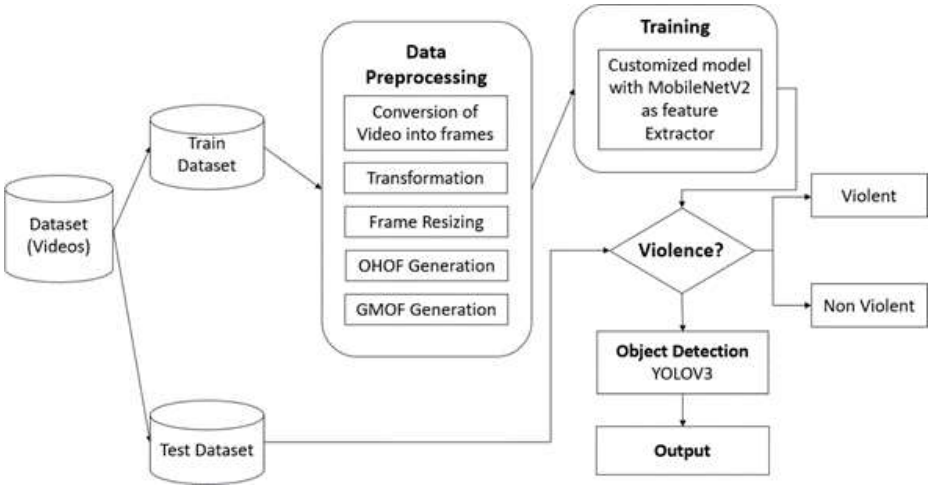

Fig. 2 Flowchart of Security Camera Analysis

As shown in Figure 2, The flowchart illustrates a security camera-based violence detection framework utilizing video datasets and advanced preprocessing techniques. Initially, video datasets are split into training and testing sets. The data preprocess- ing step involves converting videos into frames, performing transformations, resizing frames, and generating motion features like OHOF (Optical Flow Histograms) and GMOF (Global Motion Optical Flow) for effective feature extraction. A customized deep learning model based on MobileNetV2 is then trained as a feature extrac- tor to identify violent and non-violent activities. During analysis, the trained model

determines whether violence occurs in the video feed. If violence is detected, it pro- ceeds to the object detection stage using the YOLOv3 algorithm, which identifies specific objects within the frames. Finally, the system outputs results, categorizing the detected events as either violent or non-violent. This process combines advanced machine learning and computer vision techniques to enhance the accuracy and efficiency of violence detection in real-time surveillance systems.

5 Evaluation parameters and criteria

1. Accuracy

This focuses on the system's ability to correctly identify objects, individuals, or activities within the surveillance area. This includes detecting abnormal behaviors, such as violence or theft, with minimal errors. accuracy is impacted by the system's ability to process high-quality inputs. Cameras with high resolution (e.g., 1080p or higher), good frame rates, and adaptive features like low-light performance enhance detection capabilities. Advanced models such as MobileNetV2 and YOLOv3 further boost accuracy by effectively analyzing spatial and temporal features of the video. These models enable the system to capture intricate patterns in motion and object behavior, reducing errors even in dynamic or crowded scenes.

Finally, robust accuracy also depends on the quality of the training dataset used for the system. A well-curated dataset with diverse scenarios, including various lighting conditions, camera angles, and types of activities, ensures that the system generalizes well to real-world applications. By achieving high accuracy, security camera systems can reliably identify threats, reduce false alarms, and enhance safety in critical environments like public spaces, schools, and private properties.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

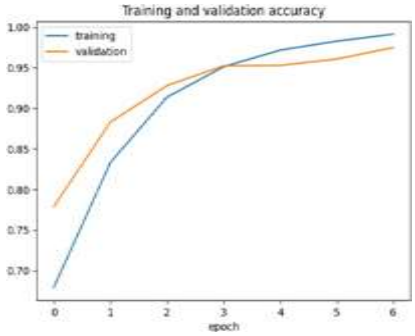

Fig. 3 Training and Validation Accuracy.

As shown in Figure 3, The training accuracy of around 99.77 percent indicates how well the model correctly classifies instances within the training dataset. This high accuracy suggests that the model has learned the patterns and features of the training data very well, making ac- curate predictions for the majority of instances it was trained on. Such a high training accuracy typically reflects a well-fitted model that has effectively minimized errors during the training phase, although it's important to ensure it generalizes well to new data.The training accuracy of around 99.77classifies instances within the training dataset. This high accuracy suggests that the model has learned the patterns and features of the training data very well, making ac- curate predictions for the majority of instances it was trained on. Such a high training accuracy typically reflects a well-fitted model that has effectively minimized errors dur- ing the training phase, although it's important to ensure it generalizes well to new data. The test accuracy of approximately 97.61unseen data that it hasn't been trained on. Despite being slightly lower than the train- ing accuracy, this value indicates strong generalization capabilities.The test accuracy of approximately 97.61 percent measures the model's performance on unseen data that it hasn't been trained on. Despite being slightly lower than the train- ing accuracy, this value indicates strong generalization capabilities.

2. Precision

In the context of machine learning and statistics refers to the proportion of true positive predictions (correctly predicted positives) out of all positive predictions made by a model. It measures the accuracy of positive predictions, indicating how reliable the model is when it identifies a positive instance (such as detecting violence in a video) among all instances it predicts as positive. A high precision indicates that when the model predicts an event (like violence), it is likely to be correct, minimizing false positives. It is calculated as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

In this formula: - True Positives are the number of correctly predicted positive in- stances. - False Positives are the number of incorrectly predicted positive instances (instances predicted as positive but are actually negative). Precision is crucial in tasks where the cost of false positives is high, such as in med- ical diagnostics or security applications like violence detection, where misidentifying a non-violent event as violent can have significant consequences.

3. Recall

Also known as sensitivity or true positive rate, is a metric used in machine learn- ing and statistics to evaluate the completeness of predictions. It measures the proportion of true positive instances (correctly predicted positives) that are correctly identified by a model out of all actual positive instances in the dataset. In simpler terms, recall answers the question: "Out of all the positive instances that exist, how many did the model correctly identify?" A high recall indicates that the model is effectively capturing most of the positive instances, minimizing false negatives (instances where the model fails to identify a positive instance that actually exists). For example, in a weapon detection system, high recall means that the system successfully identifies most instances where a weapon is present, even in challenging conditions like low light or occlusions. Recall is particularly crucial in critical applications where missing a threat could have severe consequences.

Mathematically, recall is defined as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

4. F1 Score

This is a metric that combines both precision and recall into a single value, provid- ing a balanced measure of a model's performance in binary classification tasks. It is the harmonic mean of precision and recall, designed to account for situations where either precision or recall alone might be misleading. In binary classification, precision measures the accuracy of positive predictions, while recall measures how well the model captures all positive instances. The F1 score balances these two metrics by considering their harmonic mean, which gives equal weight to both precision and recall. This makes the F1 score useful when you want to seek a balance between precision and recall, rather than favoring one over the other. Mathematically, the F1 score is defined as:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

5. Confusion Matrix

This is a table that allows visualization of the performance of a clas- sification model by presenting a summary of the model's predictions versus the actual out- comes in a tabular format. It is especially useful for evaluating the performance of machine learning algorithms in binary (two-class) or multi-class classification prob- lems. The confusion matrix is an essential evaluation tool for the security camera analysis project, particularly in the task of violence detection and classification. It provides a clear breakdown of the system's performance by categorizing predic- tions into true positives, true negatives, false positives, and false negatives. This detailed analysis helps identify specific strengths and weaknesses in the model. For instance, it highlights whether the system is more prone to false positives, such as incorrectly flagging normal activities as violence, or false negatives, where actual violent incidents are missed.

6 Performance evaluation

Performance evaluation is a critical aspect of this project, as it determines the effective- ness and reliability of the system in detecting and classifying violence in surveillance footage. Processing Speed and Latency are crucial for real-time violence detection. The system's ability to process video streams efficiently without delays is assessed by measuring frame-per-second (FPS) rates and response times for generating alerts. A well-optimized system ensures minimal latency, enabling timely interventions during critical situations.

Robustness Evaluation focuses on the system's ability to maintain high perfor- mance under varying environmental conditions, such as low lighting, occlusions, and camera angle variations. The use of techniques like Optical Flow Histogram (OHOF) and Gradient-based Motion Flow (GMOF) enhances robustness by capturing motion patterns, even in challenging scenarios. Testing the system across diverse datasets, including indoor and outdoor surveillance videos, ensures that it generalizes well to real-world environments.

Finally,Scalability and Integration are evaluated to ensure the system can handle multiple camera feeds and integrate seamlessly into existing surveillance infrastruc- tures. The system's ability to maintain consistent performance while processing data from various sources is crucial for deployment in large-scale applications, such as smart cities or institutional security systems.

Overall, performance evaluation ensures the system meets its objectives of accu- rate, real-time violence detection while being robust and scalable for practical use cases. These insights guide iterative improvements, enhancing the system's effective- ness in ensuring public safety.

6.1 Results and Discussion

The results highlight the system's high effectiveness in detecting and categorizing violence with minimal errors. The combination of MobileNetV2 for feature extraction, OHOF, and GMOF for motion analysis, and YOLOv3 for object detection contributed to the model's robust performance. The use of a well-annotated dataset, including diverse environmental conditions, further ensured the system's generalizability.

However, certain challenges were observed. The system's performance slightly declined in crowded or highly dynamic environments, where motion patterns became complex, leading to higher false negatives. This suggests a need for further refine- ment of motion-based features or integrating advanced techniques such as attention mechanisms to improve focus on relevant regions.

Additionally, the system's reliance on high-quality inputs, such as well-lit scenes, was evident. Although low-light performance was commendable, integrating infrared or thermal imaging could further enhance accuracy in such scenarios.

The processing speed and low latency indicate the system's readiness for real- time deployment, making it suitable for applications such as public space monitoring, institutional security, and smart city implementations. However, scalability tests with multiple concurrent video streams showed a minor reduction in FPS, suggesting the need for optimized resource allocation in large-scale deployments.

Overall, the project successfully demonstrates the feasibility and effectiveness of an automated violence detection system in enhancing security through intelligent video surveillance. Future enhancements, such as adaptive learning models and addi- tional contextual features, can further elevate the system's performance and expand its applications.

7. Conclusion and Future Scope

In this project, we aimed to develop a model for classifying violent and non-violent actions in videos. Leveraging deep learning techniques, including optical flow calcu- lation (OHOF and GMOF) and feature extraction using the MobileNetV2 architecture pretrained on the ImageNet dataset, along with object detection using YOLOv3, we achieved significant progress. Our model demonstrated promising results, achieving a high accuracy of approxi- mately 97.61that our model effectively learned to distinguish between violent and non-violent ac- tions in videos. Furthermore, the integration of YOLOv3 allowed us to detect objects in real-time, enhancing the model's capabilities. Overall, our model's success in accurately classifying violent and non-

violent ac- tions in videos, combined with YOLOv3's object detection capabilities, demonstrates its potential for applications in surveillance, security, and content moderation, contribut- ing to the enhancement of safety and security in various domains.

Looking beyond the current scope of the project, there are several avenues for future research and development: Multi-modal Fusion Integrating additional modalities such as audio or text along- side video data could provide richer contextual information for improved action recognition. Fusion techniques such as late fusion or early fusion could be ex- plored to effectively combine information from multiple sources.Incorporating temporal modeling techniques such as recurrent neural networks (RNNs) or 3D convo- lutional neural networks (CNNs) could capture temporal dependencies in video data more effectively. This could en- hance the model's ability to recognize complex actions and gestures over time.

8. References

[1] Irfanullah, T. Hussain, A. Iqbal, B. Yang, and A. Hussain, "Real time violence detection in surveillance videos using convolutional neural networks," Multimedia Tools and Applications, vol. 81, pp. 38 151–38 173, 2022.

[2] R. Vijeikis, V. Raudonis, and G. Dervinis, "Efficient violence detection in surveil- lance," Sensors, vol. 22, no. 6, p. 2216, 2022.

[3] K. B. Sahay, B. Balachander, B. Jagadeesh, G. A. Kumar, R. Kumar, and L. R. Par- vathy, "A real time crime scene intelligent video surveillance systems in violence detection framework using deep learning techniques," Computers and Electrical Engineering, vol. 103, p. 108319, 2022.

[4] A. F. Otoom, H. Gunes, and M. Piccardi, "Feature extraction techniques for aban- doned object classification in video surveillance," in 2008 15th IEEE International Conference on Image Processing. IEEE, 2008, pp. 1368–1371.

[5] T. Aremu, L. Zhiyuan, R. Alameeri, M. Khan, and A. E. Saddik, "Ssivd-net: A novel salient super image classification detection technique for weaponized violence," arXiv preprint arXiv:2207.12850, 2022.

[6] B. Omarov, S. Narynov, Z. Zhumanov, A. Gumar, and M. Khassanova, "State- of-the-art violence detection techniques in video surveillance security systems: a systematic review," PeerJ Computer Science, vol. 8, p. e920, 2022.

[7] T. Zhang, Z. Yang, W. Jia, B. Yang, J. Yang, and X. He, "A new method for violence detection in surveillance scenes," Multimedia Tools and Applications, vol. 75, pp. 7327–7349, 2016.

[8] M. Perez, A. C. Kot, and A. Rocha, "Detection of real-world fights in surveil- lance videos," in ICASSP 2019-2019 IEEE International Conference on Acous- tics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 2662–2666.

[9] C. Ding, S. Fan, M. Zhu, W. Feng, and B. Jia, "Violence detection in video by using 3d convolutional neural networks," in Advances in Visual Computing. ISVC 2014, ser. Lecture Notes in Computer Science, vol. 8888. Springer, Cham, 2014.

[10] L. Ye, L. Wang, H. Ferdinando, T. Sepp¨anen, and E. Alasaarela, "A video-based dt–svm school violence detecting algorithm," Sensors, vol. 20, no. 7, p. 2018, 2020.

[11] J. Mahmoodi and H. Nezamabadi-pour, "A spatio-temporal model for violence detection based on spatial and temporal attention modules and 2d cnns," Pattern Analysis and Applications, vol. 27, p. 46, 2024.

[12] A. Names, "Proposed system for criminal detection and recognition on CCTV data using cloud and machine learning," in 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), 2019.

[13] D. J. Samuel R., F. E, G. Manogaran, V. G.N, T. T, J. S, and A. A, "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional lstm," Computer Networks, vol. 52 151, pp. 191–200, 2019.

[14] S. A. Jebur, K. A. Hussein, H. K. Hoomod, and L. Alzubaidi, "Novel deep feature fusion framework for multi-scenario violence detection," Computers, vol. 12, no. 9, 2023.

[15] S. Vosta and K.-C. Yow, "A cnn-rnn combined structure for real-world violence detection in surveillance cameras," Applied Sciences, vol. 12, no. 3, p. 1021, 2022.