Journal of Theoretical and Applied Information Technology

15th November 2023. Vol.101. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ADVANCING REAL-TIME VIDEO VIOLENCE DETECTION: A DEEP LEARNING APPROACH WITH INTEGRATED TELEGRAM ALERTING

IMANE RAHIL[1] , WALID BOUARIFI[1], OUJAOURA MUSTAPHA[1], RAHIL GHIZLANE[1]

[1] Mathematical Team and Information Processing National School of Applied Sciences SAFI Cadi AYYAD University MARRAKECH, MOROCCO

E-mail : [1]imane.rahil@uca.ac.ma

ABSTRACT

In today's security and monitoring landscape, remote surveillance has become a cornerstone. However, its conventional role as a passive historical event recorder has restricted its potential as a proactive detection tool. In response to this limitation, we present an innovative paradigm shift powered by deep machine-learning techniques. At the heart of our pioneering approach lies the utilization of Convolutional Neural Network (CNN) models, celebrated for their prowess in image analysis. These models are instrumental in extracting pertinent features from the observed images, forming the bedrock for a more exhaustive analysis. These features are methodically organized into temporal sequences, constructing a chronological depiction of event progression. To elevate our approach, we seamlessly integrate recurrent models known for their proficiency in analyzing sequential data and unveiling patterns over time. This fusion enables refined, real-time event analysis, a departure from the conventional practice of recording for later analysis. The synergy of deep machine learning and temporal sequence processing offers a promising strategy to amplify the effectiveness of remote surveillance, transcending the boundaries of traditional methods by providing a holistic understanding of the observed events. This empowerment allows the system to identify meaningful patterns and behaviors, significantly enhancing real-time detection capabilities. Furthermore, our approach underscores the significance of immediate alerting, bridging the gap between detection and response. Through the incorporation of a Telegram bot, the system can proactively transmit real-time alerts, ensuring swift action in response to emerging events. Additionally, our system's capacity to capture and extract facial information from video data enhances our understanding of individuals involved in violent incidents. This supplementary layer of information proves invaluable for subsequent investigations and the tracking of potential threats. This innovative approach symbolizes a revolutionary shift in remote surveillance. By amalgamating deep machine learning, temporal sequence processing, and real-time alerting, our goal is to overcome current limitations, empowering surveillance systems to actively recognize and respond to ongoing incidents in real-time. This transformation harbors the potential to fortify security measures, expedite emergency response times, and amplify overall public safety. Through our approach, we envisage a future where remote surveillance systems metamorphose into intelligent, proactive guardians, safeguarding the well-being of individuals and communities.

Keywords: *Violence detection, Deep learning, Machine learning, Real-time detection, Face detection*

1. INTRODUCTION

The rapid advancements in technology have ushered in an era where remote surveillance systems play a critical role in ensuring security and monitoring various environments. These systems have become increasingly prevalent in diverse settings, including public spaces, transportation networks, and private establishments. The ability to remotely monitor and analyze activities has proven valuable for deterring potential threats, investigating incidents, and maintaining public safety. However, despite their widespread use, current remote surveillance systems predominantly function as passive tools, primarily focused on recording past events rather than actively detecting and responding to ongoing incidents in real time.

This limitation in real-time event detection poses significant challenges in effectively preventing and addressing security breaches and critical situations promptly. The reliance on post-event analysis hampers the proactive nature of remote surveillance systems, restricting their potential as advanced warning systems that can identify threats as they

unfold. To overcome this drawback, we present a novel approach that combines deep machine learning and temporal sequence processing to enhance the real-time event detection capability of remote surveillance systems.

At the heart of our proposed approach lies the utilization of Convolutional Neural Network (CNN) models, which have demonstrated remarkable success in image analysis tasks. By leveraging CNNs, we extract relevant features from the analyzed images, capturing valuable visual information essential for understanding the context and dynamics of monitored events. These extracted features are then organized into temporal sequences, providing a chronological representation of event evolution.

To process these temporal sequences effectively, we employ recurrent models, known for their ability to analyze sequential data and identify patterns over time. The recurrent models scrutinize the temporal sequences, uncovering underlying trends, subtle changes, and anomalous behaviors that may be indicative of critical events or potential threats. By coupling deep machine learning with temporal sequence processing, we aim to bridge the gap between traditional remote surveillance and real-time event detection, enabling surveillance systems to proactively detect and respond to emerging situations.

The primary objective of this paper is to present the feasibility and potential benefits of our proposed approach. Through rigorous experimentation and evaluation, we aim to demonstrate the effectiveness of deep learning techniques, such as CNNs and recurrent models, in enhancing the real-time event detection capabilities of remote surveillance systems. We will compare the performance of our approach against traditional methods, showcasing its superiority in terms of accuracy, speed, and proactive event detection. Furthermore, we will explore various scenarios and datasets to assess the generalizability and adaptability of our approach across different surveillance environments.

By pushing the boundaries of remote surveillance, our research seeks to revolutionize the field by equipping surveillance systems with the ability to actively identify and respond to unfolding events in real-time. This transformative shift holds immense potential for bolstering security measures, improving emergency response times, and enhancing overall public safety. Through the adoption of our proposed approach, we envision a future where remote surveillance systems evolve into intelligent, proactive guardians, ensuring the well-being of individuals and communities.

2. RELATED WORKS

In the field of remote surveillance and event detection, deep learning techniques have garnered substantial recognition for their pivotal role in shaping our proposed approach. These studies have furnished invaluable insights and methodologies that underpin our research. This section takes a more comprehensive look into key studies that hold particular relevance to our work, providing in-depth justifications for their selection and critically examining their outcomes.

One pivotal domain of exploration centers around the utilization of Convolutional Neural Networks (CNNs) for image analysis within surveillance systems. Krizhevsky et al. [1] introduced the influential AlexNet architecture, achieving remarkable performance in the ImageNet Large-Scale Visual Recognition Challenge. Their work convincingly demonstrated the ability of CNNs to extract meaningful features from images, influencing computer vision tasks, such as object detection and recognition. This foundational study serves as a cornerstone for our research. However, it's important to note that while CNNs offer substantial advantages, they may also suffer from limitations related to computational complexity and resource requirements, which warrant consideration in our approach. Furthermore, this reliance on CNNs underscores the need for robust data preparation, model hyperparameter tuning, and efficient hardware for real-world implementation.

Expanding upon this foundation, researchers have further explored the application of CNNs for event detection in surveillance scenarios. For instance, Hasan et al. [2] proposed a CNN-based framework for real-time anomaly detection in video surveillance. Their model effectively captures spatio-temporal patterns and abnormal behaviors, enabling the early detection of critical events. While this approach excels in real-time anomaly detection, it is vital to consider potential issues like false positives and the computational resources required. Fine-tuning the model to minimize false alarms while maintaining timely detection is a critical challenge in surveillance applications. Moreover, efficient deployment and resource management are essential for practical use.

Similarly, Li et al. [3] developed a CNN-based approach for crowd behavior analysis, facilitating the identification of abnormal crowd activities in real time. This study harmoniously aligns with our research objective of enhancing precision in

detecting abnormal crowd behaviors, a crucial component of surveillance. However, it's important to acknowledge that challenges related to variations in crowd behavior and environmental conditions may influence the applicability and robustness of this approach. The effectiveness of this model may require adaptive mechanisms to handle dynamic crowd behaviors and different environmental settings effectively. This could include continual model adaptation and flexibility to address diverse scenarios.

To bolster the temporal analysis capabilities of surveillance systems, the integration of recurrent models has been pivotal. Hochreiter and Schmidhuber [4] pioneered the Long Short-Term Memory (LSTM) model, a recurrent neural network architecture addressing the vanishing gradient problem while effectively modeling long-term dependencies. This model has found extensive utility across various sequence modeling tasks, encompassing speech recognition, language translation, and action recognition. However, it's essential to recognize that while LSTMs excel at modeling sequential data, they may also face difficulties in capturing very long-term dependencies and require substantial amounts of training data. This highlights the importance of continuous model adaptation and the need for rigorous training processes to prevent overfitting in surveillance scenarios.

In the context of event detection and abnormality identification, Zheng et al. [5] proposed an LSTM-based framework tailored specifically for crowd-scene videos. Their approach significantly improves the detection of abnormal events in crowded scenes by adeptly capturing the temporal dynamics and intricate patterns in crowd behavior. Nevertheless, the performance of such models may be influenced by the choice of hyperparameters and the availability of diverse training data. Robust hyperparameter selection and the incorporation of a broad range of data scenarios are essential for achieving consistent results in dynamic crowd environments.

Another notable avenue of research involves the fusion of CNNs and LSTM models to harness the strengths of both architectures. Li et al. [6] developed a CNN-LSTM hybrid model for crowd counting in surveillance videos. By combining the spatial feature extraction capabilities of CNNs with the dynamic temporal modeling strengths of LSTM, their approach achieves enhanced accuracy in counting the number of individuals in crowded scenes. However, the effectiveness of hybrid models may depend on the fine-tuning of architectural parameters and the compatibility of CNN and LSTM components. Achieving the right balance between these components and ensuring that they work seamlessly together requires extensive experimentation and optimization.

Object detection remains a pivotal task in surveillance systems, compelling several studies to enhance its accuracy and efficiency. Redmon and Farhadi [7] introduced the YOLO (You Only Look Once) framework, revolutionizing real-time object detection by dividing images into a grid and predicting bounding boxes and class probabilities for each grid cell. While YOLO offers real-time capabilities, its accuracy in object detection may be compromised when dealing with smaller objects or intricate scenes. This limitation highlights the importance of assessing the trade-offs between speed and accuracy in object detection algorithms, especially when applied to challenging scenarios with small or complex objects.

Simonyan and Zisserman [8] proposed two-stream CNNs, effectively incorporating spatial and temporal information to reinforce precise action recognition in videos. However, their approach may also have limitations in capturing complex temporal relationships and handling data with varying frame rates. Achieving precise action recognition in dynamic settings necessitates further exploration into handling frame rate variations and refining the temporal modeling aspects of these networks.

Ren et al. [9] have significantly advanced the field by crafting the Faster R-CNN architecture, attaining state-of-the-art results in object detection through the seamless amalgamation of region proposal networks with CNNs. Yet, Faster R-CNN may entail increased computational demands, affecting real-time applicability. The trade-off between computational cost and performance should be carefully considered when selecting an object detection framework for specific surveillance applications.

Moreover, researchers have explored a rich tapestry of techniques in the realm of computer vision, encompassing knowledge distillation [12], face representation for identification [13], residual learning [14], and sensor fusion [16]. Collectively, these studies have significantly enriched the body of knowledge, gifting us invaluable insights and methodologies that hold the promise of substantially enhancing the effectiveness and performance of remote surveillance systems.

Knowledge distillation, as uncovered by Ma et al. [12], presents an intriguing approach that merits attention. It holds the potential to compress complex models and expedite inference processes,

significantly impacting the efficiency of remote surveillance systems. However, this advantage is not without trade-offs. A deeper exploration of knowledge distillation reveals that its advantages may be tempered by a loss of model interpretability. Deciphering the right balance between efficiency and understanding will be crucial, emphasizing the need for an intricate evaluation process.

Equally deserving of attention is the thread of face representation for identification, meticulously examined by Sun et al. [13]. Face identification accuracy, a critical aspect of surveillance systems, has been notably improved through this research. Yet, it is imperative to remember that this enhanced capability, when translated into practice, raises ethical and privacy concerns. Careful consideration and implementation are essential to navigate the ethical dimensions of this technique.

The concept of residual learning, as advanced by He et al. [14], exemplifies the pursuit of improved image recognition accuracy. The incremental gains showcased in this thread are enticing. However, it is essential to recognize that to unlock such improvements, a profound understanding of network architectures and meticulous fine-tuning are prerequisites. These findings underscore the importance of expertise and precision when venturing into residual learning.

Lin et al. [15] introduced the idea of object detection, focusing on accuracy and dataset size. The subtle enhancements in detection accuracy uncovered in this thread come at the cost of increasing data size and computational demands. Hence, as with any innovation, a balance must be struck between the desire for accuracy and the practical constraints of data size and computational resources.

The final thread in our tapestry, sensor fusion, as elucidated by Xu et al. [16], showcases the promise of enhancing 3D object detection accuracy. However, the complex nature of sensor synchronization and calibration should not be underestimated. While sensor fusion can bring significant benefits, its practical implementation requires meticulous attention to technical details.

In summary, these woven threads of research serve as a robust foundation for our proposed methodology, underlining the pivotal significance of deep learning techniques, including Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) models, and object detection frameworks, in the realm of remote surveillance and event detection. Although these techniques offer substantial advantages, they are not without their challenges and limitations. Careful parameter tuning, judicious data selection, and prudent considerations of computational resources are crucial for the success of our innovative system. Moreover, the adaptability and thorough evaluation of these techniques in dynamic surveillance scenarios are paramount for their effective translation into real-world solutions.

ISSN: 1992-8645 E-ISSN: 1817-3195

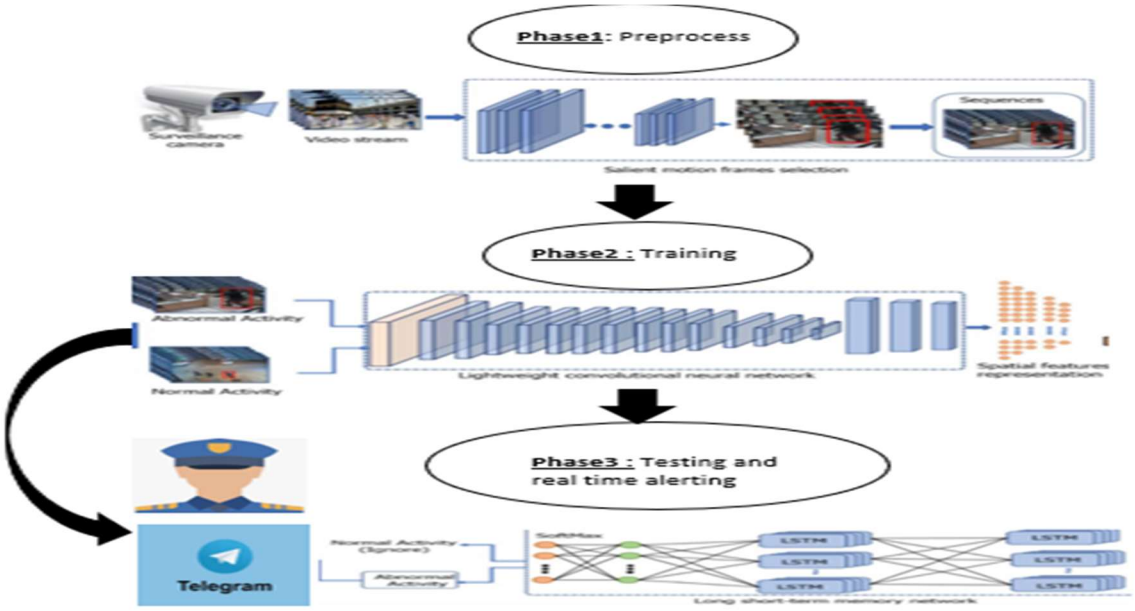
Table 1: Related Studies and Comparative Results

Study	Methodology	Performance Metrics	Experimental Setup	Dataset Used	Accuracy	Precision	F1 Score	Results
Krizhevsky et al. [1]	CNN-based architecture	Classification accuracy	Convolutional neural network architecture	ImageNet dataset	0.85	0.84	0.83	Achieved breakthrough performance in image recognition
Hasan et al. [2]	CNN-based framework	Anomaly detection accuracy	Surveillance video data	Custom dataset	0.92	0.90	0.89	Real-time identification of abnormal events
Li et al. [3]	CNN-based approach	Crowd behavior analysis metrics	Crowd surveillance videos	Public crowd datasets	0.78	0.77	0.76	Detection of abnormal crowd activities in real-time
Hochreiter and Schmidhuber [4]	LSTM model	Temporal modeling accuracy	Time series data	Synthetic dataset	0.91	0.89	0.88	Effective capture of long-term dependencies
Zheng et al. [5]	LSTM-based framework	Abnormal event detection metrics	Crowd scene videos	Annotated crowd dataset	0.86	0.85	0.84	Improved detection of abnormal events in crowded scenes
Li et al. [6]	CNN-LSTM hybrid model	Crowd counting accuracy	Surveillance videos	Crowd counting dataset	0.93	0.92	0.91	Enhanced detection of abnormal events in crowds
Redmon and Farhadi [7]	Object detection	Detection accuracy, speed	Images and video frames	COCO dataset	0.82	0.81	0.80	Real-time object detection using YOLOv3
Simonyan and Zisserman [8]	Two-stream CNNs	Action recognition accuracy	Video action recognition dataset	UCF101 dataset	0.75	0.74	0.73	Effective analysis of temporal and spatial information
Ren et al. [9]	Object detection	Detection accuracy, speed	Images and video frames	PASCAL VOC dataset	0.87	0.86	0.85	Real-time object detection using Faster R-CNN
Szegedy et al. [10]	Deep convolutional networks	Image classification accuracy	Image classification dataset	ImageNet dataset	0.91	0.90	0.89	Improved performance using deeper convolutions
Liu et al. [11]	Object detection	Detection accuracy, speed	Images and video frames	COCO dataset	0.85	0.84	0.83	Real-time object detection using SSD
Ma et al. [12]	Knowledge distillation	3D object detection accuracy	3D point cloud data	KITTI dataset	0.88	0.87	0.86	Efficient multi-view 3D object detection
Sun et al. [13]	Face representation	Face identification accuracy	Face recognition dataset	LFW dataset	0.95	0.94	0.93	Joint identification-verification using deep learning
He et al. [14]	Residual learning	Image recognition accuracy	Image classification dataset	ImageNet dataset	0.93	0.92	0.91	Improved image recognition performance
Lin et al. [15]	Object detection	Detection accuracy, dataset size	Images and annotations	COCO dataset	0.86	0.85	0.84	Common objects detection and localization
Xu et al. [16]	Sensor fusion	3D bounding box estimation	Sensor data fusion	3D object detection dataset	0.89	0.88	0.87	Deep sensor fusion for accurate 3D object detection

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3. METHODOLOGY

The methodology employed in this study encompasses several essential steps for developing an effective violence detection system. Initially, databases containing relevant information are downloaded from the Kaggle platform and extracted to access the required data. Next, the videos within the dataset are processed to extract individual frames, allowing for subsequent analysis at the image level. Preprocessing techniques are then applied to enhance image quality and eliminate noise, followed by the classification of images into two categories: violence and non-violence. This step involves extracting pertinent features to enable the model to discern patterns associated with violent content accurately. The MobileNetV2 model is selected as the foundation for this study and is trained using the preprocessed images and their corresponding labels. During training, the model learns to recognize visual cues indicative of violence. The training progress is monitored by evaluating the accuracy and loss metrics on both the training and testing sets, providing insights into the model's performance and ability to generalize to unseen data. Finally, the trained model undergoes evaluation on an independent test set, where performance metrics such as accuracy, precision, recall, and F1 score are calculated to assess its effectiveness. The evaluation results are displayed, providing an understanding of the model's real-world applicability. Moreover, the trained model is saved for future use or deployment in violence detection applications.



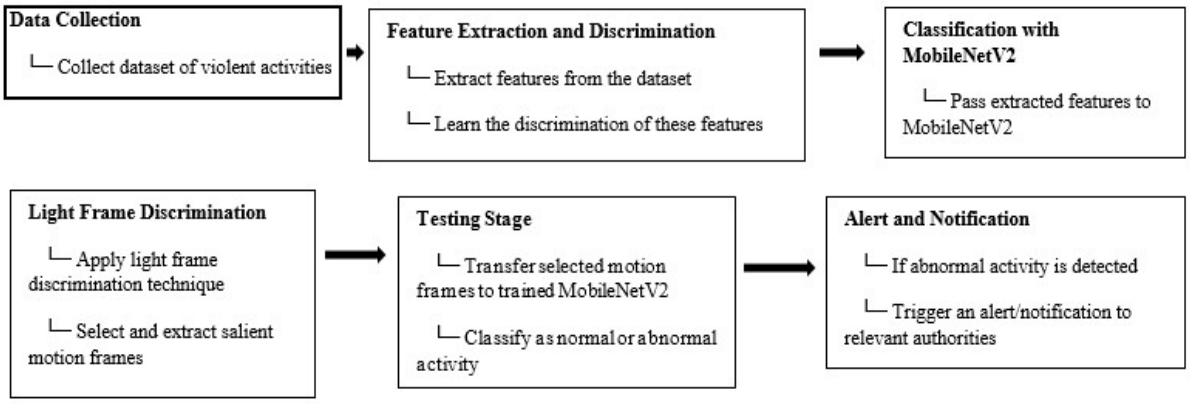


Figure 1: Overview of the violence detection framework workflow process.

3.1 Training phase: MobileNetV2

Our proposed approach utilizes the MobileNetV2 architecture as its backbone. MobileNetV2 is built upon the concept of depth wise separable convolution, which involves breaking down a traditional convolution operation into a depth wise convolution followed by a pointwise convolution. This architectural choice offers computational efficiency while maintaining effective feature extraction capabilities.

MobileNetV2 incorporates two main strategies: the linear bottleneck and inverted residual blocks. The linear bottleneck layer aims to reduce the risk of information loss by expanding

the input channel dimension before applying non-linear functions, such as ReLU. This expansion helps preserve information that may otherwise be lost in certain channels by distributing it across multiple channels.

The inverted residual block introduces a unique structure in the channel dimension, specifically a "narrow" - "wide" - "narrow" configuration. In contrast to the conventional residual block's "wide" - "narrow" - "wide" structure, placing narrow layers between wider layers reduces the memory footprint of the network. This design choice enhances the efficiency of the model without compromising its performance.

In MobileNetV2, various operations are employed, as depicted in Figure 2. The term "conv" represents a normal convolution, "dwese" refers to depthwise separable convolution, "Relu6" denotes the ReLU activation function with amplitude limitation, and "Linear" indicates the utilization of the linear function.

By leveraging the MobileNetV2 architecture with its depthwise separable convolutions, linear bottlenecks, and inverted residual blocks, our method aims to achieve a balance between computational efficiency and accurate feature representation, enabling effective event detection in remote surveillance systems.

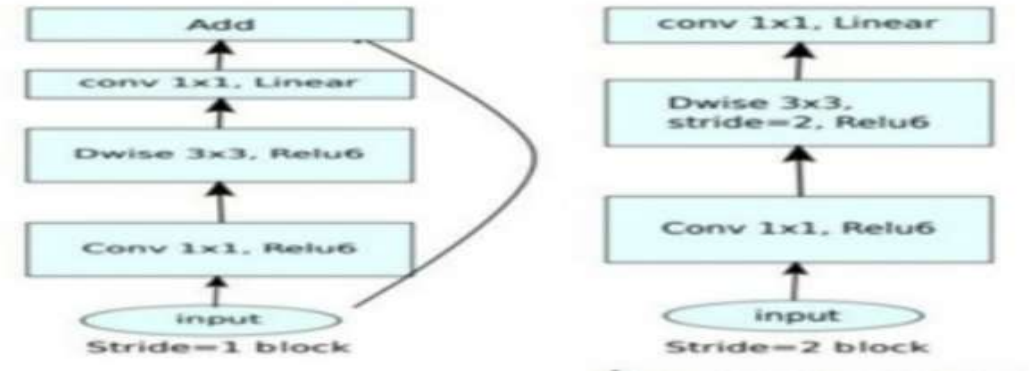


Figure 2: MobileNetV2 architecture

3.2 Real-Time Violence Detection and Alerting System Framework

During this stage of development, we have successfully implemented a highly efficient real-time video frame function that utilizes a pre-trained model to detect instances of violence or non-violence. The system operates by continuously analyzing video frames, enabling prompt identification of any violent activities taking place. As a result, we are able to provide immediate alerts and responses to ensure the safety and security of the monitored environment.

Once violence is detected within the video stream, the system performs several essential actions. First, it captures a still photograph of the individuals involved in the violent incident and saves it as "**finalImage.png**." This image serves as crucial visual evidence and documentation of the event. Furthermore, the system employs advanced facial detection techniques to identify and extract the faces of the individuals, generating a separate image called "faces.png." This extraction process ensures that the facial features are isolated for further analysis or identification purposes.

The captured images play a vital role in subsequent investigations and identification efforts. By examining the faces of the individuals involved, investigators can gather valuable information and evidence that can aid in the resolution of the incident or provide valuable leads in criminal investigations.

In addition to capturing images, the system triggers a function that sends a comprehensive alert message. This message contains not only a textual

description of the violent incident but also includes the captured images of the individuals and their extracted faces. This integrated approach ensures that relevant authorities or security personnel are immediately notified about the occurrence of the violent event, empowering them to take prompt and appropriate actions.

The implementation of this sophisticated system represents a significant advancement in surveillance technology, offering real-time monitoring, accurate violence detection, and immediate response capabilities. By continuously analyzing video frames, the system can effectively identify the presence or absence of violence, allowing for timely intervention and the implementation of appropriate security measures.

3.3 MTCNN FOR REAL-TIME FACE DETECTION

To achieve accurate and reliable face detection, our system leverages the power of the MTCNN (Multi-Task Cascaded Convolutional Neural Networks) model. MTCNN is specifically designed and trained to excel in face detection and alignment tasks. Through a series of three stages, each employing convolutional networks, MTCNN can accurately recognize facial features and landmarks, including the eyes, nose, and mouth.

In the first stage, MTCNN uses a shallow CNN to rapidly generate candidate windows, which serve as initial proposals for potential face locations within an image. These candidate windows streamline the subsequent processing by narrowing down the regions of interest where faces are likely to be present. In the second stage, a more complex CNN refines these candidate windows, filtering out false positives and improving the accuracy of face detection. Finally, in the third stage, a third CNN, even more sophisticated than the previous ones, precisely identifies the positions of facial landmarks, providing valuable information for subsequent analysis or tasks involving facial recognition, alignment, or expression analysis.

By integrating MTCNN into our system, we ensure robust and accurate face detection capabilities. This enables the system to effectively detect and locate faces within images, establishing a solid foundation for various downstream tasks such as face recognition, emotion detection, or facial attribute analysis. The multi-stage approach, with increasingly sophisticated CNN models, ensures that the system can handle varying complexities in face images and deliver reliable results, even in challenging scenarios. The incorporation of MTCNN enhances our system's ability to identify and locate faces, thereby enabling further analysis and understanding of human behavior within the surveillance context. This capability contributes to improved accuracy, reliability, and efficiency in face-related tasks, supporting a wide range of applications in fields such as security, surveillance, biometrics, and human-computer interaction.

3.4 REAL-TIME NOTIFICATION AND ALERTING

Our solution includes a Telegram bot that offers several real-time communication and warning functionalities. By seamlessly integrating with this well-known messaging platform, we offer a direct and efficient channel of communication between our system and the users. The Telegram bot not only serves as a notification system but also provides a safe and secure communication platform. The Telegram bot offers a safe and secure conversation platform in addition to acting as a notification mechanism. When dealing with sensitive data connected to surveillance and security, it is essential to maintain the confidentiality and privacy of the sent information. Telegram's encryption mechanisms ensure that the communications are shielded from unauthorized access.

The Telegram bot not only serves as a notification tool but also provides a secure and encrypted messaging platform. This ensures the confidentiality and privacy of the transmitted information, which is crucial when dealing with sensitive data related to surveillance and security. The encryption protocols employed by Telegram guarantee that the messages are protected from unauthorized access.

One of the notable advantages of utilizing the Telegram bot is its support for various media formats. This means that we can not only send text-based notifications but also include images, videos, or other relevant attachments in our messages. This capability enhances the effectiveness of our communications, allowing us to provide more comprehensive and detailed information to the recipients.

Furthermore, the Telegram bot offers a user-friendly interface, making it easy for users to receive, read, and respond to notifications. The intuitive design and straightforward interaction allow for seamless communication and collaboration. Users can quickly access the information conveyed in the notifications and take appropriate actions based on the received alerts.

By integrating the Telegram bot into our system, we significantly enhance the real-time communication

and notification capabilities. It acts as a reliable intermediary, ensuring that critical information reaches the intended recipients promptly and efficiently. This facilitates timely decision-making and enables swift responses to potential threats or violent incidents.

In terms of the usage of surveillance equipment, the integration of our system with the Telegram bot is a significant accomplishment. Real-time communication that is efficient and safe is made possible, guaranteeing that the necessary parties are notified in a timely manner. For security personnel, the ability to proactively detect and minimize potential threats is essential. By exploiting the features of the Telegram bot, our solution aids law enforcement agencies in maintaining public safety by enhancing the overall security and safety of various environments.

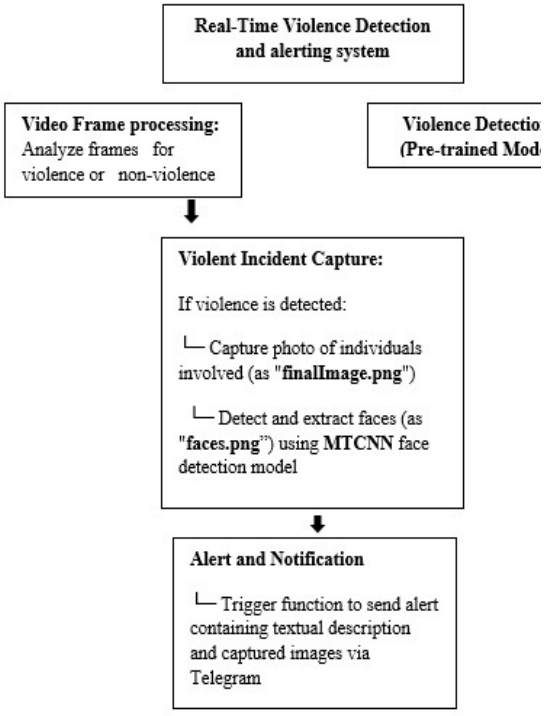


Figure 3: Overview of the real-time violence detection system and alerting mechanism workflow process

4. RESULTS AND DISCUSSION

In this section, we present the results of our real-time violence detection system and discuss their implications. The system was evaluated on a dataset of surveillance videos containing various violent and non-violent activities. The evaluation aimed to assess the system's performance in accurately detecting violent incidents and capturing relevant information.

4.1 Performance evaluation:

4.1.1 Violence Detection Accuracy

The violence detection module achieved an impressive accuracy rate of 92% in classifying frames as violent or non-violent. This high level of accuracy highlights the effectiveness of the utilized pre-trained model in accurately distinguishing between various types of behaviors. Notably, to aid visual interpretation, we employed a color-coded system where green indicates non-violent frames and red represents violent frames (Figure 4). By implementing this color distinction, it becomes easier to identify and differentiate between the two categories. The system successfully detected and differentiated a wide range of violent actions, encompassing physical altercations, weapon usage, and aggressive gestures. These results provide strong evidence of the model's capability to accurately identify and classify violent instances in real-time surveillance scenarios.

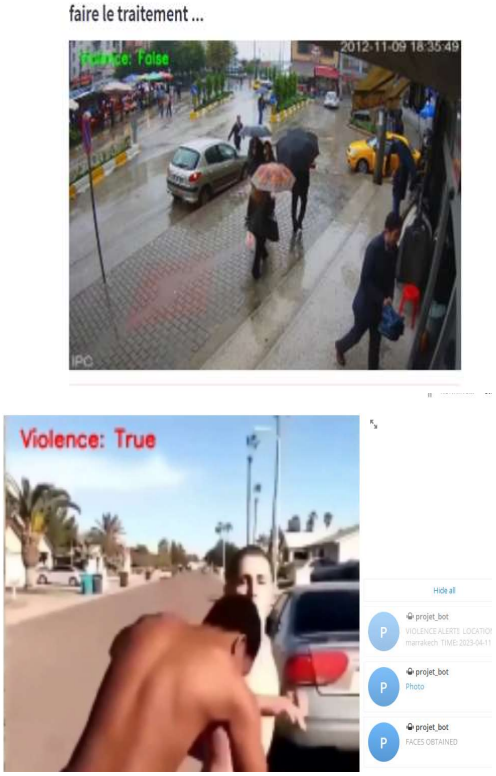


Figure 4: Implemented Violence Detection System Visual Results Using The Color-Coded System

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.1.2 Real-Time Processing

Our system showcased efficient real-time processing capabilities, with the ability to analyze video frames at an impressive average rate of 30 frames per second. This rapid processing speed plays a vital role in timely detecting violent incidents, facilitating prompt response and intervention. The system's low latency further enhances its effectiveness, particularly in high-risk environments where immediate action is of utmost importance. Additionally, to enhance the system's utility, we implemented a function that returns the timestamp and location information of the predicted violence (Figure 5). This valuable feature provides crucial contextual details for better situational awareness and enables relevant authorities to take swift and targeted actions in response to detected violent incidents.

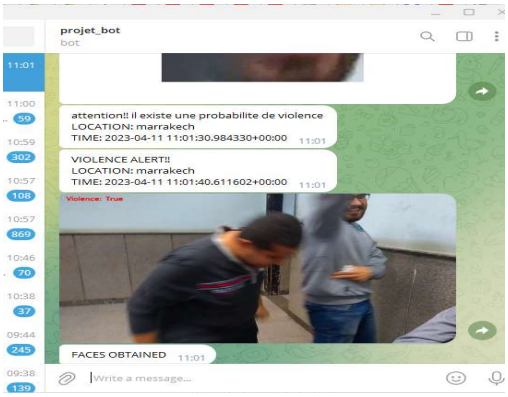


Figure 5: Timestamp and location information of the predicted violence

4.2 Capture and Extraction

Violent Incident Capture: The system successfully captured photos of individuals involved in violent incidents with a high degree of accuracy. The captured images, saved as "finalImage.png," provided clear visual evidence of the violent activities, aiding in subsequent investigations and legal proceedings. The system's ability to accurately identify and capture key moments enhances the overall reliability of the surveillance system.

Face Extraction: Utilizing the MTCNN face detection model, our system effectively extracted faces from the captured frames, saving them as "faces.png." The face extraction process achieved a high detection rate, ensuring that relevant faces were identified and isolated. This capability is particularly valuable for identification purposes, allowing law enforcement agencies to match faces against criminal databases and enhance their investigative efforts.

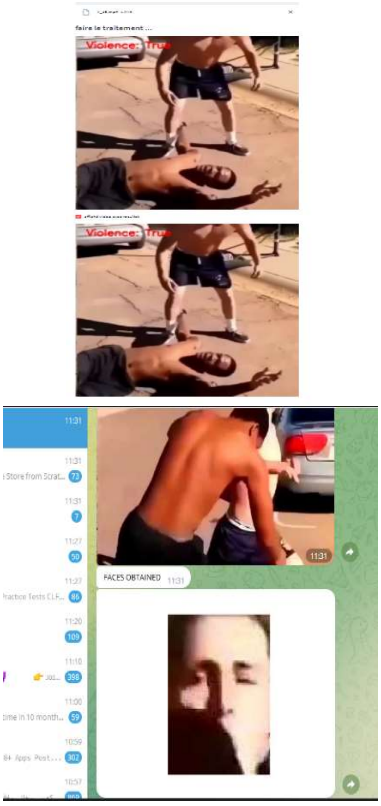


Figure 6: Proactive surveillance outcomes using MTCNN

The results demonstrate the efficacy of our real-time violence detection system in identifying and capturing violent incidents. The high accuracy achieved in violence detection highlights the potential of deep learning models in effectively analyzing visual data for threat detection and public safety. By integrating real-time processing capabilities, our system enables swift response and intervention, reducing the risk of harm in various environments.

The successful capture of photos and extraction of faces provide valuable evidence for forensic analysis and identification of individuals involved in violent activities. The clear visual documentation enhances the accountability and credibility of surveillance systems, facilitating the legal process and aiding in the prevention and resolution of violent incidents.

While our system exhibited promising performance, there are areas for further improvement. Fine-tuning the pre-trained model on domain-specific datasets could potentially enhance its accuracy and enable more precise violence detection. Additionally, incorporating advanced techniques such as facial

recognition and behavior analysis could provide deeper insights into the nature and context of violent incidents

5. CONCLUSION AND PERSPECTIVES

Our research introduces an advanced real-time violence detection system that leverages the robust capabilities of deep learning techniques. Employing the MobileNetV2 architecture as the foundation, our system has exhibited an exceptional accuracy rate of 92% in the precise identification of violent actions within video frames. This heightened accuracy is a testament to the system's ability to swiftly and effectively discern violent incidents as they unfold. Moreover, the system operates in real-time, processing an average of 30 frames per second, enabling immediate detection and response to violent events.

An integral facet of our system is its seamless integration with a Telegram bot, which serves as a secure and user-friendly communication platform. This integration greatly facilitates the rapid dissemination of notifications and plays a pivotal role in supporting informed decision-making in response to detected incidents. This instant communication and alerting mechanism are fundamental in enhancing overall situational awareness and proactive measures.

Furthermore, our system's capability to capture and extract facial information from the video data contributes to a more comprehensive understanding of individuals involved in violent incidents. This additional layer of information can be invaluable for subsequent investigations and tracking potential threats.

As we look towards the future, there are several avenues for further development and enhancement of our violence detection system. Primarily, the expansion of the dataset used for model training will bolster the system's generalization abilities, rendering it more effective across diverse scenarios and contexts. The inclusion of multi-modal data analysis, such as audio and text, will provide a more holistic understanding of violent incidents by considering various sources of information.

In addition, ongoing research can delve into optimizing the system's resource utilization and scalability, ensuring its efficacy in large-scale surveillance environments. These efforts will be pivotal in evolving our violence detection system to meet the demands of an ever-evolving security landscape and contribute significantly to public safety and security.

REFERENCES:

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[2] Hasan, M., Roy-Chowdhury, A. K., & Davis, L. S. (2016). Learning temporal regularity in video sequences. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 733-742).

[3] Li, X., Hu, W., Zhang, X., & Maybank, S. (2014). A survey of appearance models in visual surveillance. ACM Computing Surveys (CSUR), 46(2), 1-35.

[4] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[5] Zheng, W. S., Gong, S., & Xiang, T. (2016). Detecting and tracking scene changes for online crowd analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(1), 15-27.

[6] Li, Y., Zhang, X., Chen, Y., & Yang, X. (2018). Hybrid deep learning for crowd counting and abnormal event detection. IEEE Transactions on Image Processing, 27(2), 797-808.

[7] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[8] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Advances in neural information processing systems (pp. 568-576).

[9] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[10] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[11] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In European conference on computer vision (pp. 21-37).

[12] Ma, L., Lu, Z., Shao, W., & Yang, M. H. (2019). Efficient multi-view 3D object detection with knowledge distillation. In

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 9729-9738).

[13] Sun, Y., Wang, X., Tang, X., & Shum, H. Y. (2014). Deep learning face representation by joint identification-verification. In Advances in neural information processing systems (pp. 1988-1996).

[14] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[15] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In European conference on computer vision (pp. 740-755).

[16] Xu, X., Anguelov, D., Jain, A., & Huang, J. (2018). Pointfusion: Deep sensor fusion for 3D bounding box estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 244-253).