

Intelligent Video Surveillance using Deep-Learning Models

Dr. Anupama C.G¹, Gurram Pavan Kumar² and D. Chaitanya Kumar³

¹⁻³SRM Institute of Science and Technology/Computational Intelligence College of Engineering, Chennai, India
Email: anupamag@srmist.edu.in, gk6241@srmist.edu.in, cd3296@srmist.edu.in

Abstract—Closed-circuit television, or CCTV, is another name for video surveillance. It is a fast-expanding sector that has been around for more than 30 years and has seen many technological advancements. In the modern world, maintaining public safety now requires the use of video surveillance. One can define security. In a number of ways, depending on the situation, including danger of explosion, theft, violence, and so forth. In busy public spaces, "security" can refer to nearly any kind of unusual event. Depending on the user's preferences, intelligent video surveillance records unforeseen activities in homes, offices, and public spaces to provide state-of-the-art smart security. The video surveillance system will actively react to detect actions ahead of time through real-time monitoring and promptly communicate data in the event of an abnormal incident. The main focus is on the application of deep learning techniques to provide motion-activated night vision technology, high-definition picture quality, and tracking of a moving target. Additionally, the system is equipped with automatic audio and visual detection, video recording initiation, and detection of suspicious activities that can activate and alert the systems in any type of weather. The data processing model design aims to employ deep learning technology to visualize data for anomalous activities. It also suggests an intelligent surveillance system to promptly and efficiently identify activities by transmitting a video image and an alert message to the web through realtime processing. Advances in computer vision, especially in deep learning techniques, have created new avenues for these systems to explore, expanding their potential and stimulating new fields of study.

Index Terms— Convolution Neural Networks, Binary long short-term memory, Mobilenet v2, Deep Learning, Artificial Intelligence.

I. INTRODUCTION

A. Project Overview

The primary goal of video surveillance is to create intelligent video surveillance in contrast to conventional passive surveillance. This allows for the recording of unusual human activity and, following analysis, the generation of notifications via alarms, messages, or other means to deter unusual behavior. anomalous behaviors like detecting abandoned components, identifying robberies. The following scenarios necessitate the use of a smart monitoring system: health identification of patients or elder home care (i.e. fall recognition); accidents or traffic rule breaking (i.e., unauthorized U-turns, parking issues and dangerous driving recognition on the road); violence recognition (i.e., hitting, kicking, punching, slapping, shooting in public areas); and fire detection. The use of explosives by terrorists in public areas has increased in danger in recent years. Highly sensitive crowded

public spaces like airports, bus stops, train stations, government buildings, and shopping malls are frequently targeted by terrorist attacks. Terrorists visit these sites and deposit packages containing bombs for use in bomb or nuclear attacks. It is very challenging for security personnel to monitor busy public spaces and spot suspicious activity. Modern technologies fall short in their evaluation of these kinds of attacks in public spaces, which are traditionally captured on camera and subjected to investigations. Using real-time smart video surveillance, public spaces can be safeguarded by identifying and managing abandoned items and sounding an alert to notify security staff to remove the offending items. Thus, developing a completely automated, successful, and efficient smart surveillance system is highly beneficial. Artificial intelligence, machine learning, and deep learning are being applied in a new field: video surveillance. A machine that has artificial intelligence can think like a human. Making assumptions about future data and learning from training data are essential components of machine learning. The deep learning concept is preferred because large data sets and GPU (Graphics Processing Unit) processors are now available. Deep learning-powered intelligent video surveillance systems provide a revolutionary way to monitor and secure environments. These systems can automatically identify and analyze different activities and anomalies in real-time video streams by utilizing the power of deep learning algorithms. 11 Surveillance systems have the ability to function autonomously, effectively, and more accurately than traditional methods that mainly rely on human operators. This project has a lot of potential because it can improve security in a lot of different places, like public areas, transit hubs, and vital infrastructure. It also makes proactive responses to possible threats possible, which raises general security and safety standards. In addition to improving monitoring efficacy, the application of deep learning in video surveillance opens the door to more sophisticated features. Intelligent video surveillance using deep learning offers a range of advantages over traditional methods. These systems can automatically analyze video data, saving time and resources. They detect events in real-time, enabling swift action during security threats. With high accuracy, they recognize specific objects or anomalies, reducing false alarms. Deep learning algorithms are adaptable to changing environments, ensuring reliable performance. By continuously monitoring areas, they enhance security measures and deter threats. While setup costs may be higher, they can be more cost-effective long-term due to reduced manpower and improved efficiency. These systems can scale to cover large areas and integrate advanced features like facial recognition and behavior analysis

B. Deep Learning Models In Video Surveillance

A lightweight convolutional neural network architecture called MobileNetV2 was created especially for embedded and mobile vision applications. To achieve high performance at low computational cost, it makes use of efficient architecture design principles and depth wise separable convolutions. Because MobileNetV2 is so good at real-time object detection, it's a good option for intelligent video surveillance systems where efficiency is critical. TensorFlow or PyTorch as the implementation framework, MobileNetV2 pre-trained weights for transfer learning or training from scratch, OpenCV for image preprocessing and augmentation, and GPU acceleration for faster inference are the software requirements for MobileNetV2 implementation. An alternative to conventional LSTMs, binary Long Short-Term Memory (LSTM) networks employ binary activations and weights. This modification preserves the ability to capture temporal dependencies in sequential data, while requiring a significant reduction in memory and computation. Binary LSTM networks are ideally suited for activity recognition tasks in the context of intelligent video surveillance. These tasks require the networks to analyse video sequences in an efficient manner and identify patterns that indicate different types of activities. TensorFlow or PyTorch as the implementation framework, libraries for feature extraction from video sequences and sequence preprocessing, tools for model optimization and compression like TensorFlow Lite, and GPU acceleration for training and inference are the software requirements for Binary LSTM implementation. Selecting TensorFlow or PyTorch for system implementation depends on the needs of the project and the team's experience. Mobile-compatible hardware should be accessible for deployment, and high-performance GPUs should be available for effective training. To prepare video data for analysis, data preprocessing tools such as libraries for temporal segmentation, frame extraction, and data augmentation are required. In order to ensure efficiency and compatibility, model deployment requires tools like TensorFlow Lite or ONNX for converting and deploying models on mobile devices. Model performance and resource utilization should be maximized by using performance optimization techniques like quantization, pruning, and knowledge distillation. In order to safeguard sensitive video data and guarantee regulatory compliance, security and privacy measures like encryption, authentication, and privacy-preserving strategies should be put into place.

MobileNetV2 is a highly efficient and lightweight convolutional neural network architecture designed for mobile and embedded vision applications. Introduced by Google, it builds upon its predecessor, MobileNetV1, by incorporating a range of enhancements to improve performance while reducing computational complexity. The

core innovation of MobileNetV2 is the use of inverted residual blocks with linear bottlenecks, which significantly enhances the network's ability to capture intricate patterns and features in images with fewer parameters. These blocks use depthwise separable convolutions, ensuring that the network remains computationally efficient. Additionally, the architecture employs a streamlined design, enabling faster inference times and lower power consumption, making it well-suited for real-time applications on devices with limited resources, such as smartphones and IoT devices. Overall, MobileNetV2 strikes a balance between accuracy and efficiency, making it a popular choice for developers needing robust and fast image recognition capabilities on mobile platforms.

C. Related Work

When it comes to predicting scores through deep learning, it's important to go through the literature which provide better ideology to proceed further. This exploration helps to understand the various prediction models, gain valuable insights into their specific methodologies.

Yun-Xia Liu et al. [1] The paper presents an intelligent surveillance method for indoor surveillance using an R-CNN mask, which uses deep learning techniques for accurate target detection and instance segmentation. This approach not only preserves important information from the original material, but also proactively warns against detected unusual events or behavior. Despite the progress, intelligent video surveillance is still a growing field with significant challenges such as management, storage and analysis of the huge amounts of data generated by surveillance systems.

M.A.Saleem Durai et al. [2] The paper reveals that while deep learning has improved intelligent video surveillance, especially in congested areas, there are still challenges such as the demand for large labeled data sets, computational intensity, and the need for real-time video processing. This suggests that future research should aim to improve deep learning algorithms to improve large video data and explore unsupervised and transfer learning methods to minimize labeled data.

Jie Xu1 et al. [3] The paper describes an end-to end video surveillance system adept at object detection and face recognition, engineered to balance accuracy and speed while maintaining an acceptable error rate. Future enhancements could involve integrating advanced video object detection techniques and ensuring the model is trained on high-quality image sequences to align with the high-resolution images captured by surveillance cameras,

Zhenfeng Shao et al.[4] The paper discusses the difficulties of managing large surveillance video data in the era of big data and presents a new method for intelligent processing. This highlights the need for continued research in anomaly detection, spatiotemporal analysis, and system integration to improve this method and its effectiveness.

Chang-Soo Sung et al. [5] Researchers recommend adding deep learning models to distributed servers to improve surveillance and speed and accuracy, emphasizing deep learning and its promise to improve crime control. However, it also points to unsolved challenges such as traffic congestion management, real-time crime detection and algorithmic control in AI-based surveillance systems that require further research.

Vijay Bhanudas et al. [6] The work deals with a video analysis system based on deep learning, which can be used to identify suspicious behavior in surveillance footage and classify the activity as illegal, suspicious or safe. The method, dataset and results are also described. However, it acknowledges issues such as vulnerability to adversarial attacks, false positives, similar challenges and diversity of datasets. The authors call for better tracking of suspects between multiple cameras to improve performance and accuracy

II. METHODOLOGY

Designing It takes a complex team of engineers to design an intelligent video surveillance system that can identify human combat. This technology uses real-time video analysis to detect and categorize human physical disputes. This system is an example of a sophisticated use of artificial intelligence in the security and safety domain.

Frame disassembly: This system's frame disassembly is a crucial component. This is taking individual video stream frames in order to analyze possible human combat situations. The removal of frames serves as the foundation for additional monitoring process analysis. **Elimination of a feature:** Feature elimination is a critical phase in the design process. This method finds similar patterns or characteristics in the video footage that aid in differentiating between peaceful and aggressive behavior. The quality and diversity of the training data are crucial for the success of feature extraction, which frequently calls for the application of classifiers and machine learning algorithms.

Classification: The system classifies data using sophisticated algorithms and machine learning methods. In order to discern between typical behaviour and actual collisions, this method integrates pattern recognition and computer vision. The system has been trained to identify specific gestures, facial expressions, and body language that are linked to physical aggressiveness.

Technique: A complex technique is used in the development of this system. First, a wide range of video footage, including human conflict, is gathered. To enhance the quality of data, preprocessing methods including normalization and framing are applied. As it incorporates a video processing pipeline for real-time input processing and object tracking to preserve consistency and lower false positives, real-time analytics is essential for robustness. Moreover, algorithms for anomaly identification are incorporated to differentiate between appropriate and inappropriate behavior.



Figure 1 System Architectur

The system architecture includes several main components:

Camera Network: High-resolution cameras are strategically placed to ensure adequate coverage and minimize blind spots. **Camera Stream Processing:** Each camera stream is processed by the CPU, often using GPU-accelerated hardware for efficiency. **Video Analysis Module:** This module implements real-time computer vision algorithms, including motion and object detection.

Feature Extraction: Deep learning models such as CNNs are used to extract relevant features from observed individuals.

Battle Detection Algorithm: A special machine learning model is trained to detect battles and uses historical data to distinguish between combat and noncombat scenarios

Integration with Behavioral Analytics: This step integrates the battle detection algorithm with behavioral analytics to minimize false positives.

Storage and Recording: The system stores video streams and metadata, keeping a log for future analysis and review. **User interface:** A user-friendly user interface has been developed for security personnel, which includes live broadcasts, alarm notifications and access to information.

Continuous Improvement: A feedback loop is applied to continuously improve the system and accuracy. **Note.** Several important considerations are important to the system and its design

Privacy and compliance: The system must comply with data protection rules and ethical considerations. **Redundancy:**

Critical components have redundancy for reliability. **Scalability:** The system is designed for easy scaling. **Cyber security:** Strong security measures are in place to prevent unauthorized access and data breaches. **Maintenance:**

A regular maintenance schedule is prepared for the hardware components. Ethical considerations such as privacy and impartiality are crucial. The system must implement privacy protection measures, such as anonymizing individuals not involved in violent activity, and ensure the accuracy of the model and forecasts to be responsible and socially acceptable.

III. IMPLEMENTATION

A. Video source integration

Video source integration is a critical part of developing a smart video surveillance system for violence detection. This process involves accessing and processing the video stream from various sources such as CCTV cameras, IP cameras or recorded video files. The integration must be seamless so that the system can continuously analyze a live video stream or process pre-recorded material. Here's a closer look at video source integration.

Video input formats Consider a variety of video input formats, including live video streams and recorded video files. Make sure your system supports common video formats such as MP4, AVI or MJPEG for compatibility with different sources.

Camera Compatibility Determine the cameras your system is compatible with. This can include standard CCTV cameras, IP cameras or even smart cameras. Make sure the system can handle different camera resolutions and frame rates.

Identification of people Developing a robust human recognition model is an integral part of an intelligent video surveillance system for violence detection. People detection involves identifying and locating people in video frames. Various computer vision techniques and machine learning algorithms can be used for this. Below is a detailed overview of the considerations involved in implementing a people detection model

Choice of models Choose a suitable pre-trained model or design a unique model for human recognition. Popular object detection models include Faster R-CNN (Regional Convolutional Neural Network), YOLO (You Only Look Once), and SSD (Single Shot Multibox Detector).

Training Information Collect a diverse dataset containing annotated images or videos with tagged human expressions. To improve the model and generalization, make sure the dataset is representative of different locations, backgrounds, and lighting conditions.

Data pre-processing: Preprocessing training data by resizing images, normalizing pixel values and augmenting the dataset with transformations such as rotation and translation. Data preprocessing helps the model learn robust features and reduces overfitting.



Figure 2. Detection of Violence



Figure 3. Detection of Non-Violence

Detection of violence: Building an effective violence detection model is an important part of an intelligent video surveillance system. This model is responsible for analyzing human behavior in video frames and distinguishes between violent and non-violent actions.

Above is a detailed overview of the considerations involved in implementing an abuse detection model: Fig2.Detection of violence Fig3. Detection of non-violence Construction of the dataset: Create a well-annotated dataset that includes examples of both violent and nonviolent behavior. The material must be versatile and cover different types of violence, different environments and scenarios.

Data preprocessing: Data preprocessing by extracting important features and normalizing the data. Consider techniques such as frame extraction, optical flow, or spatiotemporal feature extraction to capture movement patterns suggestive of violence

Real time analysis: Real-time analysis is a critical aspect of Intelligent Video Surveillance Systems, especially when it comes to violence detection. It involves the continuous and immediate processing of video streams as they are generated, allowing for quick decision-making and timely response.

Implementing real-time analysis in a surveillance system requires careful consideration of factors such as speed, efficiency, and accuracy. Frame Rate and Speed: The frame rate at which video frames are processed directly affects the system's real-time capability. Higher frame rates result in smoother video, but they also demand more computational resources. Strike a balance between frame rate and processing speed based on the system's requirements.

Parallel Processing: Implement parallel processing techniques to enhance the speed of video analysis. Break down the video stream into smaller segments and process them simultaneously, leveraging multiple processing units or cores.

Hardware Acceleration: Utilize hardware acceleration technologies such as GPUs (Graphics Processing Units) or specialized hardware like TPUs (Tensor Processing Units) to accelerate the computation-intensive tasks involved in video analysis.

This can significantly improve realtime performance **Optimized algorithms:** Select or develop algorithms optimized for real-time processing. This may require sacrificing a certain level of complexity for speed. Algorithms that can achieve a balance between accuracy and efficiency are crucial for real-time monitoring applications.

Sources of Video Testing: Testing a video surveillance system, especially testing a system focused on violence detection, is a critical step to ensure its effectiveness and reliability in real life. Comprehensive testing involves evaluating the system and its performance under various conditions, ensuring detection accuracy and identifying potential weaknesses. Here's a closer look at the various aspects of video source testing.

Versatile composition of datasets: Develop a diverse dataset that reflects real-world scenarios. Include video footage with lighting conditions, camera angles, crowd density and different types of violence to ensure system reliability and scalability.

Abstract and basic fact: Label the dataset with truth-based labels indicating the presence or absence of violence in each frame. This annotated dataset is the basis for training and testing the violence detection model. **Training series vs. Test series:** Divide the dataset into training and test sets. The training set is used to train the model, while the test set evaluates its performance against unseen data.

Ensure that the test set represents the real scenarios that the system will encounter. **Real-time analysis:** Fig 4. **Training vs Testing** Real-time analysis is an important part of smart video surveillance systems, especially when it comes to detecting violence. This involves continuous and immediate processing of video streams as they are generated, enabling quick decision making and quick response. Implementing real-time analytics in a monitoring system requires careful consideration of factors such as speed, efficiency, accuracy.

Mobile network: MobileNet is a set of lightweight convolutional neural network (CNN) architectures designed for use on resource-constrained devices with an emphasis on computational efficiency. MobileNet, launched by Google in 2017, uses depth-separated convolutions, dividing standard convolutions into depth-based and point convolutions. This approach reduces the computational complexity and number of parameters, making it suitable for mobile and embedded applications. The architecture and modular design allow for customization, enabling different compromises between speed and accuracy. MobileNet has been widely used for tasks such as image classification, object detection, and face recognition on mobile devices, which promotes edge computing and the efficient deployment of deep learning models. **LSTM model:** Long Short-Term Memory (LSTM) is a special recurrent neural network (RNN) architecture that solves the challenges of learning long-term dependencies in sequential data. LSTMs, proposed by Hochreiter and Schmidhuber in 1997, have become central to various areas of deep learning. The main innovation is their ability to selectively retain or discard information for long periods of time using the state of the cell, which acts as a stable memory during the sequence. LSTMs with input, forget and output gates control the flow of information at each time step. The input gate controls the addition of new information to the state of the cell, the forget gate controls the removal of unnecessary information, and the output gate determines the next hidden state based on the current input and remembered information. LSTMs are particularly powerful for tasks involving sequential data, such as natural language processing and time series analysis, where capturing long-term dependencies is critical. Their widespread adoption is due to their ability to model complex relationships in sequences, which makes them indispensable for many applications in the field of deep learning.

Each layer of this architecture plays a role in processing and converting the input data (video frames) into a suitable format for classification. Using Time Distributed layers allows the network to process each frame separately, while LSTM captures temporal dependencies between frames. Dense layers act as classifiers of extracted and modified features from previous layers.

IV. RESULTS

An intelligent video surveillance system designed to detect human combat combines advanced computer vision and artificial intelligence technology. It analyzes real-time video streams to identify potential physical conflicts between individuals, using deep learning models to distinguish normal activity from violent behavior. The system recognizes patterns such as aggressive body language and sudden movements, allowing public security to proactively intervene in various environments such as public spaces and traffic intersections. In addition, it provides valuable data insights into combat events, helping to develop effective security protocols and crowd

control strategies. Overall, this vigilant system is an important tool to ensure public order and prevent violence in various situations.

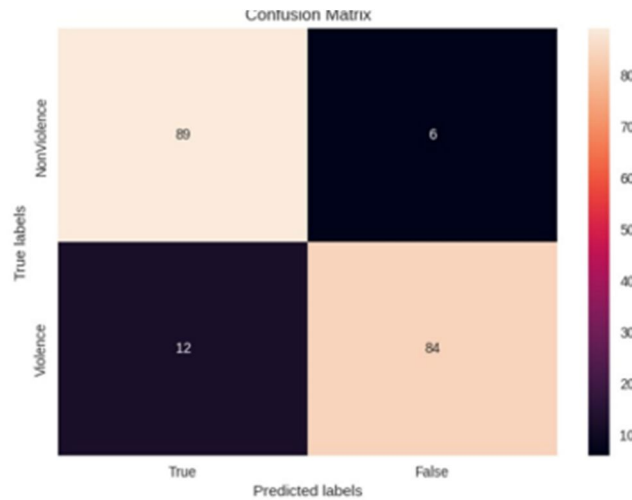


Figure 4 Confusion Matrix

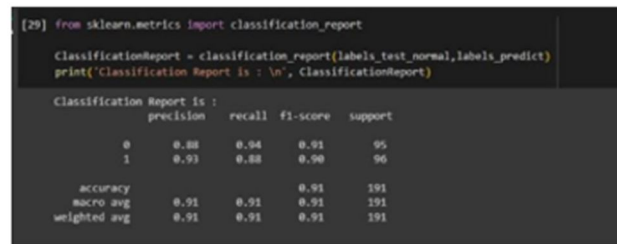


Figure 5 Accuracy Table

The accuracy score function of the sklearn metrics module to evaluate the performance of a predictive model. The function compares the predicted labels (labels_predict) with the actual labels (labels_test_normal) to calculate an accuracy score. The result is printed and the accuracy score of the model is approximately 0.90 (or 90%). This high accuracy indicates that the model's predictions were correct 90% of the time on the given test data, which is generally considered a very good performance depending on the complexity of the task and the accuracy of the baseline.

Non/Violence: our model accurately determined the 'Non/Violence' 89 times (True Negative). Our model slightly given the inaccurate results for 'Violence' 6 times (False Positive). Also in case of the Violence class: our model most accurately predicted 'Violence' 84 times (True Positive). The model incorrectly predicted 'Non/Violence' 12 times (False Negative). The below gradient shows the frequency of predictions, with darker colors indicating a higher number of predictions.

This confusion matrix shows that while the model is relatively accurate, there are still instances where it incorrectly classifies the behavior, which is particularly critical in the context of violence detection where both types of errors may have serious implications. Fig 7. Confusion matrix Fig 8. Evaluation Metrics For class '0', the model has: A precision of 0.88, meaning that when it predicts class '0', it is correct 88% of the time. A recall of 0.94, indicating that it correctly identifies 94% of all actual class '0' instances. An F1- score of 0.91, which is a harmonic mean of precision and recall, suggesting a balance between the two. A support of 95, which is the number of actual occurrences of class '0' in the dataset. For class '1', the model shows: A precision of 0.93, a high likelihood that predictions for class '1' are correct. A recall of 0.88, somewhat lower than class '0', indicating a slightly higher rate of missed actual class '1' instances. An F1-score of 0.90, which is slightly lower than that of class '0'. A support of 96, the number of actual occurrences of class '1'. Overall accuracy is 0.91, which aligns with the weighted average of precision, recall, and F1-score across both classes, indicating a consistent performance across the classes. The 'macro avg' and 'weighted avg' lines report the average metrics, with 'weighted avg' accounting for the support of each class. This model shows a high level of performance across these metrics, suggesting it is well-calibrated for both classes.

V. CONCLUSION

Ultimately In short, developing an intelligent video surveillance system to detect human behavioral violence is a multifaceted and challenging endeavor. Using computer vision techniques and machine learning algorithms, we aim to create a system that can distinguish between violent and non-violent actions in video footage. The project involved the careful collection and pre-processing of various datasets, the selection and integration of appropriate models for human detection and violence classification, and the implementation of an alarm system for real-time response. Rigorous testing and evaluation was performed throughout the process to ensure continuous optimization of the system and accuracy based on performance metrics.

Ethical considerations, including data protection laws and potential bias, were paramount in the development process. As technology evolves, the project emphasizes the importance of responsible and ethical deployment of surveillance systems, striking a delicate balance between security and the protection of the rights and privacy of individuals. The future of a smart video surveillance system to detect violence is promising as the technology continues to evolve. One expansion option is to improve the underlying machine learning models to improve accuracy and reduce false positives, thus providing a more robust and reliable system. Integration with cutting-edge computer vision technologies such as 3D object detection and scene understanding can further enhance the system's capabilities, enabling more nuanced analysis of more complex scenarios. In addition, studying real-time adaptation to dynamic environments and changes in behavior can add an additional layer to the system.

Collaboration with law enforcement and security professionals can lead to system customization to meet specific operational needs and regulatory requirements. In addition, advances in hardware such as high-resolution cameras and edge computing devices can contribute to system efficiency and realtime processing capabilities.

REFERENCES

- [1] Penet, C.; Demarty, C.H.; Gravier, G.; Gros, P. Multimodal information fusion and temporal integration for violence detection in movies. In *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 25–30 March 2012; pp. 2393–2396
- [2] Deniz, O.; Serrano, I.; Bueno, G.; Kim, T.K. Fast violence detection in video. In *Proceedings of the 2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, Lisbon, Portugal, 5–8 January 2014; Volume 2, pp. 478–485.
- [3] Bermejo Nievas, E.; Deniz Suarez, O.; Bueno García, G.; Sukthankar, R. Violence detection in video using computer vision techniques. In *International Conference on Computer analysis of Images and Patterns*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 332–339
- [4] De Souza, F.D.; Chavez, G.C.; do Valle, E.A., Jr.; Araújo, A.D.A. Violence detection in video using spatio-temporal features. In *Proceedings of the 2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images*, Gramado, Brazil, 30 August–3 September 2010; pp. 224–230.
- [5] Xu, L.; Gong, C.; Yang, J.; Wu, Q.; Yao, L. Violent video detection based on MoSIFT feature and sparse coding. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 4–9 May 2014; pp. 3538–3542.
- [6] Arceda, V.M.; Fabián, K.F.; Gutiérrez, J.C. *Real Time Violence Detection in Video*; IET: Talca, Chile, 2016
- [7] Das, S.; Sarker, A.; Mahmud, T. Violence detection from videos using hog features. In *Proceedings of the 2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, Khulna, Bangladesh, 20–22 December 2019;
- [8] Gracia, I.S.; Suarez, O.D.; Garcia, G.B.; Kim, T.K. Fast fight detection. *PLoS ONE* 2015, 10, e0120448