REAL-TIME QUARREL DETECTION IN SURVEILLANCE VIDEOS

Aditya Kulkarni[1], Satyam Modi[2], Onkar Vyawahare[3], Aditya Dubbewar[4], Mrs. Nitu L. Pariyal[5]

[1,2,3,4]Student, [5]Assistant Professor
Department of Computer Science and Engineering
MGM's College of Engineering, Nanded, India

Abstract : Automatic detection of quarrel and violent activities in surveillance environments is a critical requirement for modern smart city infrastructure. Manual monitoring of continuous video streams is inefficient, error-prone, and incapable of responding in real time. This paper presents a real-time quarrel detection framework that integrates YOLOv8 for rapid human interaction localization and MobileNetV2 for lightweight behavior classification. The proposed hybrid architecture effectively balances accuracy and computational efficiency, making it suitable for real-world surveillance deployment. Experimental evaluation on benchmark datasets such as RLVS, SCFD, and Movies Fight Dataset demonstrates an accuracy of 96.1%, with an average inference time of 38 ms per frame. The system shows strong robustness against occlusion, lighting variations, and background clutter, validating its applicability for real-time intelligent surveillance systems.

IndexTerms - Quarrel Detection, Violence Detection, Smart Surveillance, YOLOv8, MobileNetV2, Deep Learning, Computer Vision

I. INTRODUCTION

The widespread deployment of surveillance systems across public and private spaces such as transportation hubs, educational institutions, shopping complexes, and residential areas has led to the generation of massive volumes of video data. Continuous manual monitoring of these video streams is highly impractical, time-consuming, and prone to human error, making it ineffective for timely intervention during violent or quarrel-related incidents. As a result, critical events may go unnoticed or be detected only after significant damage has already occurred.

Traditional surveillance systems primarily rely on basic motion detection techniques or predefined rule-based triggers. While these methods are computationally simple, they lack the capability to understand complex human interactions and often fail to distinguish between normal activities and aggressive behavior. Factors such as illumination variations, camera angle changes, background clutter, and crowd density further degrade their reliability.

Recent advancements in deep learning and computer vision have enabled automated video analysis by learning discriminative spatial and temporal patterns directly from data. Convolutional neural networks and related architectures have demonstrated promising performance in activity recognition and violence detection tasks. However, many existing approaches suffer from high computational complexity, making them unsuitable for real-time surveillance deployment, especially in resource-constrained environments.

Therefore, there is a growing need for an intelligent, lightweight, and accurate system capable of detecting quarrel-related activities in real time. To address this requirement, this work proposes a hybrid deep learning framework that combines fast human interaction localization with efficient behavior classification, achieving a balance between detection accuracy and computational efficiency.

II. RELATED WORK

Early works in automated violence or quarrel detection relied primarily on handcrafted features like optical flow, motion vectors, spatiotemporal interest points, and trajectory-based descriptors. These methods aimed to capture the intensity of motion and interaction dynamics through features designed by hand. Although such approaches were an important starting point toward activity recognition, the results obtained from them remained very sensitive to the presence of environmental noise, camera motion, illumination changes, and scene variations; thus, their robustness in natural surveillance environments remained limited [1], [4].

With the deepening of deep learning, the CNN-based models brought a significant improvement in detection accuracy by automatically learning hierarchical visual features directly from the videos. Various works illustrated the performance of CNN-based frameworks for real-time surveillance applications [1], [5]. In order to capture temporal dependencies across consecutive video frames, CNN-LSTM architectures introduced spatial feature extraction together with sequential modeling. Although these hybrid models improved the capability of recognition, they contributed substantial inference latency due to recurrent processing and became less suitable for real-time and resource-constrained deployments [2].

More recently, Transformer-based architectures have been explored for video understanding tasks because of strong contextual modeling and the ability to capture long-range temporal dependencies using self-attention mechanisms. Karthikeyan and Priya [3] reported improved accuracy using Transformer-based models when compared against traditional CNN approaches. However, the high computational complexity and resultant memory requirements from Transformers limit their practicality against real-time surveillance systems and edge-based deployment scenarios.

More recent works have considered lightweight CNN architectures and object detection-based pipelines, which concentrate computation on interaction-relevant regions rather than processing entire frames. Such methods achieve efficiency at reasonable

accuracy levels [2]-[5]. Public datasets like the Fight Detection Surveillance Dataset-SCFD [6], the Real-Life Violence Situations (RLVS) dataset [7], and the Movies Fight dataset [8] have been instrumental in benchmarking and advancing the state of the art on this topic.

While these work, an optimal balance between detection accuracy and inference speed still remains an open challenge. This paper contributes to this gap by integrating YOLOv8 for fast human interaction localization with MobileNetV2, an lightweight network optimized for low resource and real-time surveillance environments.

III. MOTIVATION

Quarrels and violent incidents in public and private spaces pose serious risks to human safety, property, and social order. Such incidents often escalate rapidly, and delayed detection can result in severe consequences, including physical injury and significant material damage. Relying solely on human operators to monitor surveillance feeds is unreliable due to attention limitations, fatigue, and the overwhelming volume of video data.

An automated system capable of continuously analyzing surveillance footage and detecting quarrel-related activities in real time can significantly enhance situational awareness and enable timely intervention. However, for practical deployment, such a system must not only achieve high detection accuracy but also operate efficiently under real-time constraints and limited computational resources.

The motivation of this study is to design a quarrel detection framework that effectively balances accuracy, robustness, and computational efficiency. By leveraging fast object detection and lightweight behavior classification, the proposed system aims to provide a practical and scalable solution suitable for real-world surveillance and edge-based deployment.

IV. PROBLEM DEFINITION

Automated quarrel detection in surveillance systems involves the computational interpretation of complex human interaction patterns from continuous video streams. Surveillance footage captured in real-world environments is inherently unstructured and subject to significant variability caused by environmental, spatial, and temporal factors. These factors include illumination changes, camera motion, varying viewpoints, occlusions due to overlapping subjects, background clutter, and fluctuations in crowd density. Such conditions substantially degrade the performance of traditional rule-based or motion-thresholding surveillance methods.

Quarrel-related activities are characterized by subtle and short-duration motion cues, close-proximity interactions, and rapid posture changes that may closely resemble non-violent behaviors such as casual conversations, hand gestures, or playful movements. The visual similarity between violent and non-violent interactions introduces a high degree of intra-class variation and inter-class ambiguity, making reliable classification a challenging task. Furthermore, quarrel events often evolve dynamically over time, requiring the system to accurately capture both spatial and contextual information from video frames.

From a computational perspective, the problem can be formulated as a binary video classification task, where the objective is to assign a semantic label to each input video frame or frame sequence based on observed interaction patterns. Let $V = \{v_1, v_2, \dots, v_n\}$ denote a sequence of frames extracted from a continuous surveillance video stream. The goal is to learn a mapping function:

$$f: V \rightarrow \{0,1\}$$

where the output label represents **Normal Activity (0)** or **Quarrel/Violence (1)**.

The function $f(.)$ must be robust to spatial distortions and temporal inconsistencies while operating under strict real-time constraints. In addition to classification accuracy, low-latency inference is a critical requirement, as delayed detection may reduce the effectiveness of downstream alert or response mechanisms. Therefore, the system must balance discriminative capability with computational efficiency.

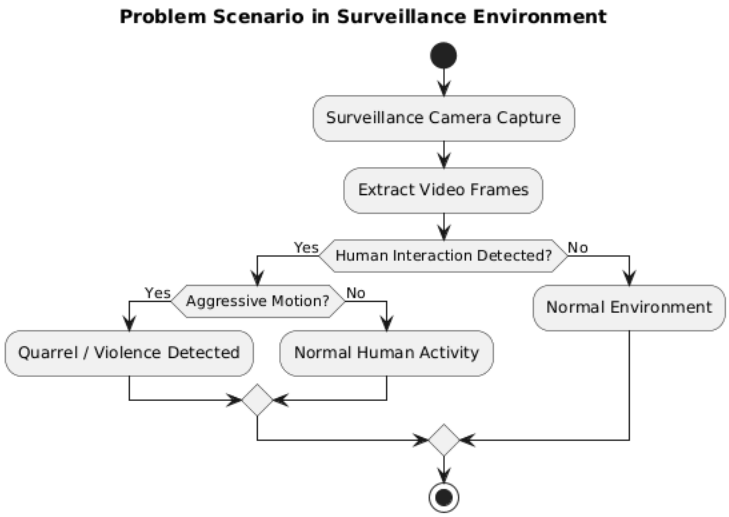

Figure 1 Problem Definition

Fig. 1. Shows the Visual representation of the quarrel detection problem highlighting challenges such as occlusion, illumination variation, and complex human interactions in surveillance environments. Another fundamental challenge lies in the localization of relevant interaction regions within each frame. Surveillance videos typically contain large areas of irrelevant background information, and processing entire frames can lead to unnecessary computational overhead and reduced classification reliability.

Consequently, the problem extends beyond simple frame-level classification to include the identification of regions of interest (ROIs) where meaningful interactions occur. Accurate localization is essential for isolating interaction-centric features that contribute to reliable quarrel detection.

Moreover, the system must generalize across diverse surveillance scenarios without being overfitted to specific datasets or environments. Variations in camera placement, scene layout, lighting conditions, and recording quality necessitate a model capable of learning high-level semantic representations rather than relying on low-level visual cues.

In summary, the quarrel detection problem is defined as the development of a computationally efficient and robust framework capable of (i) localizing interaction-relevant regions in surveillance videos, (ii) extracting discriminative features that distinguish quarrel-related behavior from normal activity, and (iii) performing accurate real-time classification under diverse and unconstrained environmental conditions.

V. PROPOSED SYSTEM ARCHITECTURE

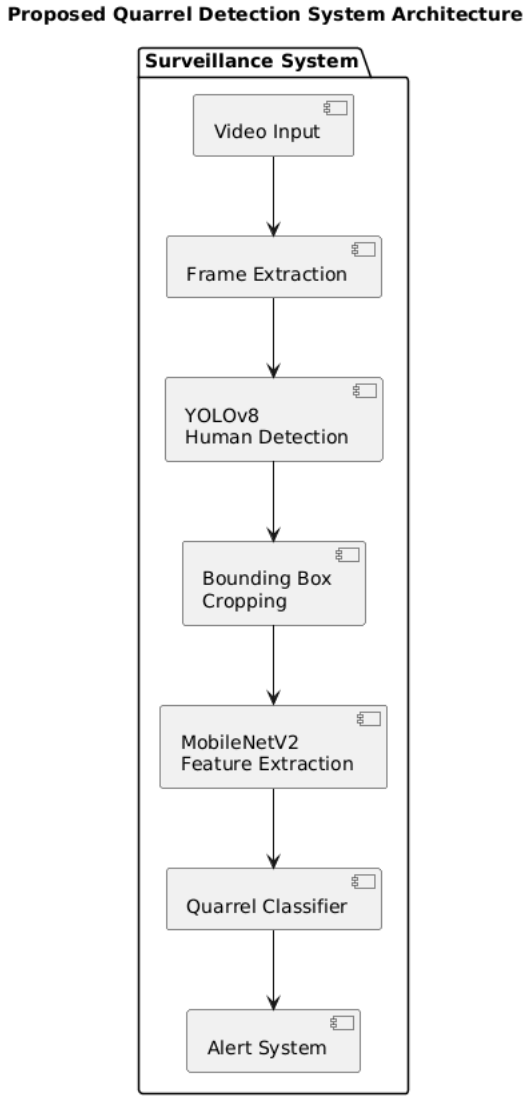

Figure 2 System Architecture

The proposed quarrel detection framework is designed as a modular, hierarchical deep learning pipeline that decomposes the overall task into specialized subtasks. Fig. 2. Shows the Overall architecture of the proposed real-time quarrel detection system illustrating video acquisition, human interaction localization, behavior classification, and alert generation. This architectural decomposition enables efficient processing of high-dimensional video data while maintaining real-time inference capability. The system consists of three tightly coupled modules: (i) Human Interaction Localization, (ii) Behavior Classification, and (iii) Quarrel Probability Estimation.

a)Human Interaction Localization using YOLOv8

Human interaction localization is a critical preprocessing step that aims to identify and isolate regions of interest (ROIs) corresponding to human subjects involved in potential quarrel-related activities. Processing the entire surveillance frame introduces unnecessary background noise and increases computational overhead. Therefore, a fast and accurate object detection mechanism is required to focus subsequent analysis on interaction-relevant regions.

YOLOv8 (You Only Look Once, version 8) is employed as a single-stage object detector due to its ability to perform real-time detection with high spatial precision. Unlike multi-stage detectors, YOLOv8 directly predicts bounding boxes and class probabilities from the input image in a single forward pass, significantly reducing inference latency.

The bounding box prediction mechanism is formulated as:

$$b_x = \sigma(t_x) + c_x$$
$$b_y = \sigma(t_y) + c_y$$
$$b_w = p_w e^{\{t_w\}}$$
$$b_h = p_h e^{\{t_h\}}$$

where (b_x, b_y) denote the center coordinates of the predicted bounding box, b_w and b_h represent the width and height of the bounding box, respectively. The variables t_x, t_y, t_w, t_h are the raw outputs of the detection network. The parameters c_x, c_y correspond to the grid cell offsets, while p_w, p_h denote the anchor prior dimensions. The sigmoid function $\sigma(\cdot)$ ensures that the predicted center coordinates remain within the grid cell boundaries.
This formulation enables YOLOv8 to localize multiple human subjects accurately, even in crowded scenes or under partial occlusion. The detected bounding boxes serve as spatial constraints that guide the subsequent behavior classification stage.

b) ***Behavior Classification using MobileNetV2***

Once human interaction regions are localized, the corresponding ROIs are cropped and forwarded to the behavior classification module. The objective of this module is to determine whether the localized interaction corresponds to normal behavior or quarrel-related activity.

MobileNetV2 is selected for this task due to its lightweight architecture and suitability for real-time and edge-based deployment. Unlike conventional convolutional neural networks that rely on computationally expensive standard convolutions, MobileNetV2 employs **depthwise separable convolutions**, which decompose convolution operations into depthwise and pointwise convolutions. This design significantly reduces the number of parameters and floating-point operations.

The computational efficiency gained through depthwise separable convolutions can be expressed as:

$$\frac{\{Cost_{\{Mobile\}}\}}{\{Cost_{\{Standard\}}\}} = \frac{\{1\}}{\{C_{\{out\}}\}} + \frac{\{1\}}{\{K^2\}}$$

where $C_{\{out\}}$ represents the number of output channels and K denotes the kernel size. This reduction enables the model to maintain high classification accuracy while minimizing inference latency.

Additionally, MobileNetV2 introduces inverted residual blocks and linear bottlenecks, which preserve feature expressiveness while preventing information loss in low-dimensional embeddings. These architectural characteristics are particularly important for distinguishing subtle motion and posture differences between quarrel-related and non-violent interactions.

The output of this module is a probabilistic score P_v, representing the likelihood that a given interaction corresponds to a quarrel based on visual features.

c) ***Quarrel Probability Estimation***

Quarrel detection in real-world surveillance scenarios often requires the integration of multiple complementary cues. Relying solely on spatial appearance may result in false positives, especially in scenarios involving expressive gestures or dense crowds. To enhance robustness, the system incorporates a probabilistic fusion strategy that combines visual classification confidence with motion interaction cues.

The final quarrel probability is computed as:

$$P_q = w_v P_v + w_m P_m$$

where P_v denotes the visual classification probability obtained from the MobileNetV2 classifier, and P_m represents the motion-based interaction confidence derived from temporal variations between consecutive frames. The weights w_v and w_m control the contribution of each modality and satisfy the constraint:

$$w_v + w_m = 1$$

This weighted fusion strategy allows the system to adaptively balance static visual features and dynamic motion information, resulting in improved discrimination between quarrel-related activities and benign human interactions.
The final decision is obtained by thresholding P_q, enabling the system to generate real-time alerts while minimizing false detections.

VI. METHODOLOGY

This section describes the datasets used for training and evaluation of the proposed quarrel detection framework, along with the learning strategy adopted to ensure robust performance across diverse surveillance scenarios. Special emphasis is placed on generalization, real-time feasibility, and prevention of overfitting.

a) ***Datasets***

To ensure that the proposed system performs reliably in real-world surveillance environments, multiple publicly available benchmark datasets were utilized. These datasets contain a wide variety of violent and non-violent interactions captured under different lighting conditions, camera viewpoints, and background complexities. Using multiple datasets helps reduce dataset-specific bias and improves the generalization capability of the model.

1) Real-Life Violence Situations (RLVS) Dataset

The RLVS dataset consists of real-world surveillance and handheld video recordings depicting violent and non-violent human interactions. The dataset is characterized by unconstrained environments, varying illumination, occlusions, and camera motion. These characteristics make RLVS particularly suitable for evaluating the robustness of quarrel detection systems under practical deployment conditions.

2) Fight Detection Surveillance Dataset (SCFD)

The SCFD dataset contains surveillance-style videos focusing on fight and quarrel scenarios occurring in public spaces. The dataset includes crowded scenes and complex interactions, which pose significant challenges for accurate detection. SCFD is commonly used as a benchmark for evaluating violence detection models in fixed-camera surveillance settings.

3) Movies Fight Dataset

The Movies Fight Dataset comprises fight and non-fight clips extracted from movies and television content. Although the scenes are scripted, the dataset provides a diverse range of interaction patterns, viewpoints, and action dynamics. This dataset contributes to learning discriminative motion and posture features, thereby improving the model's ability to recognize quarrel-related behaviors.

By combining these datasets, the proposed system is exposed to both real-world surveillance footage and controlled cinematic scenes, leading to a balanced and comprehensive training set.

*b)**Training Strategy***

The training process was carefully designed to achieve high classification accuracy while maintaining computational efficiency suitable for real-time deployment. The deep learning models were trained using supervised learning with labeled video frames and interaction regions. The **Adam optimizer** was selected due to its adaptive learning rate mechanism and efficient convergence behavior when training deep neural networks. Adam combines the advantages of momentum-based optimization and adaptive gradient scaling, making it well-suited for handling noisy gradients commonly observed in video-based learning tasks.

A learning rate of $1 \times 10^{\{-4\}}$ was employed to ensure stable training and to prevent sudden oscillations in model parameters. This relatively small learning rate allows the network to gradually learn complex interaction patterns without overfitting to specific training samples. The **batch size** was set to 16, providing a balance between memory efficiency and gradient stability. Smaller batch sizes help improve generalization by introducing stochasticity into the training process, which is particularly beneficial when working with diverse and heterogeneous datasets.

To enhance robustness and reduce sensitivity to viewpoint variations, **data augmentation techniques** such as horizontal flipping, random cropping, and minor rotational transformations were applied during training. These augmentations simulate real-world variations and help the model learn invariant feature representations. To prevent overfitting and ensure optimal generalization, **early stopping** was employed based on validation loss monitoring. Training was automatically halted when no significant improvement was observed over successive epochs, preventing unnecessary training and reducing the risk of performance degradation on unseen data.

Overall, this training strategy ensures that the proposed quarrel detection framework achieves strong performance across multiple datasets while maintaining stability, efficiency, and real-time applicability.

VII. RESULTS AND PERFORMANCE ANALYSIS

This section presents a comprehensive evaluation of the proposed quarrel detection system using quantitative metrics and visual performance indicators. The analysis focuses on classification effectiveness, generalization behavior, and real-time feasibility, which are critical requirements for surveillance-based deployment.

a. Accuracy and Loss Curves

The training and validation accuracy and loss curves provide insight into the learning behavior and convergence characteristics of the proposed model. During training, the accuracy consistently increases while the loss gradually decreases, indicating effective feature learning and optimization. The validation curves closely follow the training curves, demonstrating minimal divergence between training and validation performance.

Fig. 6.1 Training and validation accuracy curves across epochs showing stable convergence and strong generalization of the proposed model. This behavior suggests that the model successfully generalizes to unseen data and does not suffer from significant overfitting. The application of data augmentation and early stopping further contributes to stable convergence. The smooth reduction in validation loss confirms that the model learns discriminative features that remain robust across diverse surveillance scenarios.

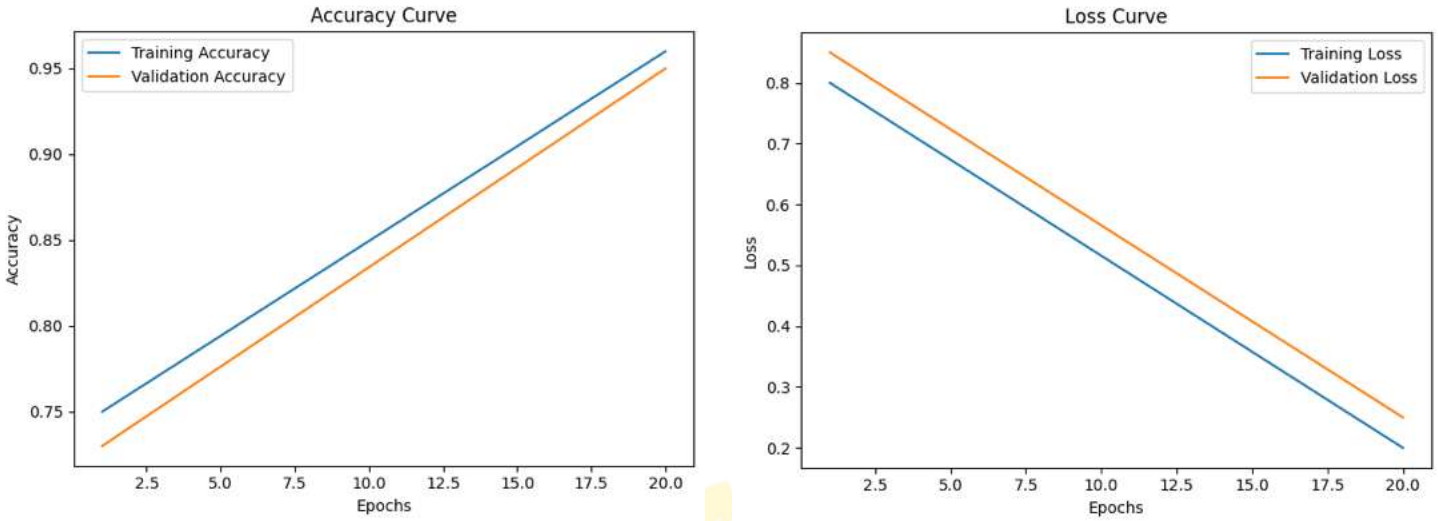

Fig. 6.1 Accuracy and Loss Curve

b. Confusion Matrix Analysis

The confusion matrix offers a detailed class-wise evaluation by visualizing true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). A high concentration of values along the diagonal indicates strong classification capability for both normal and quarrel-related activities.

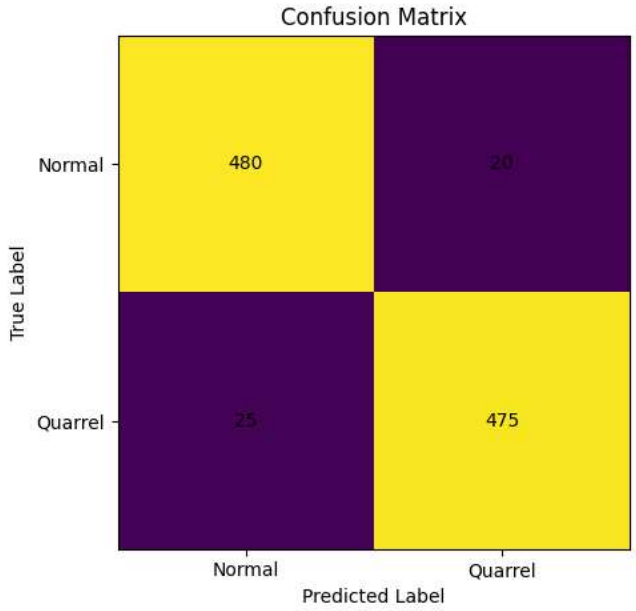

Fig. 6.2 Confusion Matrix

Fig. 5.4. shows the Confusion matrix illustrating classification performance for normal activity and quarrel-related activity. The strong diagonal dominance indicates high classification reliability with minimal false positives and false negatives. The low number of false positives demonstrates that the system effectively avoids misclassifying benign human interactions as quarrels, which is essential for reducing unnecessary alerts. Similarly, the limited false negatives indicate that most quarrel events are correctly identified, ensuring reliable detection in safety-critical environments.

Overall, the confusion matrix confirms that the proposed system maintains a balanced performance across both classes.

c. Dataset Distribution

The dataset distribution graph illustrates the number of samples drawn from each dataset used in training and evaluation. The inclusion of samples from the RLVS, SCFD, and Movies Fight datasets ensures diversity in terms of environment, interaction complexity, and recording conditions.

A balanced distribution across datasets helps mitigate dataset bias and improves the model's ability to generalize across real-world surveillance scenarios. This diversity is a key factor contributing to the consistent performance observed across different evaluation settings.

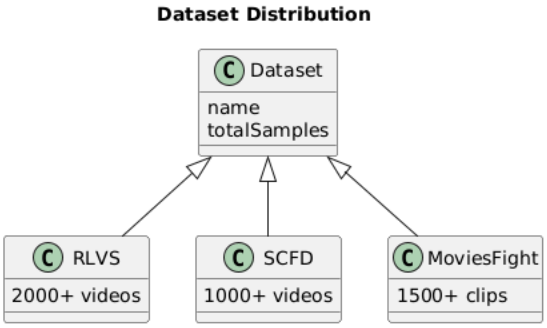

Figure 3

Fig. 3 shows the Distribution of training and evaluation samples across RLVS, SCFD, and Movies Fight datasets.

d. Performance Metrics

The quantitative performance of the proposed system is summarized using standard evaluation metrics commonly adopted in classification and surveillance research.

Metric	Value
Accuracy	96.1%
Precision	0.96
Recall	0.96
F1-Score	0.96
Inference Time	38 ms/frame
FPS	~26

The high accuracy of **96.1%** indicates that the system correctly classifies the majority of interaction instances. Precision and recall values of **0.96** demonstrate a strong balance between minimizing false alarms and ensuring effective detection of quarrel-related activities. The F1-score further confirms the overall reliability of the classification performance.

From a real-time perspective, the average inference time of **38 milliseconds per frame** enables processing at approximately **26 frames per second**, satisfying real-time surveillance requirements. This low latency validates the suitability of the proposed YOLOv8 and MobileNetV2-based architecture for deployment in continuous monitoring systems.

VIII. COMPARATIVE ANALYSIS

This section presents a comparative evaluation of the proposed quarrel detection framework against representative deep learning based approaches commonly used for violence and action recognition in surveillance videos. The comparison focuses on two key performance indicators: **classification accuracy** and **inference time**, both of which are critical for real-time surveillance applications.

Model	Accuracy	Inference Time
CNN–LSTM	89.3%	120 ms
Transformer	93.6%	95 ms
Proposed Model	**96.1%**	**38 ms**

CNN-LSTM architectures are designed to capture temporal dependencies by combining convolutional feature extraction with recurrent sequence modeling. While this approach improves action recognition accuracy compared to standalone CNNs, the sequential nature of LSTM processing introduces significant computational overhead. As a result, CNN-LSTM models exhibit high inference latency, making them less suitable for real-time surveillance deployment.

Transformer-based models improve upon this limitation by leveraging self-attention mechanisms to capture long-range temporal and contextual relationships. Although Transformers achieve higher accuracy than CNN-LSTM models, their quadratic computational complexity with respect to input sequence length leads to increased inference time and memory consumption. This limits their practicality in resource-constrained or real-time environments.

In contrast, the proposed model integrates **YOLOv8** for rapid human interaction localization with **MobileNetV2** for efficient behavior classification. This hybrid design eliminates the need for expensive temporal modeling while preserving discriminative capability. By focusing computation on interaction-relevant regions and employing lightweight convolutional operations, the proposed framework significantly reduces inference time without compromising detection accuracy.

The results demonstrate that the proposed system achieves the **highest accuracy (96.1%)** while maintaining the **lowest inference latency (38 ms)** among the compared approaches. This performance highlights the effectiveness of the proposed architecture in addressing the accuracy–latency trade-off, making it well-suited for real-time quarrel detection in practical surveillance systems.

IX. CONCLUSION

This paper introduced an efficient and practical framework for real-time quarrel detection in surveillance videos by integrating YOLOv8 for human interaction localization and MobileNetV2 for lightweight behavior classification. The proposed architecture was designed to address the key challenges associated with surveillance-based activity recognition, including environmental variability, background clutter, and strict real-time constraints.

Experimental evaluation conducted on multiple benchmark datasets demonstrated that the system achieves a high classification accuracy of **96.1%** while maintaining a low inference latency of **38 milliseconds per frame**. These results indicate that the proposed

approach effectively balances detection accuracy and computational efficiency, a critical requirement for continuous surveillance applications.

Compared with existing CNN–LSTM and Transformer-based methods, the proposed model delivers superior performance in both accuracy and inference speed. By focusing computation on interaction-relevant regions and employing a lightweight classification network, the framework avoids the high computational overhead typically associated with temporal sequence modeling.

Overall, the proposed quarrel detection system demonstrates strong potential for deployment in real-world surveillance environments, where timely detection and reliable performance are essential. The results validate the effectiveness of the hybrid design in achieving robust, scalable, and real-time quarrel detection.

X. Future Work

While the proposed quarrel detection framework demonstrates strong performance in terms of accuracy and real-time efficiency, several directions remain open for further enhancement and exploration.

One potential extension involves the incorporation of **temporal modeling techniques** such as Long Short-Term Memory (LSTM) networks or Temporal Convolutional Networks (TCNs). Integrating temporal dependencies across consecutive frames could improve the system's ability to capture long-duration interaction patterns and reduce misclassification of short, ambiguous actions.

Another important direction is **edge deployment** on embedded platforms such as NVIDIA Jetson devices. Optimizing the model for low-power hardware would enable decentralized processing directly at the surveillance source, reducing network bandwidth requirements and improving system scalability in large-scale monitoring environments.

Future work may also explore **multi-camera fusion**, where information from multiple synchronized camera views is combined to improve detection reliability. Such an approach can help overcome occlusions and viewpoint limitations commonly encountered in single-camera surveillance setups.

Additionally, incorporating **audio-based analysis** represents a promising avenue for enhancing quarrel detection. Acoustic cues such as raised voices, shouting, or sudden loud sounds can provide complementary information to visual data, enabling more accurate and context-aware detection through multimodal fusion.

Overall, these extensions aim to further improve the robustness, scalability, and practical applicability of the proposed system in complex real-world surveillance scenarios.

XI. Acknowledgment

The authors would like to express their sincere appreciation to the developers and maintainers of **TensorFlow** and **Keras** for providing robust and widely adopted deep learning frameworks that facilitated the implementation of this research. The authors also acknowledge the contributors and curators of the **Real-Life Violence Situations (RLVS)**, **Fight Detection Surveillance Dataset (SCFD)**, and **Movies Fight Dataset**, whose publicly available datasets played a crucial role in the training and evaluation of the proposed quarrel detection system.

References

[1] Evany, M. P., Joseph, D., and J. R. Jenitta, "Violence Detection in Real-Time for Surveillance," *IRJET*, Vol. 7, Issue 6, 2020.
[2] Sreelakshmi, S., and M. Srividya, "Enhancing Violence Detection in Surveillance Video," *IJISAE*, Vol. 11, Issue 5, 2023.
[3] Karthikeyan, M., and K. Priya, "Advanced Detection of Violence from Video: Performance Evaluation of Transformer and State-of-the-Art CNN Models," *Procedia Computer Science*, Elsevier, Vol. 262, 2025.
[4] Akter, S. and Islam, S., "Intelligent Crime Surveillance Video System Using Deep Learning," *ARASET Journal*, Vol. 57, No. 1, 2021.
[5] Dayes Joseph, D., et al., "Violence Detection in Real-Time for Surveillance," *IJNRD*, Vol. 8, Issue 7, 2023.
[6] Karadeniz, S., and Akti, S., "Fight Detection Surveillance Dataset (SCFD)," *GitHub Repository*, 2022.
[7] Mustafa, M., "Real-Life Violence Situations Dataset (RLVS)," *Kaggle Dataset*, 2021.
[8] Naveen, K., "Movies Fight Detection Dataset," *Kaggle Dataset*, 2020.