Security and Surveillance: Human and Suspicious Object Detection Using YOLO11

Nishchay M. Gaonkar
School of Computer Science & Engineering
KLE Technological University-Hubballi
01fe22bcs258@kletech.ac.in

Chinmay S. Avaradi
School of Computer Science & Engineering
KLE Technological University-Hubballi
01fe22bcs306@kletech.ac.in

Sushmita Math
School of Computer Science & Engineering
KLE Technological University-Hubballi
01fe22bcs184@kletech.ac.in

Harshitaa Katwa
School of Computer Science & Engineering
KLE Technological University-Hubballi
01fe22bcs257@kletech.ac.in

Shashank Hegde
School of Computer Science & Engineering
KLE Technological University-Hubballi
shashankhegde420@gmail.com

***Abstract*—Crime prevention remains a persistent challenge, particularly in environments with high population density and limited policing resources. This paper introduces an enhanced object detection model, YOLO11, designed to address these challenges by enabling the real-time detection of human activities and suspicious objects in surveillance systems. YOLO11 incorporates innovative architectural features such as the C3K2 block, the SPFF module, and the C2PSA block, optimizing both speed and accuracy in object detection tasks. Evaluated on a robust validation dataset, the model demonstrates significant performance, achieving a mean average precision (mAP) of 0.879 at IoU threshold 0.5 and a pre-processing-inference time of 11.9 ms per image. These advancements position YOLO11 as a pivotal tool in modern surveillance systems, enhancing the detection of normal actions, suspicious activities, and weapons, thereby supporting proactive crime prevention.**

Keywords: Crime Prevention, YOLO11, Object Detection, Surveillance, Real-Time Monitoring

I. Introduction

Crime has always been a social issue since its forms and patterns keep changing with time. Street crimes like theft, assault, and vandalism often terrorize the urban communities and thrive in areas of high population density with limited resources. Such problems erode the effectiveness of conventional policing because the scale of incidents is so vast, and human monitoring has so many limitations that many crimes remain unobserved and continue to haunt communities [[4], [6]].

Video surveillance systems, especially closed-circuit television (CCTV), have become a critical tool for crime prevention and resolution in recent years.Referred to as the "third eye," CCTV systems provide continuous monitoring and have proven instrumental in capturing incidents that might otherwise escape human observation. [[1], [8]]. For example, studies show that street crimes have measurably decreased in urban areas with high surveillance infrastructure [1]. Cities such as Delhi, that have over 275,000 CCTV cameras installed [12] have received acclaim for early adoption of the technology. Nonetheless, despite all the developments there are many hurdles to cross. Human monitors often fail to sift through the voluminous video feed and react in a delayed manner or even miss out on crucial situations [7]. Furthermore, the installation and maintenance of CCTV cameras is cost-intensive, which proves a hindrance in underdeveloped areas [9].

With technological advancements, methods of preventing crime also evolve. The newly emerging solutions include AI-powered video analytics and predictive policing, which are changing the surveillance landscape[[5], [11]]. They help in real-time anomaly detection, identification of potential threats, and anticipation of crime-prone areas. This is a much stronger advantage than traditional systems [7]. Furthermore, community-driven surveillance through smartphones and social media platforms helps in faster reporting and broader coverage of incidents [8].

Despite these innovations, the wide-scale implementation of surveillance technologies brings forth ethical concerns, such as privacy and data misuse [10]. There is a need to balance the effective prevention of crime and civil liberties. As technology continues to advance, solving these problems will be central to making safer and more equitable societies[[16], [17], [18], [19]].

Advanced technological innovations like You Only Look Once Version 11 (YOLO11), one of the newest object detection algorithms, present unprecedented precision and speed in surveillance systems, addressing traditional limitations. Its real-time detection capabilities and robust multi-scale object recognition make it particularly beneficial for urban spaces. YOLO11 builds on YOLOv8 with improvements like the C3K2 block, SPFF module, and C2PSA block, enhancing feature extraction and spatial processing. These innovations boost detection of small or occluded objects while maintaining efficiency, making YOLOv11 ideal for high-priority environments. [20].

The figure1 illustrates the YOLO (You Only Look

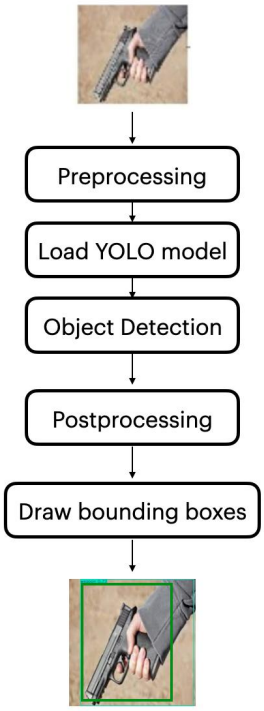

Fig. 1: System Design for YOLO11

Once) object detection system design. Initially, input images undergo preprocessing to normalize and prepare them for the model. The YOLO model is then loaded, performing object detection to identify objects within the image. Postprocessing is applied to refine detections, filter out low-confidence predictions, and remove overlaps. Finally, bounding boxes are drawn around detected objects, highlighting them in the image for interpretation.

The performance improvements of YOLO11 over YOLOv10, including higher mean average precision (mAP50) and significant gains in recall, will be discussed in a later section. YOLO11 achieves a higher mAP50 of 0.879 compared to YOLOv10's 0.857. While YOLOv10 has a faster inference time, YOLO11's enhanced accuracy justifies the slight trade-off in processing time. These advancements demonstrate YOLO11's effectiveness in handling complex object detection tasks.

This paper follows the structure is as follows: Section II describes the methodology behind this proposed YOLO11 model, which provides architecture and components of such a model.Section III gives the description of the dataset and model implemetation.In Section IV, we discuss results and performance metrics of the proposed YOLO11 in real-time surveillance. Finally, Section V concludes the paper, discussing the main results and listing future research avenues.

II. Background Study

The architecture of YOLO11 optimizes both speed and accuracy, building on advancements from YOLOv8 to YOLOv10 [13]. Key innovations include the C3K2 block, SPFF module, and C2PSA block, which improve spatial processing and maintain fast inference [14]. The design also incorporates advanced feature extraction and multi-scale feature fusion, enhancing the detection of small or occluded objects while ensuring computational efficiency [15].

A. Backbone

The backbone extracts features from input images using convolutional blocks, bottleneck structures, and advanced modules like C2F and C3K2.

1) Convolutional Block : The Convolutional Block [23]processes the input through a 2D convolution, batch normalization, and SiLU activation. The transformation is represented by Equation 1 and Equation 5.:

$$Y = \text{SiLU}(\text{BatchNorm}(\text{Conv2D}(X))) \tag{1}$$

In this context, let X represent the input tensor, and $Z = \text{Conv2D}(X)$ be the output of a 2D convolution operation. The batch normalization is applied to Z, resulting in equation 2

$$Z' = \gamma \frac{Z - \mu}{\sigma} + \beta, \tag{2}$$

where μ and σ are the mean and standard deviation of Z over the batch, respectively, and γ and β are learnable scale and shift parameters. Subsequently, the sigmoid activation function is defined in equation 3

$$\sigma(Z') = \frac{1}{1 + e^{-Z'}}. \tag{3}$$

Finally, the SiLU acivation function is applied, producing the output as shown in equation 4

$$Y = Z' \cdot \sigma(Z'). \tag{4}$$

The final expression is:

$$Y = Z' \cdot \frac{1}{1 + e^{-Z'}} \tag{5}$$

2) Bottleneck Block: This is a sequence of convolutional blocks with a shortcut parameter that decides whether to include the residual part or not. It is similar to the ResNet Block [22]. If the shortcut is set to False, no residual would be considered. If the shortcut is enabled, the output is:

$$Y = X + F(X) \tag{6}$$

Otherwise, if the shortcut is disabled, the output is:

$$Y = F(X) \tag{7}$$

The transformation is represented by Equation 6 when the shortcut is enabled, and Equation 7 when it is disabled.

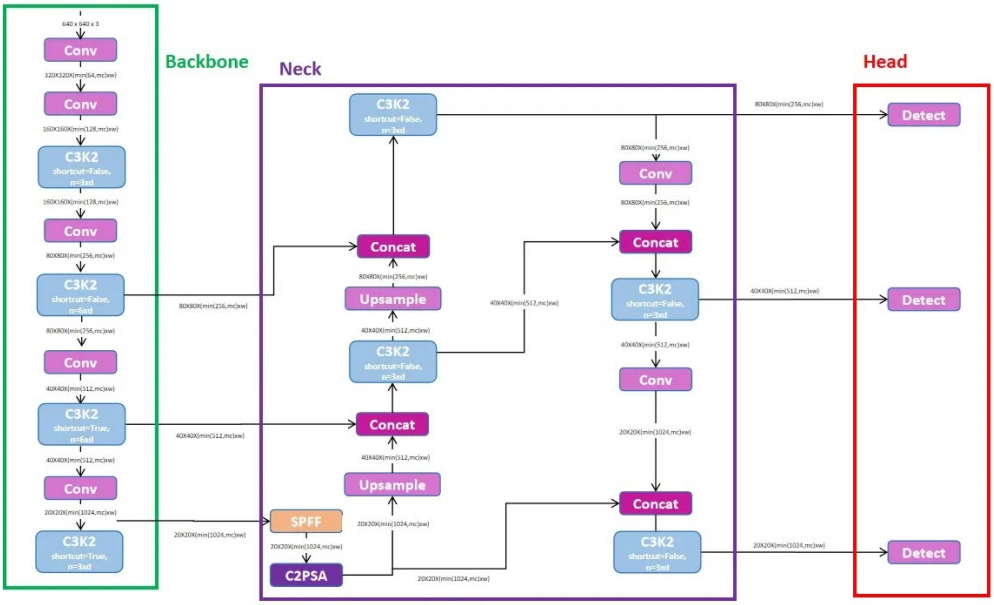

Fig. 2: YOLO11 Archictecture [21]

3) C2F Block: The C2F block (Cross Stage Partial Focus, CSP-Focus) enhances efficiency and feature map preservation. It consists of a Conv block, splits output channels into two halves, processes them with n Bottleneck layers, concatenates their outputs, and applies a final Conv block. This improves feature connections while minimizing redundancy. The transformation is represented as:

$$Y = \text{Concat}(F_1, \text{Processed}(F_2)) \tag{8}$$

In this context, let F_1 represent the first half of the input channels and F_2 represent the second half. The term Processed(F_2) refers to the result obtained by passing F_2 through a series of n Bottleneck layers. Finally, Concat denotes the concatenation of feature maps along the channel dimension.

Equation 8 represents the final output after concatenating the processed feature maps.

4) C3K2 Block: The C3K2 block in YOLOv11 enhances feature extraction with smaller 3×3 kernels for efficiency while maintaining detail. It improves upon the CSP bottleneck by splitting and processing feature maps with convolutions, then merging outputs, offering better representation with fewer parameters than YOLOv8's C2F. The structure includes initial and final convolutions, multiple C3K blocks, and efficient feature connections, optimizing detection with a balance of speed and accuracy. Its operation:

B. Neck

The Neck combines multi-scale features for better object detection.

1) Spatial Pyramid Pooling Fast (SPPF): The SPPF module uses max-pooling at multiple scales to capture spatial hierarchies and patterns across different receptive fields, improving robustness to object size and position variations. The max-pooling outputs are concatenated along the channel dimension, and a convolutional layer refines the fused map, reducing redundancy and enhancing discriminative features for multi-scale object detection.The transformation is represented as:

$$Y = \text{Concat}(\text{MaxPool}_1(X), \text{MaxPool}_2(X), \ldots, \text{MaxPool}_n(X)) \tag{9}$$

Where: X represents the input tensor, $\text{MaxPool}_i(X)$ denotes the output of the i-th MaxPooling operation applied to X, for $i \in \{1, 2, \ldots, n\}$. The term Concat refers to the concatenation of all MaxPooling outputs along the channel dimension, and Y is the final concatenated output tensor.

The process of concatenating max-pooling outputs allows for more robust feature extraction and better performance in object detection, as described in equation 9.

2) C2PSA Block: The C2PSA block uses two Partial Spatial Attention (PSA) modules on separate branches, concatenating their outputs to enhance focus on spatially significant regions. This improves fine-detail detection with high precision while maintaining efficiency, enabling YOLO11 to outperform YOLOv8.Equation 10 represents the spatial attention mechanism applied to the input tensor X.

$$\text{Attention}(X) = \text{Softmax}(\text{Conv2D}(X)) \tag{10}$$

C. Head

The Head generates predictions at three scales (small, medium, large) for multi-scale object detection represented by equation11:

$$Y_{\text{head}} = [P_3, P_4, P_5] \tag{11}$$

where P_3, P_4, and P_5 are feature maps of different granularity.

III. Proposed Work

This study utilizes the *Suspicious Detection Dataset* from Roboflow [24] to train and evaluate a YOLO11-based model designed for detecting suspicious activities in surveillance environments. The dataset (sample images shown in fig. 3) consists of 4,143 meticulously annotated images featuring a variety of unusual behaviors, each labeled with high-quality bounding boxes and class labels for objects and actions. To ensure a balanced evaluation, the dataset was divided into three subsets: 88% (3,627 images) for training, 8% (344 images) for validation, and 4% (172 images) for testing, maintaining consistency throughout the model's development pipeline.

Fig. 3: Sample images from dataset

Before training, the dataset underwent preprocessing steps tailored to maximize the effectiveness of the YOLO11 architecture. These included resizing images to fit YOLO11's input dimensions and normalizing pixel values to standardize input data. To further enhance the model's robustness and generalization, diverse data augmentation techniques such as horizontal flipping, rotation, scaling, and brightness adjustments were applied. These methods introduced real-world variations into the training data, effectively mitigating overfitting and improving the model's ability to handle unseen scenarios.

The YOLO11 architecture was chosen for its remarkable balance between real-time object detection capabilities and computational efficiency, making it ideal for surveillance applications. The training process involved detecting a range of suspicious activities, including loitering, unauthorized access, and object tampering. Hyperparameters such as learning rate, batch size, and the number of epochs were meticulously fine-tuned to achieve optimal performance, ensuring that the model could deliver accurate results while maintaining inference speed. For evaluation, the model's performance was assessed using standard metrics such as precision , recall, F1-score, and mean Average Precision (mAP).

These metrics provided a comprehensive analysis of the system's ability to accurately identify and classify suspicious activities across diverse scenarios. The results demonstrated that the YOLOv11-based system achieved high precision and recall values, highlighting its effectiveness in recognizing fine details and maintaining low false positive rates. The high mAP scores confirmed the model's reliability in detecting suspicious activities with consistency.

The comprehensive application of the *Suspicious Detection Dataset* played a pivotal role in training and evaluating the model in realistic scenarios, ensuring that the system is well-suited for practical deployment. By leveraging this dataset and the advanced YOLO11 architecture, the proposed system provides a robust solution for automated surveillance, offering enhanced safety and security across a range of environments.

IV. Results

The performance of the YOLO models on the dataset is summarized in Table I. The models were trained and validated using the Ultralytics framework on a system equipped with a Tesla T4 GPU and 15,102 MiB of memory. Evaluation metrics include precision (P), recall (R), and mean average precision at IoU threshold 0.5 (mAP@0.5).

TABLE I: Comparison of YOLOv10 and YOLO11 Performance.

Model	Class	Box(P)	R	mAP50
YOLOv10	All	0.861	0.768	0.857
	Normal-Action	0.844	0.804	0.892
	Suspicious-Suspect	0.907	0.858	0.919
	Victim	0.828	0.828	0.910
	Weapon	0.865	0.583	0.708
YOLO11	All	0.830	0.843	0.879
	Normal-Action	0.813	0.865	0.891
	Suspicious-Suspect	0.861	0.903	0.928
	Victim	0.822	0.897	0.928
	Weapon	0.825	0.709	0.770

The overall results in Table I demonstrate that YOLO11 surpasses YOLOv10 across most categories. YOLO11 achieves a higher overall mean average precision (mAP50) of 0.879 compared to YOLOv10's 0.857. Notably, the `Suspicious-Suspect` and `Victim` classes show the highest mAP50 of 0.928 in YOLO11, outperforming YOLOv10's 0.919 and 0.910, respectively. Additionally, recall (R) values improved significantly for all classes, with the `Weapon` class showing a substantial increase from 0.583 in YOLOv10 to 0.709 in YOLO11.

Fig. 4: Examples of model predictions: Weapon detection, Normal-action detection, Suspicious-suspect detection .

In terms of speed, YOLOv10 has a faster inference time of 3.1 ms per image compared to YOLO11's 11.6 ms. However, YOLO11's improved accuracy justifies the slight trade-off in processing time. These results highlight the advancements in YOLO11 for handling complex object detection tasks effectively.

V. Conclusion & Future Work

In conclusion, the YOLO11 object detection model represents a significant advancement in surveillance technology, offering a balanced trade-off between speed and accuracy through innovative modules like the C3K2 block and SPFF. It effectively detects diverse classes, including suspicious activities and weapons, with high precision and recall, making it a cost-effective and adaptable solution for scalable surveillance systems. Its real-time processing capabilities reduce the need for manual monitoring, enabling seamless integration into various security applications and enhancing community safety.

Future work will enhance YOLO11's detection of occluded and small objects using feature fusion, optimize it for edge devices. Applications include intrusion detection, emergency alerts, and monitoring restricted areas. Additionally, integrating YOLO-11 with IoT and smart systems will enable automated decision-making and expand its use in security and surveillance.

References

[1] National Crime Records Bureau, *Crime in India Report*, 2022.

[2] J. Smith, "Technological Advancements in Urban Policing," *Journal of Urban Security*, vol. 8, no. 3, pp. 45–56, 2021.

[3] P. Kumar and R. Shah, "Impact of Surveillance on Crime Prevention," *International Journal of Criminal Studies*, vol. 15, no. 2, pp. 98–105, 2020.

[4] National Crime Records Bureau, "Urban Crime Statistics," 2022.

[5] L. Jones, "AI in Policing: A New Frontier," *Technology and Society*, vol. 12, no. 1, pp. 12–20, 2023.

[6] S. Gupta, "Challenges of CCTV Implementation in Developing Nations," *Security Review*, vol. 6, no. 4, pp. 22–30, 2021.

[7] H. Lee, "Video Analytics and Real-Time Crime Detection," *IEEE Transactions on Information Forensics*, vol. 18, no. 7, pp. 120–130, 2022.

[8] A. Sharma, "Community-Driven Surveillance and its Implications," *Journal of Social Technology*, vol. 9, no. 2, pp. 34–40, 2021.

[9] N. Patel, "Cost Analysis of Urban Surveillance Systems," *Economic Studies in Technology*, vol. 10, no. 3, pp. 50–60, 2022.

[10] "Ethics of Surveillance Technology," *International Journal of Security Studies*, vol. 11, no. 1, pp. 15–25, 2023.

[11] Z. Wang, "Future of Predictive Policing," *Computing Advances*, vol. 14, no. 4, pp. 78–85, 2023.

[12] Citywide Safety Initiative, "Case Study: Delhi's Surveillance Network," 2022.

[13] M. Hussain, "YOLOv5, YOLOv8 and YOLOv10: The Go-To Detectors for Real-time Vision," arXiv, Jul. 2024. [Online]. Available: https://arxiv.org/abs/2407.02988.

[14] R. Khanam and M. Hussain, "YOLOv11: An Overview of the Key Architectural Enhancements," ResearchGate, Oct. 2024. [Online]. Available: https://www.researchgate.net/publication/385177106_YOLOv11_An_Overview_of_the_Key_Architectural_Enhancements.

[15] Unknown Author, "YOLOv11 for Vehicle Detection: Advancements, Performance, and Applications," arXiv, Oct. 2024. [Online]. Available: https://arxiv.org/pdf/2410.22898.

[16] Plural Policy, "Government Surveillance and Civil Liberties," [Online]. Available: https://pluralpolicy.com/blog/government-surveillance-civil-liberties/.

[17] Brookings Institution, "Police Surveillance and Facial Recognition: Why Data Privacy is an Imperative for Communities of Color," [Online]. Available: https://www.brookings.edu/articles/police-surveillance-and-facial-recognition-why-data-privacy-is-an-imperative-for-communities-of-color/.

[18] University of Southern California, "Ethics of Data Sharing and Digital Privacy," [Online]. Available: https://vce.usc.edu/volume-7-issue-2/ethics-of-data-sharing-and-digital-privacy/.

[19] Data Ethics Organization, "The Ethics of Surveillance and Privacy," [Online]. Available: https://www.dataethics.org/surveillance-and-privacy/.

[20] DigitalOcean, "What's New in YOLOv11: Transforming Object Detection Once Again," [Online]. Available: https://www.digitalocean.com/community/tutorials/what-is-new-with-yolo.

[21] Analytics Vidhya, "YOLOv11: Object Detection - Advancements and Applications," [Online]. Available: https://www.analyticsvidhya.com/blog/2024/10/yolov11-object-detection/.

[22] Analytics Vidhya, "Understanding ResNet and Analyzing Various Models on the CIFAR-10 Dataset," [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/understanding-resnet-and-analyzing-various-models-on-the-cifar-10-dataset/.

[23] Analytics Vidhya, "A Comprehensive Guide to Convolutional Neural Networks (CNN)," [Online]. Available: https://www.analyticsvidhya.com/blog/2018/12/guide-convolutional-neural-network-cnn/.

[24] Roboflow, "Suspicious Detection Dataset Documentation," [Online]. Available: https://universe.roboflow.com/suspicious-movement/suspicious-detection/dataset/7.