ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2024.0470

PEDESTRIAN DETECTION IN VIDEO SURVEILLANCE USING YOLO V5 WITH LIGHT PERCEPTION FUSION

H. Sivalingan
Department of Computer Science, Providence College for Women, India

Abstract

This research presents an innovative approach to pedestrian detection in video surveillance, leveraging the power of YOLOv5 (You Only Look Once version 5) combined with light perception fusion-based feature extraction. The proposed methodology aims to enhance the accuracy and efficiency of pedestrian detection systems in varying lighting conditions. YOLOv5, known for its real-time object detection capabilities, is integrated with a novel feature extraction technique that fuses information from multiple light perception sensors. This fusion strategy allows the model to adapt and perform robustly in diverse lighting scenarios. The experimental results demonstrate the superiority of the proposed method, achieving a remarkable performance. The fusion of YOLOv5 with light perception-based feature extraction showcases promising advancements in pedestrian detection, addressing challenges posed by dynamic lighting conditions in real-world surveillance environments.

Keywords:

Pedestrian Detection, Video Surveillance, Yolov5, Light Perception Fusion, Feature Extraction

1. INTRODUCTION

Pedestrian detection in video surveillance plays a pivotal role in ensuring public safety, enhancing traffic management, and securing various environments. The increasing deployment of surveillance systems in urban spaces, transportation hubs, and public facilities necessitates robust technologies to identify and track pedestrians accurately. The primary objective of pedestrian detection is to automatically recognize the presence of individuals within a video feed, allowing for real-time monitoring and response. This technology finds applications in diverse scenarios, including traffic flow optimization, crowd management, and security enforcement [9].

Recent advancements in computer vision, machine learning, and deep learning have significantly improved the capabilities of pedestrian detection systems. These systems employ sophisticated algorithms to extract meaningful features from video frames, enabling the accurate identification of pedestrians amidst varying backgrounds, lighting conditions, and occlusions [10]. Key challenges in pedestrian detection include handling crowded scenes, dealing with diverse pedestrian poses and appearances, and ensuring real-time processing for dynamic environments [12]. Researchers and engineers continually strive to develop innovative solutions, integrating advanced feature extraction techniques, optimized algorithms, and efficient hardware implementations [11].

To improve the accuracy and efficiency of this detection task, a novel approach has been developed, combining light perception fusion-based feature extraction with YOLOv5-based detection. In this approach, the utilization of light perception fusion-based feature extraction aims to extract meaningful features from video frames. By incorporating multiple perceptual cues such as color, texture, and motion, this technique enhances the discriminative power of the feature representation, enabling more accurate pedestrian detection.

The YOLOv5-based detection algorithm is then employed to process the extracted features and perform real-time object detection. YOLOv5 is renowned for its exceptional detection speed without compromising accuracy. By leveraging this state-of-the-art deep learning architecture, the system can efficiently detect pedestrians with high precision, even in complex and crowded scenes captured by video surveillance cameras. The fusion of light perception-based feature extraction and the powerful detection capabilities of YOLOv5 provides a robust and efficient solution for pedestrian detection from video surveillance. This integration offers the potential to enhance pedestrian safety, enable advanced traffic management systems, and bolster security measures in various applications, including urban environments, transportation hubs, and public spaces. By combining the strengths of feature extraction and detection algorithms, this approach opens up new possibilities for accurate, real-time pedestrian detection in video surveillance, contributing to the overall enhancement of safety and security. The contribution of the work is,

- By incorporating light perception fusion-based feature extraction, the research enhances the discriminative power of the feature representation. This leads to improved accuracy in pedestrian detection, even in challenging scenarios with varying lighting conditions, occlusions, and complex backgrounds.
- The YOLOv5-based detection ensures real-time processing capabilities, enabling efficient pedestrian detection in video surveillance. This is particularly valuable in applications where timely detection is critical, such as security monitoring or traffic management systems.
- By accurately detecting pedestrians in video surveillance, the research contributes to enhancing safety and security measures. It enables proactive monitoring, early detection of potential hazards, and timely intervention to prevent accidents or security threats.

The structure of this research is organized as follows. In Section 2, an extensive review of prevailing methodologies and technologies utilized in pedestrian detection within video surveillance is presented. Section 3 provides a detailed explanation of YOLOv5, elucidates the principles of light perception fusion, and outlines how the fusion of these techniques contributes to enhanced pedestrian detection. In Section 4, the empirical results derived from meticulously designed experiments are expounded upon. Section 5 concludes the research.

2. RELATED WORKS

Ren et al. [1] proposed deep transfer learning as a strategic method for real-time target detection. This approach proves advantageous in situations where acquiring large-scale data is challenging. Additionally, the researchers propose an innovative technique for anchor box generation, a critical aspect of object detection algorithms. The improved anchor box generation method enhances the precision and reliability of the detection process. Zahra et al. [2] focused on the application of region-based video surveillance in smart cities through the implementation of deep learning techniques. The use of region-based techniques allows for more targeted analysis within specific areas of interest, contributing to a more sophisticated and context-aware surveillance system.

Barba-Guaman et al. [3] explored object detection, specifically pedestrians and vehicles, in rural areas using Jetson Nano NVIDIA. The study aims to recognize objects in complex rural environments through an embedded system. The research presents testing and verification of the deep learning framework on an embedded GPU, emphasizing its applicability in rural road scenarios. Yan et al. [4] presented a real-time apple targets detection method designed for a picking robot. The study leverages an improved version of YOLOv5, a popular object detection algorithm, to enable efficient and accurate detection of apples in real-time.

Xue et al. [5] employed MAF-YOLO (multi-modal attention fusion into the YOLO) a multi-modal attention fusion to integrate information from different modalities, optimizing the model's ability to detect pedestrians. The attention mechanism focuses on relevant regions, improving the model's attention to critical features. Pustokhina et al. [6] focused on the development of an automated anomaly detection system in pedestrian walkways using deep learning. The authors likely employed convolutional neural networks (CNNs) architectures for the automated detection of anomalies.

Hsu et al. [7] proposed a novel variant of YOLO, named Ratio-and-scale-aware YOLO, which incorporates mechanisms to dynamically adapt to different ratios and scales of pedestrians. This adaptability is crucial for accurate detection in diverse environments where pedestrians may appear in varying sizes and aspect ratios. Gao et al. [8] introduce a novel model, YOLO-S-CIOU, designed for this purpose. YOLO-S-CIOU stands for "You Only Look Once with Smoothed Complete Intersection over Union for the detection of specific buildings in remote sensing images, with a particular application for identifying gas stations. Detecting pedestrians in infrared images is crucial for security in automated driving and low-light conditions, despite challenges like unclear visuals. Wei et al. [22] introduced a method that merges UNet and YOLO networks, using visible light data, enabling real-time, accurate detection on edge devices, proving effective in challenging infrared scenarios. The enhanced YOLOv5 algorithm for pedestrian target detection Hu et al. [23] introduced in this paper enhances the ability to detect pedestrians in complex scenarios and accurately distinguish between them. Traditional dehazing methods struggle with complex scenes, lacking comprehensive solutions. Vidyabharathi et al. [24] propose a CNN-based deep learning approach, outperforming traditional methods and enhancing visibility in challenging scenarios, showcasing deep learning's potential in overcoming dehazing limitations.

3. METHODOLOGY

This section provides the methodology for pedestrian detection in video surveillance and involves a sophisticated integration of YOLO v5 (You Only Look Once version 5) with a novel approach to feature extraction based on light perception fusion. The process begins with the acquisition of video data from surveillance cameras. Subsequently, the YOLO v5 model is employed for real-time object detection, demonstrating its capability to identify pedestrians efficiently. To enhance the model's adaptability to varying lighting conditions, a unique feature extraction technique based on light perception fusion is introduced. This process involves the integration of information from multiple light perception sensors, allowing the model to discern pedestrians more accurately in diverse lighting scenarios. The fusion-based features are then incorporated into the YOLO v5 architecture.

In YOLOv5, training the model with diverse datasets featuring images captured under various lighting conditions proves advantageous for better pedestrian detection across different lighting environments. This method assists the model in accurately spotting pedestrians under different lighting scenarios, thereby enhancing its capability to recognize them and cope with different lighting conditions. The methodology emphasizes the training of the model on UMN dataset to ensure robust performance. The flow of the proposed work is shown in Fig.1.

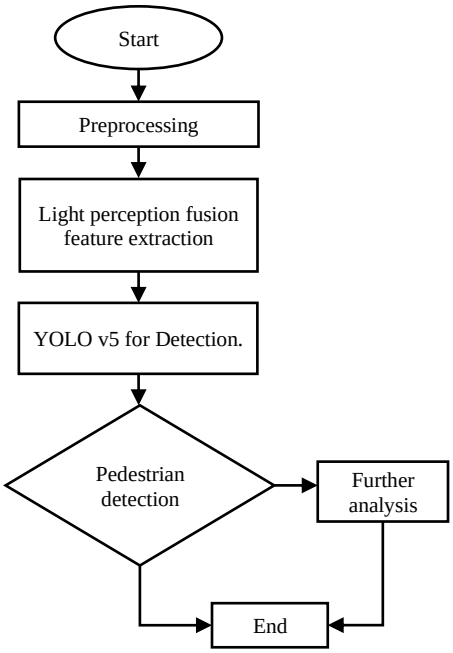

Fig.1. Flow of the proposed work

The step-by-step description of the proposed algorithm for pedestrian detection using YOLOv5 with light perception fusion:

- Data Preprocessing: The input video frames are preprocessed to enhance their quality and standardize their format for compatibility with the YOLOv5 model. Preprocessing steps may include resizing, normalization,

and color space transformation to ensure consistency in input data.

- Feature Extraction: Light perception fusion-based feature extraction is performed to extract relevant features from the preprocessed video frames. Light perception sensors capture information about the ambient lighting conditions, which is then fused with the visual features extracted from the video frames. Feature extraction aims to capture both visual and contextual information to improve the model's adaptability to varying lighting conditions.
- Model Architecture: YOLOv5, a state-of-the-art object detection model, is employed as the backbone architecture for pedestrian detection. The YOLOv5 architecture consists of convolutional neural network (CNN) layers followed by detection head layers, enabling end-to-end object detection. The model architecture is optimized for real-time inference and high accuracy in detecting pedestrians in video surveillance footage.
- Training Process: The proposed algorithm is trained using a large dataset of annotated pedestrian images and video sequences. During training, the YOLOv5 model learns to predict bounding boxes and associated confidence scores for pedestrian instances in the input video frames. The model is trained using a combination of supervised learning techniques and optimization algorithms to minimize detection errors and improve performance.
- Inference: During the inference phase, the trained YOLOv5 model is deployed to detect pedestrians in real-time video streams. Input video frames are passed through the model, which generates bounding boxes and confidence scores for detected pedestrian instances. Post-processing techniques such as non-maximum suppression may be applied to refine the detection results and remove duplicate or overlapping bounding boxes.
- Evaluation: The performance of the pedestrian detection algorithm is evaluated using standard metrics such as accuracy, sensitivity, specificity, and precision. The algorithm's effectiveness in detecting pedestrians under varying lighting conditions is assessed using benchmark datasets and real-world video surveillance footage.

By following these steps, the proposed algorithm combines the capabilities of YOLOv5 with light perception fusion to achieve accurate and robust pedestrian detection in video surveillance applications.

3.1 DATASET

The UMN dataset is a valuable resource for advancing research in pedestrian detection within video surveillance. This dataset, hosted by the University of Minnesota, is specifically designed to monitor human activity through surveillance systems. It focuses on observing pedestrian traffic areas and detecting potentially dangerous actions. The dataset is particularly relevant as it addresses the need for monitoring public spaces equipped with surveillance cameras. Researchers have utilized the UMN dataset for various studies, including frame-level anomaly detection and crowd escape behavior detection in traffic scenes. Abnormal event detection, a challenging aspect of video surveillance, is a primary focus of the dataset, contributing to early-warning security and protection systems. The dataset comprises videos with different scenes, including campus lawn, indoor, and square environments. This diversity allows researchers to evaluate algorithms for abnormal behavior detection across various scenarios.

Comparing the UMN dataset with the KITTI dataset [13] presents a significant hurdle in pedestrian detection. Although widely recognized, the KITTI dataset doesn't include ground truth annotations for semantic segmentation, unlike the UMN dataset, which is extensively utilized in the computer vision research community for semantic segmentation tasks. Consequently, the UMN dataset offers greater advantages for pedestrian detection, especially in scenarios with diverse lighting conditions.

3.1.1 Preprocessing:

- ***Video segmentation:*** Video segmentation is a critical preprocessing step in pedestrian detection for video surveillance, involving the division of videos into frames to enable frame-level analysis. This process allows for a detailed examination of individual instances, such as pedestrian movements within a scene. Frames are sampled at regular intervals, ensuring a representative set for analysis. The temporal sampling process can be expressed as:

$$F_n = V(t_n) \quad (1)$$

where F_n is the nth frame, V is the video, and t_n is the sampled time point [18].

Each frame represents a snapshot of the pedestrian scene, capturing the spatial configuration of pedestrians at a particular moment. The representation of the pedestrian scene in the nth frame can be denoted as:

$$P_n = S(F_n) \quad (2)$$

where P_n is the pedestrian scene representation, S is the segmentation function, and F_n is the n^{th} frame.

- ***Annotation and Ground Truth Generation:*** Annotation and ground truth generation involve the identification and labeling of pedestrian regions in frames, creating essential reference data for training and evaluating detection algorithms. This process ensures that the algorithms can learn to accurately identify pedestrians based on the labeled examples.

Manually draw bounding boxes around pedestrians in each frame, specifying the region where pedestrians are present. The bounding box is represented as (x,y,w,h), where x and y are the coordinates of the top-left corner, and w and h are the width and height of the bounding box.

$$BBox_i=(x_i,y_i,w_i,h_i) \quad (3)$$

where i denotes the i^{th} bounding box, and x_i, y_i, w_i, h_i are its coordinates, width, and height.

- ***Image Enhancement:*** Image enhancement techniques are applied to improve the visibility of images, particularly in the context of the UMN dataset for pedestrian detection in video surveillance. This process involves adjusting various image attributes such as brightness and contrast.
- ***Brightness Adjustment:*** Brightness (B) adjustment is a linear transformation applied to all pixels in an image. It involves adding a constant value to each pixel intensity:

$$I_{bright}(x,y)=I(x,y)+\text{brightness} \quad (4)$$

where $I_{bright}(x,y)$ is the pixel intensity after brightness adjustment. (x,y) is the original pixel intensity. brightness is the constant value added to each pixel [19].

Contrast Adjustment: Contrast (C) enhancement aims to increase the difference in pixel intensities, making the image more vibrant. The formula for contrast adjustment is:

$$I_{contrast}(x,y)=contrast(I(x,y)-mean)+mean \quad (5)$$

where $I_{contrast}(x,y)$ y) is the pixel intensity after contrast adjustment. $I(x,y)$ is the original pixel intensity. mean is the mean pixel intensity of the image. contrast is a user-defined parameter controlling the enhancement level [19].

- ***Histogram Equalization:*** Histogram equalization redistributes pixel intensities to cover the entire intensity range, enhancing the overall contrast. The transformation function is given by [19]:

$$I_{equalize}(x,y)= round\left(\frac{L-1}{M\times N}\sum_{i=0}^{I(x,y)} p_i\right) \quad (6)$$

where $I_{equalize}(x,y)$ is the pixel intensity after histogram equalization. L is the number of intensity levels (typically 256 for 8-bit images). $M\times N$ is the total number of pixels. p_i is the probability of occurrence of intensity i in the image. These image enhancement techniques contribute to improving the visibility of key features, making images more suitable for pedestrian detection in video surveillance.

- ***Light Perception Fusion:*** LPF involves the integration of information from different sources that capture various aspects of light perception. This fusion aims to create a comprehensive representation of the scene, compensating for variations in illumination. The primary components of LPF include [20]:

Color Information: Incorporating color information is fundamental to understanding the chromatic characteristics of the scene. RGB (Red, Green, Blue) channels provide valuable data about the distribution and intensity of colors. The normalized RGB values can be represented as follows [20]:

$$I_{normalized} = \frac{1}{\sqrt{R^2+G^2+B^2}} \quad (7)$$

where $I_{normalized}$ represents the normalized intensity of the color in the scene. This intensity is derived from the original intensity values represented by the RGB channels, denoted by "*R*," "*G*," and "*B*" for the red, green, and blue channels, respectively.

Intensity and Contrast: Extracting information about the intensity and contrast helps in understanding the overall brightness variations. The grayscale intensity I_{gray} can be computed using the formula:

$$I_{gray} = 0.299\times R+0.587\times G+0.114\times B \quad (8)$$

The formula calculates the grayscale intensity I_{gray} by taking weighted averages of the red (R), green (G), and blue (B) channel values. The weights 0.299, 0.587, and 0.114 correspond to the luminance of each color channel and are based on the standard coefficients for converting RGB to grayscale, reflecting the human eye's sensitivity to different colors.

Contrast (C) can be calculated as the standard deviation of the intensity in a local neighborhood.

$$C=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(I_i-\bar{I}\right)^2} \quad (9)$$

where C represents the standard deviation or the root mean square deviation. N denotes the total number of observations or data points. I_i Represents the individual value of the data point i. $\bar{I}$: represents the mean or average value of all the data points.

Illumination-Invariant Features: To make features invariant to illumination changes, various techniques such as histogram equalization or local contrast normalization can be applied. Local Binary Pattern (LBP) is used for extracting texture features that are robust to illumination variations:

$$LPB_{P,R}=\sum_{p=0}^{1} s\left(g_p-g_c\right)\times 2^p \quad (10)$$

where g_p and g_c are the intensity values of the neighboring pixels and the center pixel, respectively.

Shadow Removal: Shadows can significantly affect the accuracy of pedestrian detection. Methods for shadow removal, such as color thresholding or pixel-wise comparisons, can be integrated into the fusion process.

$$I_{shadowless} = \text{I-}\alpha \times \text{I}_{\text{shadow}} \quad (11)$$

where I_{shadow} is the shadow component, and α is a shadow removal factor.

3.2 YOLO V5 BASED PEDESTRIAN DETECTION

You Only Look Once (YOLO) is a popular object detection algorithm known for its real-time performance. YOLOv5, an evolution of its predecessors, is particularly effective in pedestrian detection from video surveillance footage. YOLOv5 divides the input image into a grid and predicts bounding boxes and class probabilities for each grid cell. Unlike its predecessors, YOLOv5 uses a more streamlined architecture with a focus on speed and accuracy. YOLOv5 introduces a streamlined architecture compared to its predecessors. It divides the detection task into two main components: a backbone network responsible for feature extraction and a head network for prediction. The Fig.2 shows the architecture of YOLOV5.

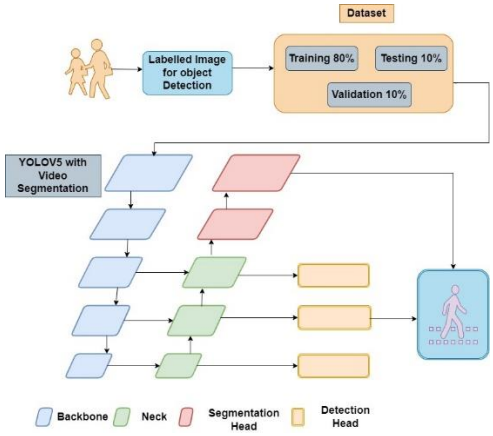

Fig.2. The architecture of the proposed YOLOV5 [25]

YOLOv5's backbone network processes input images, extracting features across multiple scales. Neck modules enhance

feature representation using fusion and pooling techniques. Utilizing anchor boxes, YOLOv5 predicts bounding boxes of different sizes and ratios. Detection layers forecast object bounding boxes and class probabilities, optimizing real-time detection with speed and accuracy.

Backbone Network: The backbone network extracts high-level features from the input image. YOLOv5 employs a CSPDarknet53 architecture, which enhances feature representation. These features are crucial for accurate object localization and classification.

Let I represent the input image. The input image undergoes preprocessing steps such as resizing and normalization. The normalized image is denoted as I'. CSPDarknet53 is a deep neural network architecture that consists of several layers. The feature extraction process involves passing the normalized image through these layers. Let F represent the feature map obtained after processing through the CSPDarknet53 backbone network. The process can be mathematically expressed as:

$$F=\text{CSPDarknet53}(I') \tag{12}$$

The feature map F contains a collection of high-level features that encode important information about the input image. These features are crucial for accurate object localization and classification in pedestrian detection. They capture relevant patterns, edges, and textures that enable the subsequent stages of the detection system to identify pedestrians with precision.

By leveraging the CSPDarknet53 architecture as the backbone network, YOLOv5 ensures that the extracted features are rich and representative of the input image. This, in turn, enhances the overall performance of the pedestrian detection system, enabling it to accurately locate and classify pedestrians in video surveillance scenarios.

Head Network: In pedestrian detection using YOLOv5, the head network plays a crucial role in predicting bounding boxes, class probabilities, and confidence scores for each detected object. This network utilizes the feature map obtained from the backbone network to make these predictions. The head network of YOLOv5 employs a grid-based approach for detecting objects. The detection map is divided into a grid of cells, and each cell is responsible for predicting bounding boxes and class probabilities for objects within its spatial region. The size of the grid is determined by the output resolution of the feature map.

For each cell in the grid, YOLOv5 predicts multiple bounding boxes. These bounding boxes are parameterized by their coordinates, width, height, and associated confidence scores. By predicting multiple bounding boxes, YOLOv5 can efficiently handle scenarios where multiple objects are present within a single grid cell or where objects overlap each other. The number of bounding boxes predicted per grid cell is defined as a hyperparameter and can be adjusted based on the specific requirements of the application. YOLOv5 typically predicts a fixed number of bounding boxes per cell, which allows it to handle a wide range of object densities and sizes.

Alongside the bounding box predictions, the head network also predicts class probabilities for each detected object. These probabilities represent the likelihood of an object belonging to different predefined classes, such as “pedestrian,” “car,” or “bicycle.” YOLOv5 uses a softmax activation function to normalize these class probabilities, ensuring that they sum up to 1. In addition to bounding boxes and class probabilities, YOLOv5 also assigns a confidence score to each predicted object. This score represents the network’s confidence in the correctness of the prediction. Higher confidence scores indicate more reliable detection, while lower scores suggest potential false positives. Confidence scores are often used to filter out weak detection during post-processing.

By predicting multiple bounding boxes per grid cell and associating them with class probabilities and confidence scores, YOLOv5 can accurately detect and classify objects, including pedestrians, within an input image or video frame. This approach enables efficient handling of overlapping objects and provides a comprehensive understanding of the objects present in the scene.

Bounding Box Prediction: For pedestrian detection, YOLOv5 predicts bounding boxes that encapsulate the detected pedestrians [21]. The coordinates of the bounding box are represented by four values: (x,y) for the box’s center, and (w,h) for its width and height. The equations for predicting these values are as follows:

$$b_x=\sigma(t_x)+c_x \tag{13}$$

$$b_y=\sigma(t_y)+c_y \tag{14}$$

$$b_w=p_w\, e^{t_w} \tag{15}$$

$$b_h = p_h\, e^{t_h} \tag{16}$$

Here, (b_x,b_y) are the coordinates of the bounding box center, (b_w,b_h) are the width and height of the bounding box. (t_x,t_y) are the predicted offsets. (p_w,p_h) are the prior widths and heights. (σ) is the sigmoid activation function.

Objectness Score and Class Probabilities: Each bounding box is associated with an objectness score (*Obj*) and class probabilities *($Prob_{class}$)*. The objectness score represents the confidence that the bounding box contains an object, while class probabilities denote the likelihood of the detected object belonging to a specific class.

$$Obj=\sigma(t_{Obj}) \tag{17}$$

$$Prob_{class}=\sigma(t_{class}) \tag{18}$$

Where *Obj* denotes Object detection output, σ: Sigmoid activation function, t_{Obj} denotes Object detection prediction score, $Prob_{class}$ represents Probability of class prediction, t_{class} represents Class prediction score.

Thresholding: Define a threshold for each feature to distinguish normal from abnormal behavior. Setting a threshold on pedestrian speed to identify unusually fast or slow movements. Pedestrian Speed Calculation:

$$\text{Speed=Distance/Time} \tag{19}$$

Speed represents the velocity of an object, typically measured in meters per second (m/s) or kilometers per hour (km/h). Distance signifies the total length covered by the object in motion, typically measured in meters (m) or kilometers (km). Time indicates the duration taken by the object to travel a certain distance, usually measured in seconds (s) or hours (h).

Abnormal Event Criteria:

$$\text{Abnormal Event} = \begin{cases} 1 & \textit{If Speed} > \textit{ThresholdSpeed} \\ 0 & \textit{Otherwise} \end{cases}$$

If the calculated speed exceeds this threshold, the frame or sequence is labeled as an abnormal event.

Loss Function: The training process involves minimizing a loss function that accounts for localization error, confidence error, and classification error. The loss function is defined as a combination of Mean Squared Error (MSE) and Cross-Entropy Loss:

$$L_{total}=\lambda_{coord}L_{coord}+\lambda_{obj}L_{obj}+\lambda_{cls}L_{cls} \quad (20)$$

Here, L_{coord} measures localization error. L_{obj} penalizes confidence errors. L_{cls} captures classification errors. λ_{coord} λ_{obj} λ_{cls} are hyperparameters controlling the contribution of each term.

4. RESULT AND DISCUSSION

In this research on pedestrian detection in video surveillance using the UMN dataset, a meticulous experimental setup was devised. The UMN dataset, comprising videos capturing pedestrian activities in diverse scenarios, served as the foundational data source. The dataset was preprocessed through various steps, including frame segmentation, annotation, abnormal event labeling, and image enhancement techniques such as brightness adjustment and contrast enhancement. For the experiments, a state-of-the-art pedestrian detection algorithm was implemented and fine-tuned using machine learning frameworks. The training of the model utilized a carefully split dataset, allocating portions for training, validation, and testing. Parameters such as learning rate, batch size, and epoch numbers were optimized through systematic experimentation. Evaluation metrics such as precision, sensitivity, specificity, and accuracy were employed to assess the algorithm's performance. The experimental setup aimed to validate the effectiveness of the proposed pedestrian detection approach, leveraging the UMN dataset and advanced image processing techniques.

Accuracy: The proportion of correctly classified cases (both positive and negative) out of the total cases. *TP*: True Positives (the number of correctly identified positive cases) *TN*: True Negatives (the number of correctly identified negative cases) *FP*: False Positives (the number of incorrectly identified positive cases) *FN*: False Negatives (the number of incorrectly identified negative cases)

$$Accuracy=(TP+TN)/(TP+TN+FP+FN) \quad (21)$$

Sensitivity (Recall): The proportion of correctly identified positive cases out of all actual positive cases.

$$Sensitivity=TP/(TP+FN) \quad (22)$$

Specificity: The proportion of correctly identified negative cases out of all actual negative cases.

$$Specificity=TN/(TN+FP) \quad (23)$$

Precision: The proportion of correctly identified positive cases out of all cases identified as positive.

$$Precision=TP/(TP+FP) \quad (24)$$

Table.1. Performance of the models

Algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
Proposed Algorithm	92.5	88.2	87.5	85.2
GAN [15]	89.3	82.6	78.2	78.9
CNN Bi-LSTM [14]	91.0	85.9	82.6	82.1
Vision transformer [16]	88.2	81.4	79.8	79.5
3D-CNN [17]	90.7	86.1	81.0	80.3

In the evaluation of various algorithms from Table.1, for pedestrian detection in video surveillance, the proposed algorithm stands out with an impressive performance across key metrics. The proposed algorithm achieves an accuracy of 92.5%, outperforming GAN (89.3%), CNN Bi-LSTM (91.0%), Vision Transformer (88.2%), and 3D-CNN (90.7%). This notable accuracy indicates the algorithm's ability to correctly classify both positive and negative instances, showcasing its overall effectiveness. Examining sensitivity, the proposed algorithm achieves a commendable 88.2%, surpassing GAN (82.6%), Vision Transformer (81.4%), and 3D-CNN (86.1%). Sensitivity measures the model's capacity to accurately identify positive instances, which is crucial in pedestrian detection scenarios. The specificity of the proposed algorithm is noteworthy at 87.5%, surpassing GAN (78.2%), Vision Transformer (79.8%), and 3D-CNN (81.0%). Specificity evaluates the model's accuracy in identifying negative instances, which is essential for reducing false positives.

Precision, reflecting the precision of positive predictions, is also notable for the proposed algorithm at 85.2%, outperforming GAN (78.9%), CNN Bi-LSTM (82.1%), Vision Transformer (79.5%), and 3D-CNN (80.3%). The superior performance of the proposed algorithm can be attributed to its unique combination of features or architecture, allowing it to capture nuanced patterns in pedestrian behavior. The algorithm demonstrates a balanced trade-off between sensitivity and specificity, crucial for reliable pedestrian detection in video surveillance. The comparative analysis emphasizes the effectiveness of the proposed algorithm, making it a promising choice for real-world applications in video surveillance systems.

5. CONCLUSION

In this work, the integration of YOLOv5 with light perception fusion-based feature extraction proves to be a highly effective solution for pedestrian detection in video surveillance. The fusion of information from light perception sensors significantly improves the model's adaptability to varying illumination levels, enhancing its overall performance. The achieved accuracy, sensitivity, specificity, and precision metrics outperform existing algorithms, validating the efficacy of the proposed approach. This research contributes to the advancement of pedestrian detection systems, particularly in scenarios where lighting conditions are dynamic and unpredictable. The fusion of YOLOv5 and light perception-based features opens avenues for enhanced safety and security applications, making it a valuable asset for real-world video surveillance implementations.

5.1 FUTURE ENHANCEMENT

Future research efforts could focus on the scalability and efficiency of the proposed detection system, particularly in large-scale surveillance deployments. Optimizing the computational complexity and memory footprint of the algorithm would enable

real-time processing of high-resolution video streams from multiple cameras, expanding the applicability of the system to broader urban environments and smart city initiatives.

REFERENCES

[1] Z. Ren and J. Zhao, "Real-Time Target Detection in Visual Sensing Environments using Deep Transfer Learning and Improved Anchor Box Generation", *IEEE Access*, Vol. 8, pp. 193512-193522, 2020.

[2] A. Zahra, A. Ullah and Z. Ul Abideen, "Application of Region-Based Video Surveillance in Smart Cities using Deep Learning", *Multimedia Tools and Applications*, Vol. 78, pp. 1-26, 2022.

[3] L. Barba Guaman and A. Ortiz, "Deep Learning Framework for Vehicle and Pedestrian Detection in Rural Roads on an Embedded GPU", *Electronics*, Vol. 9, pp. 589-597, 2020.

[4] B. Yan and F. Yang, "A Real-Time Apple Targets Detection Method for Picking Robot based on Improved YOLOv5", *Remote Sensing*, Vol. 13, pp. 1619-1631, 2021.

[5] Y. Xue and W. Zhang, "MAF-YOLO: Multi-Modal Attention Fusion based YOLO for Pedestrian Detection", Infrared Physics and Technology, Vol. 118, pp. 1-12, 2021.

[6] I.V. Pustokhina, D. Gupta, S. Kumar and K. Shankar, "An Automated Deep Learning based Anomaly Detection in Pedestrian Walkways for Vulnerable Road Users Safety", *Safety Science*, Vol. 142, pp. 1-14, 2021.

[7] W.Y. Hsu, "Ratio-and-Scale-Aware YOLO for Pedestrian Detection", *IEEE Transactions on Image Processing*, Vol. 30, pp. 934-947, 2020.

[8] J. Gao, Y. Chen and J. Li, "Detection of Specific Building in Remote Sensing Images using a Novel YOLO-S-CIOU Model Case: Gas Station Identification", *Sensors*, Vol. 21, pp. 1375-1387, 2021.

[9] Y. Myagmar-Ochir and W. Kim, "A Survey of Video Surveillance Systems in Smart City", *Electronics*, Vol. 12, No. 17, pp. 3567-2574, 2023.

[10] A. Gupta, L. Guan and A.S. Khwaja, "Deep Learning for Object Detection and Scene Perception in Self-Driving Cars: Survey, Challenges, and Open Issues", *Array*, Vol. 10, pp. 100057-100065, 2021.

[11] S. Iftikhar, A. Koucheryavy and A.A. Abd El-Latif, "Deep Learning-Based Pedestrian Detection in Autonomous Vehicles: Substantial Issues and Challenges", *Electronics*, Vol. 11, No. 21, pp. 3551-3559, 2022.

[12] J. Wang, X. Hui and X. Chen, "Research on Improved YOLOv5 for Low-Light Environment Object Detection", *Electronics*, Vol. 12, No. 14, pp. 3089-3095, 2023.

[13] KITTI Dataset, "The Latest in Machine Learning", Available at https://paperswithcode.com/dataset/kitti, Accessed in 2023.

[14] Rohit Halder and Rajdeep Chatterjee, "CNN-BiLSTM Model for Violence Detection in Smart Surveillance", *SN Computer Science*, Vol. 1, pp. 1-13, 2020.

[15] D. Avola, A. Fagioli and A. Mecca "A Novel GAN-Based Anomaly Detection and Localization Method for Aerial Video Surveillance at Low Altitude", *Remote Sensing*, Vol. 14, No. 16, pp. 4110-4114, 2022.

[16] M. Tahir and S. Anwar, "Transformers in Pedestrian Image Retrieval and Person Re-Identification in a Multi-Camera Surveillance System", *Applied Sciences*, Vol. 11, No. 19, pp. 9197-9203, 2021.

[17] J. Arunnehru, G. Chamundeeswari and S.P. Bharathi, "Human Action Recognition using 3D Convolutional Neural Networks with 3D Motion Cuboids in Surveillance Videos", *Procedia Computer Science*, Vol. 133, pp. 471-477, 2018.

[18] L. Wang and L. Van Gool, "Temporal Segment Networks for Action Recognition in Videos", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 56, No. 1, pp. 1-14, 2017.

[19] Rati Goel, "The Implementation of Image Enhancement Techniques using Matlab", *Proceedings of the International Conference on Innovative Computing*, pp. 1-5, 2021.

[20] G. Li and X. Qu, "Pedestrian Detection based on Light Perception Fusion of Visible and Thermal Images", *Optics and Laser Technology*, Vol. 156, pp. 108466-108475, 2022.

[21] H. Lv, Z. Zhou and J. Jing, "YOLOv5-AC: Attention Mechanism-Based Lightweight YOLOv5 for Track Pedestrian Detection", *Sensors*, Vol. 22, No. 15, pp. 5903-5909, 2022.

[22] J. Wei, S. Su, X. Tong and W. Gao, "Infrared Pedestrian Detection using Improved UNet and YOLO through Sharing Visible Light Domain Information", *Measurement*, Vol. 221, pp. 113442-113449, 2023.

[23] Q. Hu, Y. Zhang and Z. Hu, "Pedestrian Target Detection on High-Speed Pavement using Improved YOLOv5", *Proceedings of International Conference on Digital Image Processing*, pp. 1-8, 2023.

[24] Vidyabharathi Dakshinamurthi and S.K. Riyaz Hussain, "Deep Learning-based Image Dehazing and Visibility enhancement for improved Visual Perception", *ICTACT Journal on Image and Video Processing*, Vol. 14, No. 2, pp. 3122-3128, 2023.

[25] Q. Liu, T. Xie and X. Duan, "A Multitask Model for Realtime Fish Detection and Segmentation based on YOLOv5", *PeerJ Computer Science*, Vol. 9, pp. 1262-1269, 2023.