MM_FAST_RCNN_RESNET: CONSTRUCTION OF MULTIMODAL FASTER RCNN INCEPTION AND RESNET V2 FOR PEDESTRIAN TRACKING AND DETECTION

Reference NO. IJME 1381, DOI: 10.5750/ijme.v1i1.1381

Johnson Kolluri*, Department of CSE, National Institute of Technology Mizoram, Aizawl, Mizoram, India, **Sandeep Kumar Dash,** Department of CSE, National Institute of Technology Mizoram, Aizawl, Mizoram, India and **Ranjita Das,** Department of CSE, National Institute of Technology Mizoram, Aizawl, Mizoram, India

*Corresponding author. Johnson Kolluri (Email): (johnson.kolluri@gmail.com)

KEY DATES: Submission date: 20.12.2023 / Final acceptance date: 27.02.2024 / Published date: 12.07.2024

SUMMARY

Pedestrian identification and tracking is a crucial duty in smart building monitoring. The development of sensors has led to architects' focus on smart building design. The image distortions caused by numerous external environmental factors present a significant problem for pedestrian recognition in smart buildings. It is difficult for machine learning algorithms and other conventional filter-based image classification methods, such as histograms of oriented gradient filters, to function efficiently when dealing with many input photos of pedestrians. Deep learning algorithms are now performing substantially better when processing an enormous amount of image data. This article evaluates a novel multimodal classifier-based pedestrian identification method. The proposed method is Multimodal Faster RCNN Inception and ResNet V2 (MM Fast RCNN ResNet). The collected attributes address a tracking problem and establish the foundation for several object recognition tasks (novelty). Our method's neural network is regularized, and the feature representation is automatically adjusted to the detection assignment, resulting in high accuracy (superior to the proposed method). The proposed method is assessed using the PenFudan dataset and contemporary techniques regarding several factors. It is discovered that the recommended MM Fast RCNN ResNet obtains precision, recall, FPPI, FPPW, and average precision of 0.9057, 0.8629, 0.0898, and 0.0943.

KEYWORDS

Pedestrian detection, Inception, Neural network, Convolution layer, Pooling layer, Tracking

NOMENCLATURE

VOC	Visual Object Classes
NMI	Non-Maximum Inhibition
F	Frequency
ASIFT	Affine Scale Invariant Feature Transform

1. INTRODUCTION

Automatic numbering systems must not be used. A wide range of applications is made possible by accurate pedestrian recognition and tracking in surveillance photos, including recognizing anomalous behavior, path prediction, intruder identification, public safety on mobile platforms, and crowd counting [1]. Significant improvements in the detection of pedestrians have just been reported. Unfortunately, it is not easy to directly apply these conventional approaches to difficult monitoring photographs. This is because of the intrinsic characteristics of these security applications, which frequently use many cameras to cover large areas adequately. Such circumstances place strict limitations on the hardware: Wide-angle, low-cost cameras are used, positioned high for a partially downward angle bird's eye view [2]. Therefore, it is not easy to comprehend and analyze these pictures. Indeed, surveillance photographs are frequently compressed and taken at low resolution. Traditional background subtraction-based object identification algorithms produce fewer results when using these significant compression ratios [3].

Additionally, these methods do not distinguish between individuals and other items. Typical pedestrian detectors, trained and assessed on forward-looking photos, are similarly unable to yield appropriate detection accuracy on these images because of their particular point of view (and wide-angle lens). Additionally, a few of the pedestrians appear very little in the image due to perspective distortions, which remains among the most challenging difficulties for pedestrian trackers now in use [4].

Real-time computing rates are also necessary. The challenge arises from the reality that pedestrians are distinct from physical quantities but desire to move from the available present in the shortest amount of time in an emergency or hostile environment [5]. The presence of

pedestrians in an emergency or violent atmosphere has irregular escape spaces and fluctuates positions due to variations in their common posture, clothing, lighting, and background. This results in a very challenging judgment of their position detection and emergency exits. Moreover, data-driven approaches have recently gained traction [6]. It has significantly advanced associated algorithms for monitoring and determining the best exit strategy for the mass panic crowd. The most sophisticated target identification techniques are built on the premise of boundary boxes, the resampling of pixels or characteristics for each box, and the use of an elevated classifier during object identification [7].

The PASCAL Visual Object Classes (VOC), Conventional Objects in Context (COCO), and ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) datasets were used to apply the method after that for evaluation. The most sophisticated algorithms generally consist of Faster R-CNN (convolutional neural network), which was designed depending on Fast R-CNN [13], You Only Look Once (YOLO), Single Shot MultiBox Detector (SSD), etc. The technique has a very high detection rate but also a very expensive hardware requirement and a very slow detection time. Only 7 frames per second, even for the highest high detectors, can operate. Many researchers have worked to enhance the Fast R- CNN performance. However, increasing speed comes at the expense of decreased detection performance. Thus, it is unable to reach the optimal state. Single-shot detection methods include YOLO and SSD [14]. The YOLO model views the challenge of object recognition as a probability-like prediction problem with a bounding box that is physically separated. When evaluating the entire image, Border boxes and quasi-probabilities can be designed to predict by a single neural network.

The SSD framework applies a basic network with several feature levels that forecast the alignments to default boxes of various scales. The anchor boxes used for the Faster R-CNN [15] are comparable to the standard boxes of SSD. Multimodal representations are distributed vectors that combine several informational modalities into a unified mathematical space, with the similarities of the samples being indicated by the distances between them. When creating multimodal spaces, using a single unimodal representation technique for each Mode is typically adequate. Since combining complementary data from visible and infrared sensors provides reliable human target detection in both daytime and nighttime environments, multimodal pedestrian detection has recently attracted much attention. Although the availability of modalities gives additional information, learning from unimodal data presents two significant challenges: 1) models must learn the complicated intramodal and cross-modal interactions for predictions, and 2) models must be adaptable to unexpectedly missing or noisy modalities during testing.

Occlusion has been one of the biggest obstacles to pedestrian recognition in complicated contexts in reality, particularly in crowded settings during abnormal or real emergencies. As a result, this increases the responsive detection method based on a non-maximum inhibition (NMI) threshold in the busy scene. It is impossible to fix the dense occlusion issue by changing the NMS threshold. While a low NMS threshold results in many failed checks, a big NMS threshold causes numerous false checks.

Work contributions are as follows:

It is suggested that the network incorporate the benefits of feature fusion and residue architecture in quicker R-CNN and that it be optimized from two sections of the Feature generation module using RoI pooling design and Weight-aware module.

- The data points are extracted by scaling and rotating for each labeled pedestrian in the dataset.
- Constructing the feature generation is concerned with detecting valuable local and global trends to create new features as a supplement to natural features.
- The ResNet V2 is optimized, and the detection rate of various models is examined. The approaches of residual structure, feature fusion, and hole convolution are suggested for optimization.
- The feature recalibration is accomplished by multiplying the source convolutional features by the channel weight coefficients.
- The rest of the essay is structured as follows: Literature surveys are described in Section 2. Section 3 explains the proposed approach. A discussion of performance assessment is included in Section 4. Section 5 presents a conclusion and future projects.

2. RELATED WORKS

In general, the detection and tracking of pedestrians is a particularly active research area. Initially suggested using HOGs (Histograms of Oriented Gradients) to identify pedestrians. Even now, most advanced pedestrian detectors still depend on HOG features, but in a more subdued fashion, thanks to their discoveries that opened the door for numerous derivative techniques (e.g., in combination with other features). Unimodal sets of numbers are those with only one Mode, bimodal sets of numbers have two Modes, trimodal sets of numbers have three Modes, and multimodal sets of numbers have four or more Modes.

2.1 PEDESTRIAN DETECTION AND TRACKING USING NEURAL NETWORK

To resolve the shortcomings of the prior methodology, [16] offer the new resilient Scale Illumination Rotation and Affine invariant Mask R-CNN (SIRA M-RCNN). The suggested program's initial stage addresses brightness variation through histogram analysis. Additionally, this

work creates the rotationally invariant features using the directional filter bank and the contourlet transformation. To locate translational and scale-invariant points, this work then utilizes the Affine Scale Invariant Feature Transform (ASIFT).

In [17], the logic underlying our hypothesis can be used to support the progression of identification methods from the earliest fully convolutional neural networks (FCNNs), such as the Faster R-CNN, to the latest configuration techniques, and it can also assist us in comprehending why certain architectures seem to be better than others. Although it is significantly simpler on certain techniques than others, our approach has the advantage that it can be implemented in the majority of the traditional models with some modifications.

All original photos in [18] were cropped using a sliding window to achieve pre-detection findings for this paper. To create multi-scale photos, the pictures are cropped once more to place the object in the center using label files distributed in the same scene. The superfluous shredded boxes brought on by picture cropping are eventually eliminated using the region Normalized Mean Subtraction technique, an integration method about the outcomes of converting small images into big pictures.

Novel methods were presented in [19] to merge two of the most effective deep learning models. The Mask R-CNN is widely acknowledged as the best effective deep learning model for two-stage object recognition. Although this method uses high accuracy, it has drawbacks like slow detecting speed and high computing cost. A more lightweight Mask R-CNN with MobileNetV2 structure has been created to address this issue. A Convolution Operation has been substituted by the Depth-wise Separable Convolution Operation inside the Region Proposal Network (RPN) to activate the Mask R-CNN light.

Weak segmentation masks autonomously produced by depth pictures are proposed in [20] using spatial-contextual deep network architecture (SC-DNN). This allows you to recognize pedestrians and perform joint semantic segmentation using only real-world boundary boxes for training. The research shows that the pedestrian detector performs better due to this cooperative training. Furthermore, research demonstrates that combining the segmentation masks generated and the classification network's outputs improves detection capability even further.

2.2 MULTIMODAL CONCEPT ON PEDESTRIAN DETECTION

New spatial-contextual deep network design in [21] can effectively utilize multimodal data. Extracting features from the two modalities comprise 2 unique deformable ResNeXt-50 encoders. A neural network-based unit and many groups of a combination of Graph Attention Networks make up a multimodal feature embedding module (MuFEm), in which the fusion of these two encoded characteristics occurs. Two CRFs receive the outcomes of MuFEm's last feature fusion unit handed to them for spatial refinement.

The multimodal data YOLOv3 (MDY) method is used in [22] for embedded device detection and recognition. By optimizing anchor frames and including short target detection branches, the Multiple Dyuonate Yield method enhances pedestrian detection performance using YOLOv3 as its fundamental framework. The method is then sped up using TensorRT technology to enhance embedded devices' real-time performance.

Introduce an attention-guided multimodal and multi-scale fusion (AMSF) module in [23] to synchronously compatible local traits as an example dispersed across multimodal and multi-scale layers and flexibly combine with fine-grained attention to properly exploit multiple modalities for superior multi-scale prediction results.

Using all available multimodal characteristics, [24] proposes a cross-modal object tracking network based on Gaussian Cross Attention (GCANet) to enhance detection capability. The important multimodal characteristics are successfully highlighted through the bidirectional coupling of local features from distinct modalities, increasing the detection performance by realizing feature engagement and fusion between different modalities.

Using a Separation and combining technique, [25] proposes a cross-modal feature learning (CFL) module to systematically examine both the common and modality-specific concepts of matched RGB and infrared pictures. To learn the cross-modal interpretations at various semantic levels, research integrates the suggested CFL module into many layers of a two-branch-based pedestrian recognition network. The multimodal network is developed end-to-end by jointly maximizing a multi-task loss by adding a separation-based secondary task.

In [26], the authors offer a unique single-stage detection system that uses multi-label learning to acquire input state-aware characteristics by tagging each input image with a different label based on its current state. They also describe a brand-new augmentation technique synthesizing unpaired multispectral pictures using geometric modifications. Table 1 shows the Comparison of several approaches.

The researchers primarily resolved the relevant issues using variable component models because of the insufficient data for pedestrians under occlusion. The number of model simulations has increased significantly even though the detection performance has improved slightly. Using partial model-assisted global identification is an efficient way to improve pedestrian recognition under interference, but it comes with a higher computing cost and slower detection

Table 1. Comparison of several approaches

Author/ year	Method	Advantage	disadvantage
Gawande et al., [2022]	Scale Illumination Rotation and Affine invariant Mask R-CNN (SIRA M-RCNN)	Robust to photometric changes	Requires more computing resources to process the image
Saeidi et al;., [2022]	Fully convolutional neural networks (FCNNs)	It is possible to maintain the same geometric and optical deformation.	More expensive
Li et al., [2022]	Normalized Mean Subtraction technique	Suitable for crowded scenarios	Unsuitable for long-term prediction
Sahu et al., [2023]	Depthwise Separable Convolution Operation inside the Region Proposal Network	It is possible to maintain the grayscale rotation and invariance.	Complex and slow
Guo et al. [2021]	Joint semantic segmentation	More robust discrimination	Ignores the pixel distribution
Dasgupta et al., [2022]	Graph Attention Networks	The procedure slows down computational frame efficiency and produces observable outcomes at grey levels.	Cannot handle occlusion
Wang et al., [2022]	Multiple Dyuonate Yield method	It can effectively manage sceneries' dynamic variation.	High computational complexity
Bao et al. [2020]	Attention-guided multimodal and multi-scale fusion (AMSF)	For issues with shadows, it produces observable results.	Not suitable for scenarios with abrupt changes in lighting
Peng et al. [2020]	Gaussian Cross Attention (GCANet)	It is adaptive to dynamic variation of occlusions	High computational complexity
Liu et al. [2021]	Cross-modal feature learning (CFL) module	It effectively removes noise from the scenes	Less accuracy
Kim et al. [2021]	Multi-label learning	It achieves better local performance	More errors

speed. As a result, one of the primary areas for research in this technology is to increase the detector's recognition rate for obstructed pedestrians while preserving the recognition rate.

3. PROPOSES SYSTEM MODEL

Data may be extracted from all photos with a multimodal range after the compile time of the PenFudan dataset. After data points are retrieved, they are passed to the MultiModal Faster RCNN Inception and ResNet V2 feature creation module, where n features are extracted. In the fusion step, these collected characteristics are combined and provided to the weight-aware module for further recognition.

3.1 DATA POINTS EXTRACTION FROM PEDESTRIAN IMAGES

The monitoring photos look like the pedestrians are resized and rotated. Both factors solely rely on location in the image because the monitoring camera's orientation concerning the ground plane is constant. If each pixel position comes rotation and typical pedestrian height are known, $x=[x,y]$. This research can use this scene knowledge to quickly and precisely identify pedestrians. In order to employ a real-time, precise pedestrian detector, This research deforms each patch to an upright, fixed-scale image patch for each validly proposed region using the transformation factors. A one-time offline calibration is required to obtain these transformation factors. As described here, the scene calibration is simple to carry out and unnecessary. In order to do this, This research took the scaling and rotating for each labeled pedestrian in the dataset and extracted their rotation and height. After that, This research used a second-order 2D polynomial function to estimate the data points $f_i(x)$ for both parameters:

$$f_i(x) = p_0 + p_1 x + p_2 y + p_3 x^2 + p_4 xy + p_5 y^2 fi(x) \quad (1)$$

Both $f_{scale}(x)$ and $f_{rotation}(x)$ are used as Lookup functions (LUFs): the intended region attributes and transformation parameters are defined for each place in the picture. Figure 1 shows Proposed multimodal pedestrian tracking anddetection method.

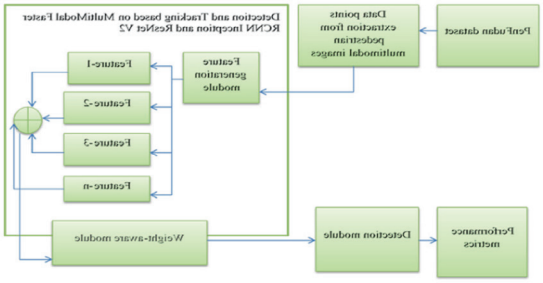

Figure 1. Proposed multimodal pedestrian tracking and detection method

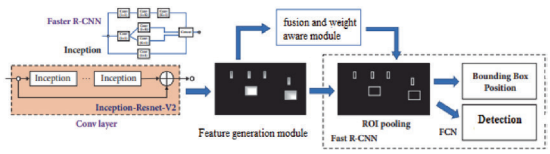

Figure 2. Architecture of faster RCNN inception and ResNet V2

3.2 MULTIMODAL FASTER RCNN INCEPTION AND RESNET V2-BASED DETECTION

The presentation of a multimodal fusion architecture for cooperative training in pedestrian identification and tracking is illustrated in Figure 2. It comprises three main parts: pedestrian recognition, segmentation monitoring, and feature learning/fusion.

3.3 FEATURE GENERATION MODULE

Feature generation is concerned with detecting valuable local and global trends to create new attributes as a supplement to raw attributes. Each sample is denoted as an embedding matrix $E \in R^{n_f \times k}$ via attributes embedding, where n_f is the number of fields and k is the embedding size. To simplify, modify the embedding matrix's shape as $E^1 \in R^{n_f \times k \times 1}$ as the input sequence for the initial convolutional layer. By convolving a matrix, a convolutional layer is created to collect the neighbor feature correlations $WC^1 \in R^{h^1 \times 1 \times 1 \times m_c^1}$ with nonlinear activation functions. Here, the result of the initial convolutional layer is represented as $C^1 \in R^{n_{f \times k \times m_c^1}}$. The convolutional layer can be constructed as follows:

$$C^1_{p,q,i} = tanh(\sum_{m=1}^{1}\sum_{j=1}^{h^1} E^1_{p+j-1,q,m} WC^1_{j,1,1,i} \tag{2}$$

$$\tan(h) = \frac{\exp(x) - exp(-x)}{\exp(x) + exp(-x)} \tag{3}$$

Where $C^1_{p,q,i}$ i represents the i-th feature map in the primary convolutional layer and p, q are the row and column index of the i-th feature map. Note that padding, used in practice, is not included in the formula above. A max-pooling layer is utilized, followed by the initial convolutional layer, to capture the most significant feature correlations and minimize the number of factors. This research introduces h_p as the altitude of pooling layers (width=1). The output in the starting pooling layer is $S^1 \in R^{\left(\frac{n_f}{h_p}\right) \times k \times m_c^1}$:

$$S^1_{p,q,i} = max\left(C^1_{p,h_p,q,i}, \ldots C^1_{p,h_p,+h_p-1,q,i}\right) \tag{4}$$

The source for the (i + 1)-th convolutional layer will consist of the pooling outcome of the i-th pooling layer: $E^{(i+1)} = S^i$.

Following, the first convolutional layer and pooling layer, $S^1 \in R^{\left(\frac{n_f}{h_p}\right) \times k \times m_c^1}$, contains the patterns of neighbor features. Global non-neighbor feature correlations will be disregarded if S 1 is considered as the created additional features because of the nature of CNN. In order to merge patterns in the nearby neighborhood features and create substantial new features, This research, therefore, constructs a completely connected layer. This research represents the weight matrix as $WR^1 \in R^{\left(\frac{n_f}{h_p}\right) \times kMC \times m_c^1}$ and the bias as BR^l where m_c^1 is the number of attributes maps in the initial CNN layer, and m_r^1 is the number of unique attributes map in the starting recombination Layer. Therefore, in the i-th reintegrating layer, $n_f / h_p^i m_r^i$ features are created.

$$R^1 = tanh\left(S^1.WR^1 + BR^1\right) \tag{5}$$

Consider there are n_c pooling, recombine, and convolutional layers, and $N_i = n_f / h_p^i m_r^i$ fields of features are produced by i-th round considered as R^i. The entire novel features $R \in R^{N \times k}$ (where, $N = \sum_{i=1}^{n_c} N_i$) produced by Feature Generation is calculated as:

$$R = \left(R^1, R^2, \ldots R^{n_c}\right) \tag{6}$$

Raw characteristics and new features are then combined to form

$$E = \left(E'^T, R^T\right)^T \tag{7}$$

where E' is the embedding matrix of raw features for the Fusion and Weight-aware module

3.4 FUSION AND WEIGHT-AWARE MODULE

Choosing the best discriminative features and their passage through the coarse-to-fine strategy is the key challenge in salient pedestrian identification. Nevertheless, combining

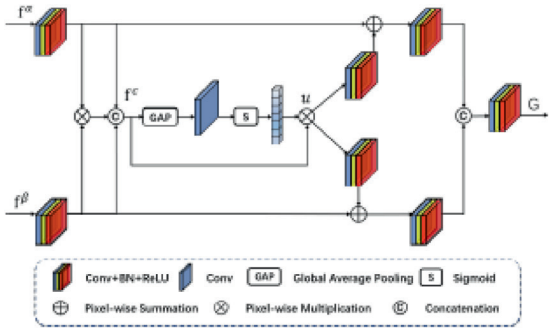

Figure 3. Weight -aware fusion layer

characteristics from various levels in an encoder-decoder way frequently results in either missing information or ambiguous features, preventing the network from optimizing. To achieve this, This research suggests a brand-new Weight-Aware Fusion module (WAF), which, as shown in figure 3, adaptively chooses the discriminative characteristics for object identification.

Initially, provide two input feature types $f^{\propto}, f^{\beta} \in R^{w' \times H' \times C'}$. This research employs pixel-wise multiplication to improve the frequent feature map pixels while reducing the unclear ones. A lightweight encoder is then used to combine the improved features with the modified features $\vartheta(.)$ It can be referred to as a:

$$f^{C} = \vartheta_{\alpha}\left(f^{\propto}\right) © \vartheta_{\beta}\left(f^{\beta}\right) © \left(\vartheta_{\alpha}\right) \otimes \vartheta_{\beta}\left(f^{\beta}\right)) \tag{8}$$

Where © and $\otimes$ represent the feature combination process and pixel-wise multiplication Correspondingly. Every encoder $\vartheta_{\{\alpha,\beta\}}$ typically consists of a 3 × 3 convolutional layer after a Batch Normalization and a ReLU activation. When combining the multi-level features, the features $\propto$ and f^{β} are initially upsampled to a similar scale. After acquiring a valuable feature $f^{C} \in R^{W' \times H' \times 3C'}$ by (1), the second major issue is choosing the extremely responsive features in the segmentation objective. So, This research suggests using global features for a situational comprehension of the attention weights, motivated by the channel attention process. The f^{C} is then compressed with a global average pooling, normalized using a sigmoid curve σ, and aligned with feature channels by turning into a vector shape. This sequential operation has the following form:

$$g = \frac{1}{W' \times H'} \sum_{i=1}^{W'} \sum_{j=1}^{H'} f_{i,j}^{c} \tag{9}$$

$$u_{i,j} = f_{i,j}^{c} \otimes \mu\left(\tau_{c}\left(g_{i,j}\right)\right) \tag{10}$$

The u symbol represents the learned concentration weighted features, and the τ is a linear transformation to reorder the pooling features. Therefore, characteristics pertinent to the salient target may stand out in all source characteristics $f^{\propto}$ and f^{β}. A Weight-aware attention mechanism can accomplish this. The weight coefficients w^{aving} and w^{max} of these two compression methods are evaluated as follows:

$$w^{avg} = fc2\left(fc1\left(GAP(F)\right)\right) \tag{11}$$

$$w^{max} = fc2\left(fc1\left(GMP(F)\right)\right) \tag{12}$$

Where $F \in R^{H \times W \times N}$ is a feature map; *H* and *W* denote the height and width of attributes maps, correspondingly; *N* refers to the number of feature channels; and $fc2(\cdot)$ and $fc1(\cdot)$ is the fully connected layers, with the aim of better fitting the intricate relationships between feature channels. A nonlinear unit between two completely linked layers is applied using the ReLU activation function. $GAP(\cdot) and GMP(\cdot)$ represent the global max pooling function and average pooling function. The relevance of each feature channel is then differentiated using sigmoid activation only after two weight vectors have been added element by element. The feature recalibration is accomplished by multiplying the source convolutional features by the channel weight coefficients, or w. The particular calculations are written as follows:

$$w = \sigma\left(w^{avg} + w^{max}\right) \tag{13}$$

$$\mathrm{F}' = \mathrm{w.F} \tag{14}$$

where w denotes the last channel correlation coefficients, σ represents the sigmoid activation function, and F' is a weighted feature map.

3.5 DETECTION MODULE

The classification method assigns the image features that were extracted from the pictures. ResNet V2-based recognition carries over the classification process. Assume the feature map f_{in_i} extracted by ResNet V2 at layer *i*with dimension $h_i \times \omega_i \times c_i, i \in N$. Let $F_{1,2}^{i}$ be the new function that completes the mapping from f_{in_i} to f_{out_i}, shown as:

$$F_{1}^{i} = f_{in\_i} \rightarrow f'_{in\_i}, F_{2}^{i} = f'_{in\_i} \rightarrow f_{out\_i} \tag{15}$$

Where the two similar operations, F_{1}^{i} and F_{2}^{i}, are combined of two operations accordingly. Parameters $H and W$ are the height and width of the input image. The proposed $loss_{rep}$ is to keep the two calculated boxes $(pred\,box_1 and\ pred\,box_2)$ distant from other nearby pedestrians' points of contact. There are just two classes present here: pedestrian and background. Thus the parameter K, which determines the number of pedestrians one suggestion anticipates, should be set to 2. Let $P = \{P_1, P_2, \ldots. P_n\}$ shows the suggestions that have retracted positively from anchors. The $loss_{rep}$ is represented as:

$$loss_{rep} = loss\left(gt_3, predbox_1\right) + loss\left(gt_3, predbox_2\right) \tag{16}$$

Simly, for the n-th suggest P_n, its repugnance except for its two authorized targets, ground truth is described as the pedestrian reality for which it shares the biggest IoU area.

$$G_{rep}^{p_n} = argmax\, IoU\left(G, P_n\right) \tag{17}$$

Where $G=\{G\}$ is represented as the collection of all fundamental truths in the picture, $G_{1,2}^{P_n}=argmax\,IoU\left(G,P_n\right)$ are the two ground truth pedestrians according to the suggested P_n. The $loss_{rep}$ is between the ground truth $G_{rep}^{P_n}$ and two identified boxes $B_{1,2}^{P_n}$ retracted by P_n, shown as:

$$loss_{rep}=\frac{\sum_{P_n\in p}smooth_{ln}\left(IoG\left(B_{1,2}^{P_n},G_{rep}^{P_n}\right)\right.}{2(P)} \tag{18}$$

Where,

$$IoG\left(B,G\right)=\frac{area\left(B\cap G\right)}{area\left(G\right)}\in[0,1]$$

And

$$smooth_{ln}=\begin{cases}-\ln\left(1-x\right) & x\le\rho\\ \dfrac{x-\rho}{1-\rho}-\ln\left(1-\rho\right) & x>\rho\end{cases} \tag{19}$$

$smooth_{ln}$ is a smoothed ln function, which can be differentiated continually in (0,1). The parameter ρ is the smooth parameter to modify the repulsion loss's sensitivity to outliers. The greater the proposal seems to intersect with a non-target ground truth pedestrian, according to Equations (11) and (13), the larger penalty the $loss_{rep}$ will enhance the bounding box regression model. The $loss_{rep}$ can successfully block the anticipated bounding boxes from migrating to nearby pedestrians who are not their objectives. As a result, this research anticipates a higher $loss_{rep}$, which indicates that the predicted boxes are different from other ground facts and makes it easier to identify substantially obscured pedestrians.

4. PERFORMANCE ANALYSIS

The experiment was conducted in the PYTHON tool. The metrics used for evaluation are precision, recall, false positive per image (FPPI), false positive per window (FPPW), average precision, PR curve, and ROC curve. The spatial-contextual deep network architecture (SC-DNN) [20] and the attention-guided multimodal and multi-scale fusion (AMSF) [23] are two widely used methods that are contrasted with the proposed MM Fast RCNN ResNet.

Recall: It gives a ratio between the value of the right forecast and all possible predictions. In the equation, it is stated.

$$Recall=\frac{TP}{TP+FN}$$

FPPI: The FPPI curve, whose value is nearer to the classifier's actual use, depicts the average number of successful picture retrievals.

Receiver operating characteristic (ROC): To quantify the impact of statistical distribution on the approach's effectiveness, a curve was created with the horizontal axis, the genuine positive ratio, and the false positive rate as the vertical axes.

Dataset description- In this evaluation, the Penn-Fudan Database was created to identify human involvement in these experiments described in [27]. 345 humans have been identified in 170 shots; 96 of these pictures were taken close to the University of Pennsylvania, and the other 74 were taken near Fudan University. Since they were taken in academic and metropolitan street settings, every picture will have at least one person.

Precision: The proportion of true positive values to all anticipated values is provided. In the equation, it is indicated. Table 2 shows an analysis of precision

$$\text{Precision}=\frac{TP}{TP+FP}$$

Figure 4 compares the suggested MM Fast RCNN ResNet with the conventional SC-DNN, AMSF in precision. The precision values and the number of frames are displayed on the X and Y axes correspondingly. When compared, the suggested MM Fast RCNN ResNet technique obtains 0.9057 precision, which is 0.1447 greater than the SC-DNN technique and 0.0387 better than the AMSF method. The conventional SC-DNN and AMSF techniques reach 0.761 and 0.867 precision, accordingly. Table 3 shows an analysis of recall.

Figure 5 compares the suggested MM Fast RCNN ResNet recall model to the conventional SC-DNN, AMSF. Accordingly, the number of frames and recall values are

Table 2. Analysis of precision

Number of frames	SC-DNN	AMSF	MM_ Fast_RCNN_ResNet
10	0.75	0.85	0.92
20	0.74	0.841	0.91
30	0.756	0.854	0.90
40	0.76	0.82	0.91
50	0.75	0.83	0.90

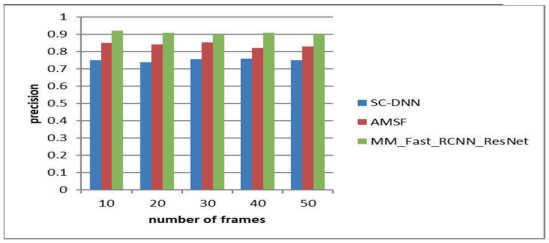

Figure 4. Comparison of precision

displayed on the X and Y axes. When contrasted, the suggested MM Fast RCNN ResNet technique obtains 0.1099 of recall, which is 0.1447 greater than the SC-DNN technique and 0.0509 better than the AMSF method. The conventional SC-DNN and AMSF techniques obtain 0.753 and 0.812 recall correspondingly. Table 4 shows an analysis of FPPI.

Table 3. Analysis of recall

Number of frames	SC-DNN	AMSF	MM_Fast_ RCNN_ResNet
10	0.73	0.82	0.87
20	0.72	0.81	0.86
30	0.75	0.80	0.89
40	0.71	0.85	0.87
50	0.70	0.86	0.88

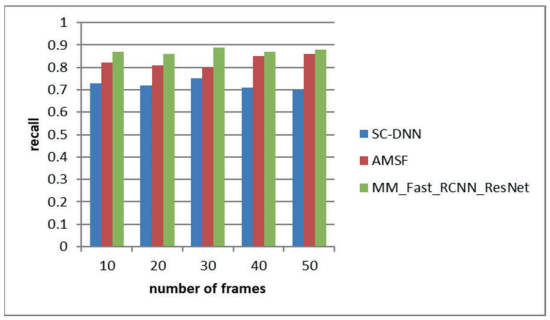

Figure 5. Comparison of recall

Table 4. Analysis of FPPI

Number of frames	SC-DNN	AMSF	MM_Fast_ RCNN_ResNet
10	0.74	0.83	0.95
20	0.76	0.83	9.843
30	0.78	0.84	0.861
40	0.71	0.821	0.84
50	0.73	0.85	0.82

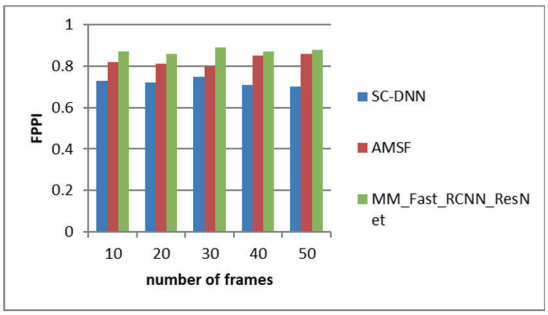

Figure 6. Comparison of FPPI

Figure 6 shows the FPPI evaluation of the proposed MM Fast RCNN ResNet and the existent SC-DNN, AMSF. The FPPI values and the number of frames are accordingly displayed on the X and Y axes. The suggested MM Fast RCNN ResNet approach obtains 0.0898 of FPPI, which is 0.6532 greater than the SC-DNN technique and 0.7032

Table 5. Analysis of FPPW

Number of frames	SC-DNN	AMSF	MM_Fast_ RCNN_ResNet
10	0.091	0.098	0.097
20	0.089	0.0979	0.0984
30	0.0891	0.097	0.093
40	0.092	0.092	0.095
50	0.081	0.093	0.094

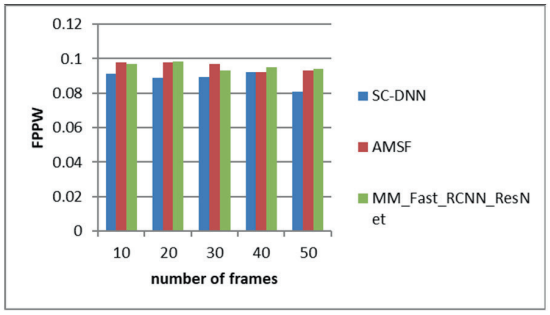

Figure 7. Comparison of FPPW

Table 6. Analysis of average precision

Number of frames	SC-DNN	AMSF	MM_Fast_ RCNN_ResNet
10	0.643	0.76	0.80
20	0.657	0.77	0.79
30	0.634	0.75	0.80
40	0.62	0.79	0.76
50	0.61	0.78	0.798

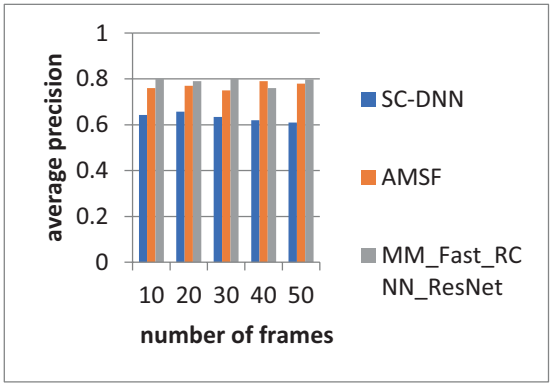

Figure 8. Comparison of average precision

better than the AMSF method compared to the existing SC-DNN and AMSF methods, which each obtain 0.743 and 0.793 of FPPI, correspondingly. Table 5 shows an analysis of FPPW.

The suggested MM Fast RCNN ResNet is compared with the existing SC-DNN, AMSF, and FPPW in Figure 7. The FPPW values and the number of frames are accordingly displayed on the X and Y axes. The suggested MM Fast RCNN ResNet technique obtains 0.0943 of FPPW, which is 0.0055 superior to the conventional SC-DNN method and 0.0048 better than the AMSF method. In comparison, the existing SC-DNN and AMSF techniques obtain 0.0998 and 0.0991 of FPPW, correspondingly. Table 6 shows analysis of average precision.

Figure 8 compares the average precision of the SC-DNN AMSF that is currently in use with the MM Fast RCNN ResNet that has been suggested. The average precision values and the number of frames are displayed on the X and Y axes, respectively. Comparatively, the SC-DNN and AMSF techniques now in use reach average precisions of 0.657 and 0.784, respectively. In contrast, the suggested MM Fast RCNN ResNet approach obtains an average precision of 0.807, 0.15 better than the SC-DNN technique and 0.18 better than the AMSF technique. Table 7 shows an analysis of RMSE.

Figure 9 contrasts the RMSE of the proposed MM Fast RCNN ResNet with the RMSE of the SC-DNN AMSF that exists in use. The X and Y axes show the RMSE values and the number of frames. In Comparison, the employed SC-DNN and AMSF procedures achieve RMSE of 0.43 and 0.59, respectively. However, the recommended MM Fast RCNN ResNet strategy achieves an RMSE of 0.24, which is 0.15 better than the SC-DNN technique and 0.18 better than the AMSF methodology.

Table 7. Analysis of RMSE

Number of frames	SC-DNN	AMSF	MM_Fast_ RCNN_ResNet
10	0.45	0.56	0.24
20	0.46	0.54	0.21
30	0.48	0.59	0.25
40	0.43	0.59	0.21
50	0.44	0.53	0.20

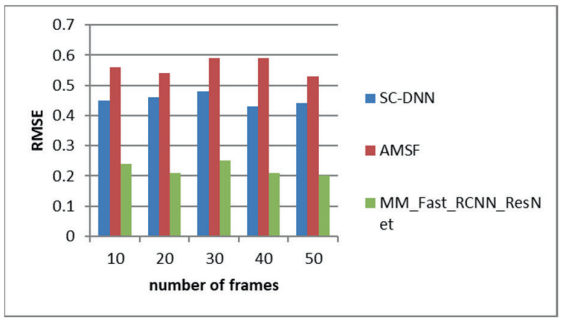

Figure 9. Comparison of RMSE

The precision-recall curve for the suggested MM Fast RCNN ResNet method is shown in Figure 10, where recall is indicated on the x-axis and precision is indicated on the y-axis. It is discovered that the curve begins to slope downward at precision values of 0.96, 0.94, and 0.6.

The suggested MM Fast RCNN ResNet method's ROC curve is shown in Figure 11, with the x-axis indicating false positive rates and the y-axis indicating true positive rates. The curve is discovered to begin rising at an FPR of 0.04 and TPR of 0.8.

Figure 12 and 13 shows examples of camera views in which pedestrian detection works well. In each image, people are relatively visible in the camera view, and the camera is positioned high enough to minimize occlusions. Table 8 shows the overall comparative analysis.

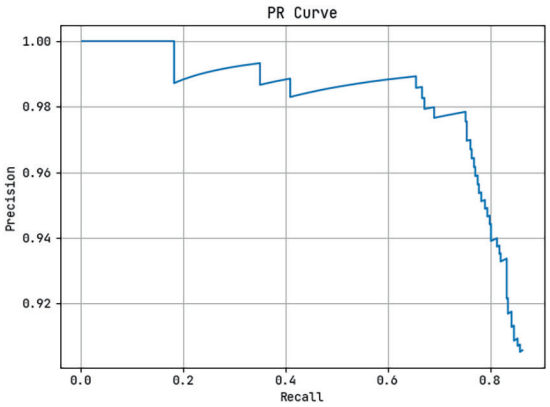

Figure 10. Comparison of PR curve

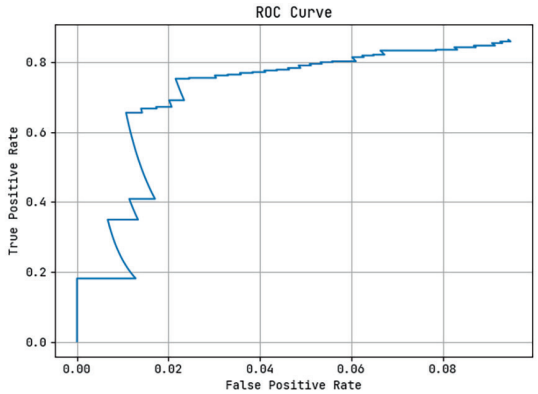

Figure 11. Comparison of ROC curve

Figure 12. Detection of humans

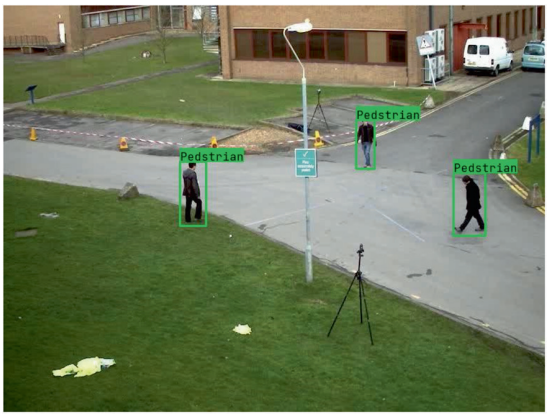

Figure 13. Pedestrian detection

Table 8. Overall comparative analysis

Parameters	SC-DNN	AMSF	MM_Fast_RCNN_ ResNet
Precision	0.761	0.867	0.9057
Recall	0.753	0.812	0.8629
FPPI	0.743	0.793	0.0898
FPPW	0.0998	0.0991	0.0943
Average precision	0.657	0.784	0.807
RMSE	0.43	0.59	0.24

5. CONCLUSION

The main challenge in pedestrian detection are identifying useful features that enable accurate and quick identification. This work used Multi-Modal Faster RCNN Inception and ResNet V2 (MM Fast RCNN ResNet) to overcome these issues using a unique parallel architecture. Convolutional characteristics and semantic features taken from the network are combined for better pedestrian detection. With the utilization of several modules, it is possible to distinguish between some confusing pedestrian assumptions that could be challenging to identify from the convolutional feature maps. This research used a pedestrian dataset to evaluate our system and demonstrate the effectiveness of our methodology. It is discovered that the suggested MM Fast RCNN ResNet obtains precision, recall, FPPI, FPPW, and average precision of 0.9057, 0.8629, 0.0898, and 0.0943. Future work focuses on examining the performance of segmentation and identification using the proposed technique with more sophisticated deep neural networks, such as the residual network.

6. REFERENCES

1. BEECK, K. V. and GOEDEMÉ, T. (2015, March). *Pedestrian detection and tracking in challenging surveillance videos.* In International Joint Conference on Computer Vision, Imaging and ComputerGraphics (pp. 356-373). Springer, Cham, https://doi.org/10.1007/978-3-319-29971-6_19
2. HU, W., TAN, T., WANG, L., MAYBANK, S. et al.(2004). *A survey on visual surveillance of object motion and behaviors*. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 34(3), 334-352, https://doi: 10.1109/TSMCC.2004.829274
3. ELHABIAN, S. Y., EL-SAYED, K. M., AHMED, S. H. et al. (2008). *Moving object detection in spatial domain using background removal techniques-state-of-art.* Recent patents on computer science, 1(1), 32-54, https://doi:10.2174/1874479610801010032
4. JIN, Y., ZHANG, Y., CEN, Y., LI, Y., MLADENOVIC, V., VORONIN, V. et al. (2021). *Pedestrian detection with super-resolution reconstruction for low-quality image.* Pattern Recognition, 115, 107846, https://doi.org/10.1016/j.patcog.2021.107846
5. ARGOUL, P. and KABALAN, B. (2017). *Pedestrian trajectories and collisions in crowd motion*. In Collisions Engineering: Theory and Applications (pp. 79-144), Springer, Berlin, Heidelberg, https://doi.org/10.1007/978-3-662-52696-5_6
6. ZHOU, M., DONG, H., WANG, F. Y., WANG, Q., YANG, X., et al. (2016). Modeling and simulation of pedestrian dynamical behavior based on a fuzzy logic approach. Information Sciences, 360, 112-130, https://doi.org/10.1016/j.ins.2016.04.018
7. ZHANG, H., NIU, M., CHEN, X., WU, J., ZHANG, Y., LIU, C., et al.(2022). *Target recognition and localization based on lightweight single-shot multibox detector network for robotics*. Journal of Electronic Imaging, 31(6), 061803, https://doi.org/10.1117/1.JEI.31.6.061803
8. M. EVERINGHAM, L. VAN GOOL, C. K. I. WILLIAMS, J. WINN, AND A. ZISSERMAN., et al. (2010). *The pascal visual object classes (Voc) challenge,* International Journal of Computer

Vision, vol. 88, no. 2, pp. 303–338, 2010, https://doi.org/10.1007/s11263-009-0275-4

9. T.-Y. LIN, M. MAIRE, S. BELONGIE et al. (2014). *Microsoft Coco: common objects in context," in Computer Vision – ECCV 2014, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds.,* Springer International Publishing, New York, NY, USA, https://doi.org/10.1007/978-3-319-10602-1_48
10. S. REN, K. HE, R. GIRSHICK, J. SUN, et al. (2017). *Faster R-CNN: towards real-time object detection with region proposal networks.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, doi: 10.1109/TPAMI.2016.2577031
11. J. REDMON, S. K. DIVVALA, R. B. GIRSHICK, A. Farhadi (2015) *You only look once: unified, real-time object detection*. in Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788, https://doi.org/10.48550/arXiv.1506.02640, Las Vegas, NV, USA.
12. W. LIU, D. ANGUELOV, D. ERHAN et al., *Ssd: single shot multibox detector. in Computer Vision–ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Springer International Publishing, New York, NY, USA, 2016, https://doi.org/10.1007/978-3-319-46448-0_2
13. R. GIRSHICK, "Fast R-CNN(2015). *in* The IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448, Santiago, Chile, December 2015.
14. SUN, S., YIN, Y., WANG, X., XU, D., WU, W., GU, Q. et al. (2018). *Fast object detection based on binary deep convolution neural networks.* CAAI transactions on intelligence technology, 3(4), 191-197, doi:10.1049/trit.2018.1026
15. C. M. BUKEY, S. V. KULKARNI, R. A. CHAVAN. (2017).*Multi-object tracking using kalman filter and particle filter. in* Proc. IEEE Int. Conf. Power, Control, Signals Instrum. Eng. doi: 10.1109/ICPCSI.2017.8392001
16. GAWANDE, U., HAJARI, K., GOLHAR, Y. et al. (2022). *SIRA: Scale illumination rotation affine invariant mask R-CNN for pedestrian detection*. Applied Intelligence, 1-19, doi:10.1007/s10489-021-03073-z
17. SAEIDI, M., and ARABSORKHI, A. (2022). *A novel backbone architecture for pedestrian detection based on the human visual system.* The Visual Computer, 38(6), 2223-2237, https://doi.org/10.1007/s00371-021-02280-6
18. LI, L., GUO, X., WANG, Y., MA, J., JIAO, L., LIU, F., LIU, X. et al. (2022). *Region NMS-based deep network for gigapixel level pedestrian detection with two-step cropping.* Neurocomputing, 468, 482-491, https://doi.org/10.1016/j.neucom.2021.10.006
19. SAHU, S., SAHU, S. P., DEWANGAN, D. K. (2023). *Pedestrian Detection Using MobileNetV2 Based Mask R-CNN.* In IoT Based Control Networks and Intelligent Systems (pp. 299-318), https://doi.org/10.1007/978-981-19-5845-8_22, Springer, Singapore.
20. GUO, Z., LIAO, W., XIAO, Y., VEELAERT, P., PHILIPS, W., et al. (2021). *Weak segmentation supervised deep neural networks for pedestrian detection*. Pattern Recognition, 119, 108063, https://doi.org/10.1016/j.patcog.2021.108063
21. DASGUPTA, K., DAS, A., DAS, S., BHATTACHARYA, U., YOGAMANI, S., et al. (2022). *Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving.* IEEE Transactions on Intelligent Transportation Systems, https://doi.org/10.48550/arXiv.2105.12713
22. WANG, C., LIU, Y. S., CHANG, F. X., LU, M., et al. (2022). Pedestrian detection based on YOLOv3 multimodal data fusion. Systems Science & Control Engineering, 10(1), 832-845, https://doi.org/10.1080/21642583.2022.2129507
23. BAO, W., HUANG, M., HU, J., XIANG, X., et al. (2022). *Attention-Guided Multi-modal and Multi-scale Fusion for Multispectral Pedestrian Detection.* In Chinese Conference on Pattern Recognition and Computer Vision (PRCV) (pp. 382-393), https://doi.org/10.1007/978-3-031-18907-4_30 Springer, Cham.
24. PENG, P., MU, F., YAN, P., SONG, L., LI, H., CHEN, Y., XU, T. (2022). GCANet: A Cross-Modal Pedestrian Detection *Method Based on Gaussian Cross Attention Network. In Science and Information Conference* (pp. 520-530), https://doi.org/10.1007/978-3-031-10464-0_35, Springer, Cham.
25. LIU, T., LAM, K. M., ZHAO, R., QIU, G., et al. (2021). *Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection*. IEEE Transactions on Circuits and Systems for Video Technology, 32(1), 315-329, DOI:10.1109/TCSVT.2021.306016
26. KIM, J., KIM, H., KIM, T., KIM, N., CHOI, Y. et al. (2021). MLPD: *Multi-Label Pedestrian Detector in Multispectral Domain*. IEEE Robotics and Automation Letters, 6(4), 7846-7853, DOI:10.1109/LRA.2021.3099870
27. WANG L., SHI J., SONG G., et al. (2007) *Object detection combining recognition and segmentation[C]. Asian Conference on Comput-3-er Vision*, 2007: 189-199, DOI:10.1007/978-3-540-76386-4_17

