I.J. Image, Graphics and Signal Processing, **2025, 1, 31-44**
Published Online on February 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2025.01.03

Multimodal Image Analysis Based Pedestrian Detection Using Optimization with Classification by Hybrid Machine Learning Model

Johnson Kolluri*
Department of CSE, National Institute of technology Mizoram, Aizawl, Mizoram, India
Email: johnson.kolluri@gmail.com
ORCID iD: https://orcid.org/0000-0001-8266-1766
*Corresponding Author

Ranjita Das
Department of CSE, National Institute of technology Agartala, Agartala, India
Email: ranjita.nitm@gmail.com
ORCID iD: https://orcid.org/0000-0001-6184-6294

Received: 11 April, 2023; Revised: 27 May, 2023; Accepted: 19 June, 2023; Published: 08 February, 2025

Abstract: In recent times People commonly display substantial intra-class variability in both appearance and position, making pedestrian recognition difficult. Current computer vision techniques like object identification as well as object classification has given deep learning (DL) models a lot of attention and this application is based on supervised learning, which necessitates labels. Multimodal imaging enables examining more than one molecule at a time, so that cellular events may be examined simultaneously or the progression of these events can be followed in real-time. Purpose of this study is to propose and construct a hybrid machine learning (ML) pedestrian identification model based on multimodal datasets. For pedestrian detection, the input is gathered as multimodal pictures, which are then processed for noise reduction, smoothing, and normalization. Then, the improved picture was categorized using metaheuristic salp cross-modal swarm optimization and optimized using naive spatio kernelized extreme convolutional transfer learning. We thoroughly evaluated the proposed approach on three benchmark datasets for multimodal pedestrian identification that are made accessible to the general public. For several multimodal image-based pedestrian datasets, experimental analysis is done in terms of average precision, log-average miss rate, accuracy, F1 score, and equal error rate. The findings of the studies show that our method is capable of performing cutting-edge detection on open datasets. proposed technique attained average precision of 95%, log-average miss rate of 81%, accuracy of 61%, F1 score of 51%, equal error rate of 59%.

Index Terms: Pedestrian Detection, Multimodal Image Analysis, Classification, Optimization, Hybrid Machine Learning, Salp Cross Modality

1. Introduction

In recent decades, pedestrian detection has become a popular computer vision research topic. In order to correctly locate individual pedestrian instances, a pedestrian detection system is needed to produce bounding boxes from photos recorded in various real-world surveillance scenarios. It has important features that make it easier to use a wide range of applications that focus on humans, like video surveillance and autonomous driving [1]. Even though significant advancements have been made over the past few years, developing a robust pedestrian detection solution that is ready for use in real-world situations is still a difficult task. It has been noted that the majority of currently available pedestrian detectors are trained solely on visible data, making their performance susceptible to changes in illumination, weather, and obstructions [2]. Numerous research efforts are devoted to creation of multimodal pedestrian detection solutions with goal of facilitating robust human target detection for use around clock [3]. The underlying assumption is that effective fusion of multispectral images, such as thermal and visible ones, can result in more robust and accurate detection results because they provide complementary data about objects of interest. A wide range of applications, such as intelligent driver assistance systems and video surveillance, rely on the ability to reliably identify individuals in

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

real-world settings [4]. In order to learn more about a specimen than is feasible from just one imaging modality, it is common practice to analyze together several pictures of the same specimen taken using various imaging modalities. Because the specimen is typically physically transported from one imaging instrument to another or because at least some alterations to the optical route are required that might affect the geometrical parameters of pictures, multi-modal imaging frequently necessitates image alignment. To better comprehend various specimen attributes or to improve observed object segmentation, the pictures from the various modalities might be aligned [5]. However, pedestrians are particularly difficult to spot due to their highly variable appearance caused by lighting, clothing, and shape characteristics that vary depending on the viewpoint. Because they dramatically alter the shape of a pedestrian, occlusions from carried items like backpacks and clutter in crowded scenes can make this task even more difficult [6].Making predictions and decisions is the emphasis of the branch of artificial intelligence (AI) known as machine learning. The main goal is to enable the computer to develop further without any assistance from humans; it will train using observations or data. The machine learning algorithms are frequently divided into two categories: supervised and unsupervised. Supervised machine learning applies past knowledge to new data using labeled examples.

Motivating the research: Although pedestrian detection research is not new, the issue of potential bias has not been examined in an algorithmic context. As a result, our main contribution is an algorithmic analysis of the pedestrian detection issue in terms of possible bias. People of all skin tones will feel safer if we look at which methods are less biased. Our paper can also help the automotive industry figure out how to improve their systems to take into account people of different skin tones so that drivers and passengers alike can feel safe on the road.

Contribution to the research:

- Due to human variation in appearance and posture, pedestrian detection frequently encounters significant intra-class variability. This study aims to develop hybrid machine learning methods for multimodal image dataset-based pedestrian detection.
- The image was classified and optimized using metaheuristic salp cross modality swarm optimization and naive spatio kernelized extreme convolutional transfer learning in this instance. To accomplish more precise outcomes, warm pictures have been broadly taken advantage of as integral data to help traditional RGB-based identification.
- There is a lack of research that focuses on examining features that are unique to each modality, despite the fact that numerous fusion strategies have been developed using the complementary features of existing methods.
- Because of this, the features that are unique to one modality can't be used to their full potential, and the fusion results could easily be dominated by the other modality, which limits the ability to discriminate.

2. Literature Review

In this section, Issue directing against pedestrian localization, scale prediction, and classification is decoupled from the task into Where, What, and whether using multi-modal learning in the proposed W3Net [7]. As a result, we examine the most recent research on pedestrian detection, both with and without multi-modal data, and contrast our approach to the most recent cutting-edge ones. The majority of popular pedestrian detectors are based on framework for general object detection, such as Faster RCNN and SSD, and they use pedestrian-oriented characteristics to deal with obstacles like occlusion as well as scale variation in pedestrian detection task [8]. Part-based techniques, such as DeepParts, FasterRCNN+ATT, or ORCNN, take full advantage of pedestrian part data, particularly visible body parts, in order to support robust feature embeddings [9]. TLL [10] breaks away from pre-defined anchors and proposes line localization after carefully determining that pedestrians typically maintain an upright posture. This greatly advances the development of pedestrian detection. Researchers have lately begun to pay attention to multi-modal learning, which provides the opportunity to identify correspondences between various modalities and develop a thorough understanding of natural events [11]. SDSRCNN [12] proposes a segmentation infusion network to enable combined supervision of semantic segmentation as well as pedestrian identification. A variant of the Faster RCNN architecture called F-DNN+SS [13] suppresses background proposals by adding pixelwise semantic segmentation as a post-processing step. To adjust the confidence in the detector proposals, F-DNN2+SS [14] make use of a semantic segmentation network and an ensemble learning strategy. Numerous specialists have involved different data for body parts, for example, leg, arm, and head, notwithstanding different ways to deal with distinguish walkers with the assistance of complete body signals [15]. However, in order to address problem of occlusion, researchers have classified pedestrian detection into two categories: specifically the deep neural network approach and conventional methods. Histogram of oriented flow and gradient, Haar wavelet, local binary variant, SVM [16], and other techniques are utilized in the conventional approach. to get features out. A deep CNN of finger-vein as well as finger shape modalities was proposed in Work [17]. The images of the fingers were taken by the near-infrared camera sensor. Weighted sum, product, and perceptron techniques were used to combine the features' matching distance scores. Additionally, in [18], face, fingerprint, and iris biometrics identification methods were designed utilizing DL template matching algorithms. Extracted features were combined by the weighted rank-level algorithm. Multi-level feature abstraction was used to combine the extracted elements of face, iris, and fingerprint modalities from multi-stream CNNs for identification in [19]. Both a deep belief network (DBN)-

based facial feature extraction as well as recognition method and a CNN-based left and right iris feature extraction as well as classification approach have been proposed in study [20].

Although the terms "classification" and "detection" are sometimes used interchangeably in articles that discuss pedestrian detection, they don't mean the same thing. The former relies primarily on a classifier and correlate feature-space, while the latter must deal with unknowns in position, size, and scale and depends on a variety of factors, including, but not limited to: hypothesis generation (e.g., clustering, salient region generation), hypothesis confirmation (a classifier), and post-processing (e.g., non-maximum suppression). However, the classification problem is specifically addressed in our study since we are curious to find out how the suggested model contributes [21, 22].

3. Proposed Pedestrian Detection Model based on Multimodal Datasets

In this section, a novel hybrid machine learning technique for multimodal image-based pedestrian detection is discussed. Here, naive spatio kernelized extreme convolutional transfer learning was used to classify the input image, and metaheuristic salp cross modality swarm optimization (MSCMSO) was used to optimize it. Figure 1 depicts the proposed model in detail.

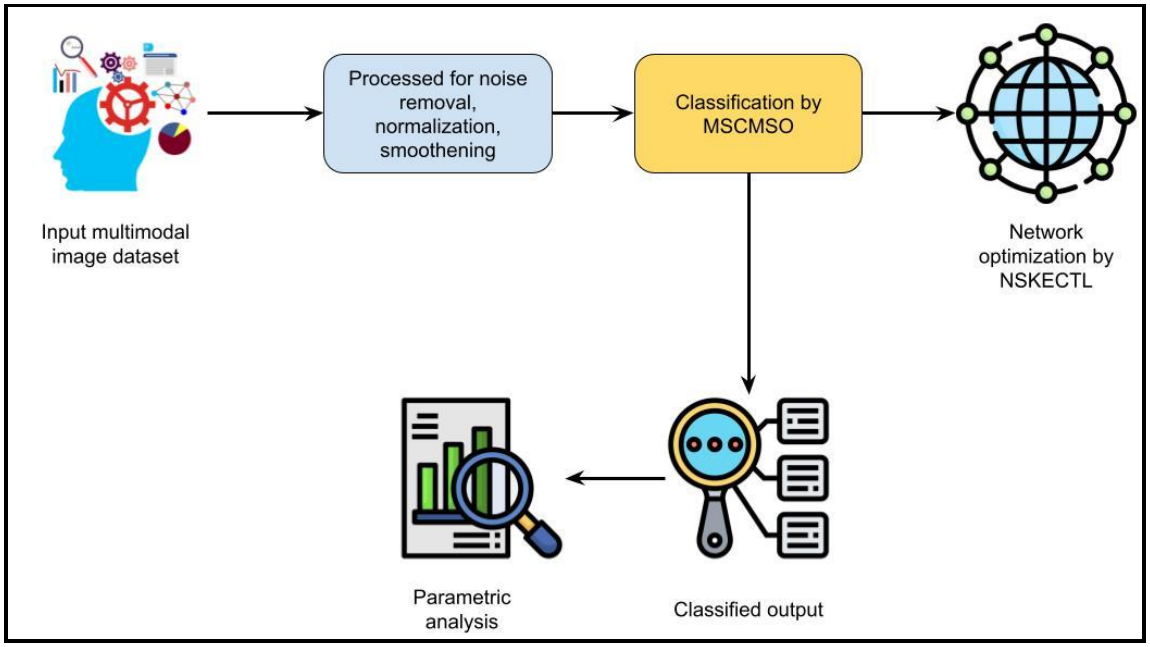

Fig. 1. The Proposed pedestrian detection model

Because random noise that is prevalent in infrared images typically hinders detection process, we first preprocess input images. Using default specifications of matlab command "imguidedfilter," we employ guided filtering technique to preserve as many structure details as possible in the original images while denoising because structure of clutters as well as targets is crucial for subsequent detection steps. After identifying mask object in input occluded facial image, Binary Mask Generation Network creates a binary mask. At Stage-I, binary mask lpre_mask is created using input image lc by the generator G1.

For reasons that will become clear shortly, eq. (1) will be adjusted slightly to utilize the l2 norm of the filter responses:

$$T = \sum_{k=1}^{K} q_k^{\,2} M_k \tag{1}$$

Where Mk is the filter orientation tensors and qk are answers from single scale loglets.

Boundaries denoising as well as smoothing are applied to original and mask images. Photographs are improved using the picture imperatives near to the picture skeleton after the skeletons of the source and cover pictures have been separated. In this work, the angle-based filtering technique is used for the boundary smoothing process, and scale-invariant intrinsic variables of edge pixel connection serve as filtering independent variable. Smoothness of local plane picture, as well as trend and degree of curvature of global edge vector, can also be reflected in the directional corner. Because it takes into account the geographic distance between the edge vectors, bilateral filtering alters angle of edge vectors further to application scenario and method. The discrete curve's intrinsic scale-invariant variables are written as eq. (2).

$$\Omega=\left\{\left(r_i,\theta_i\right)\right\}_{i=1}^{n-1}\begin{cases} r_i=l_i/l_{i-1}=\left\|\overline{v_{l+1}v}\right\|\Big/\left\|\overline{v_{l+1}v}\right\| \\ \theta_i\in(-\pi,\pi] \end{cases} \tag{2}$$

where$\|\overline{v_{\iota+1}v_\iota}\|/\|\overline{v_{\iota-1}v_\iota}\|$ is ratio of two edge vectors' modulus lengths.It has vertex vi in common, where i is directed angle of rotation of two nearby edge vectors. Positive rotation is anticlockwise. As an illustration, consider the vertex vi, and discrete representation of angle bilateral filtering is represented by eq. (3).

$$\theta_i^*=\begin{cases} \dfrac{\theta_1+c\left(v_2,v_1\right)s\left(\theta_2,\theta_1\right)\theta_2}{1+c\left(v_2,v_1\right)s\left(\theta_1,\theta_1\right)}, i=1 \\ \dfrac{c\left(v_{i-1},v_i\right)s\left(\theta_{i-1},\theta_j\right)\theta_{i-1}+\theta_i+c\left(v_{i+1},v_j\right)s\left(\theta_{i+1},\theta_j\right)\theta_{i+1}}{c\left(v_{i-1},v_i\right)s\left(\theta_{i-1},\theta_i\right)+1+c\left(v_{i+1},v_i\right)s\left(\theta_{i+1},\theta_i\right)}, i=2,3,\ldots\ldots,n-2 \\ \dfrac{c\left(v_{n-2},v_{n-1}\right)s\left(\theta_{n-2},\theta_{n-1}\right)\theta_{n-2}+\theta_{n-1}}{1+c\left(v_{n-2},v_{n-1}\right)s\left(\theta_{n-2},\theta_{n-1}\right)}, i=n-1 \end{cases} \tag{3}$$

The unique circumstance determines amount of weight factor σd, σr. Impact of distance size increases with the size of σd. The impact of the angle difference increases with increasing σr.

You can consider the predicted residue to be local noise. According to conventional estimating theory, if there are many estimations of the same signal, they should be combined using weights that are inversely proportional to the corresponding noise variances. This is the rationale for the suggestion that local certainty, c, be calculated as eq. (4) in practise.

$$c=\frac{\gamma\left(E_{\text{tot}}-E_{\text{res}}\right)}{E_{\text{tes}}+\gamma E_{\text{tot}}} \tag{4}$$

whereEtot is the total energy of the signal picked up by the filter set in use, Eres is the energy of the residue, and g is a measure of how far the actual filters in use vary from the ideal filter form. This probability estimate will fall between 0 and 1. By using, the estimate's sensitivity can be changed.

A. Na ïve Spatio Kernelized Extreme Convolutional Transfer Learning With Metaheuristic Salp Cross Modality Swarm Optimization:

Assume we have a detector model Mw with w as a parameter. $\mathcal{D}_{tr}=\{(x_i,y_i)\}_{i=1}^{n}$ and $\mathcal{D}_{\text{val}}$ are used to denote training set as well as validation set, respectively. Then, we represent loss function and data augmentation policy as two distributions. The loss function Lϕ and data augmentation policy Oθ are parameterized by θ and respectively. As stated in equations (5,6), our goal is to maximize method Mw's rewards R (Mw; Dval) with respect to θ and ϕ.

$$\theta,\varphi=\arg\max_{\theta,\varphi}R\left(M_{w^*};\mathcal{D}_{tal}\right) \tag{5}$$

$$s.t.w^*=\arg\min_w\sum_{(x,y)\in\mathcal{D}_{tr}}\mathcal{L}_\varphi\left(M_w\left(\mathcal{O}_\theta(x)\right),y\right) \tag{6}$$

If we think of hyper-parameters θ and ϕ weight of entire model as two optimization goals. The generally used pedestrian detection assessment metric, unlike genetic object detection, is log-average miss rate on false positive per image (FPPI) in $[10^{-2},10^{0}]$.and better it will be, lower it is. Our reward R (Mw; Dval) could be modeled as eq. (7) in order to guarantee that a higher reward can result in a relative lower MR:

$$R\left(M_w;\mathcal{D}_{val}\right)=1-MR\left(M_w;\mathcal{D}_{val}\right) \tag{7}$$

where M R() denotes the detection performance in terms of MR.

Let P: [0, 1] be a probability distribution on such that 0 P(A), 0 P(B), and, obviously, P() = 1 are all equal to 1. These incidents can be shown in a Venn diagram (Fig. 1a). The event that either A or B or both occur is indicated by the union of the events A and B, denoted by A B. We can state following equation (8,9) by assuming that |A| 6= 0 and |B| 6= 0:

$$P(A \mid B) = \frac{|A \cap B|}{|B|} = \frac{\frac{|A \cap B|}{|\Omega|}}{\frac{|B|}{|\Omega|}} = \frac{P(A \cap B)}{P(B)} \tag{8}$$

$$P(B \mid A) = \frac{|B \cap A|}{|A|} = \frac{\frac{|B \cap A|}{|\Omega|}}{\frac{A|}{|\Omega|}} = \frac{P(A \cap B)}{P(A)} \tag{9}$$

Eqs. 10 and 11 make it readily clear that

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A) \tag{10}$$

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{11}$$

Generalized Bayes' in equation (12) will be applied to each Ai.

$$P\left(A_i \mid B\right) = \frac{P\left(B \mid A_i\right)P\left(A_i\right)}{\sum_{j=1}^{n} P\left(B \mid A_j\right)P\left(A_j\right)} \tag{12}$$

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P\left(B \mid A^c\right)P\left(A^c\right)} \tag{13}$$

Due to total probability theorem (Eqs. 12 and 13) both eqs. 11 and 12 follow from Eq. 13. It is frequently believed that data can result from two opposing hypotheses, H1 and H2, with P(H1) = 1 P(H2), depending on the data. Equation (14) then gives posterior probability of the hypothesis (or model) H1.

$$P\left(H_1 \mid D\right) = \frac{P\left(D \mid H_1\right)P\left(H_1\right)}{P\left(D \mid H_1\right)P\left(H_1\right) + P\left(D \mid H_2\right)P\left(H_2\right)} \tag{14}$$

and eq. (15) represents the posterior probability of H2.

$$P\left(H_2 \mid D\right) = \frac{P\left(D \mid H_2\right)P\left(H_2\right)}{P\left(D \mid H_1\right)P\left(H_1\right) + P\left(D \mid H_2\right)P\left(H_2\right)} \tag{15}$$

From Equations 14 and 15, we derive Equation (16).

$$\underbrace{\frac{P\left(H_1 \mid D\right)}{P\left(H_2 \mid D\right)}}_{\text{posterior odds?}} = \underbrace{\frac{P\left(D \mid H_1\right)}{P\left(D \mid H_2\right)}}_{\text{Bayes factor } B_{12}} \cdot \underbrace{\frac{P\left(H_1\right)}{P\left(H_2\right)}}_{\text{prior odds?}} \tag{16}$$

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x, y)}{f_Y(y)} \tag{17}$$

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x, y)}{f_X(x)} \tag{18}$$

eq. (19) is utilized to express Bayes' theorem for continuous variables.

$$f_{X|Y}(x \mid y) = \frac{f_{Y|X}(y \mid x)\, f_X(x)}{f_Y(y)} \tag{19}$$

where $f_Y(y) = \int_X f_{Y|X}(y \mid x) f_X(x) dx = \int_{-\infty}^{+\infty} f_{XY}(x, y) dx$ because of total probability theorem.

Equation (20) is produced by applying Bayes' theorem (Eq. 17) and somewhat simplifying notation.

$$P\left(y_j \mid \mathbf{x}_i\right) = \frac{P\left(\mathbf{x}_i \mid y_j\right) P\left(y_j\right)}{P\left(\mathbf{x}_i\right)} \tag{20}$$

The model employs a fixed input image size of 224×224 pixels per channel and nine learnt layers. With 32 filters, convolutional layer serves as the network's foundational layer. An output feature map is created by convolving input volume with these 3×3 filters. Each layer's feature map size is determined using the formula in equation (21):

$$Feature\ Map\ out = \frac{W - F + 2P}{S} + 1 \tag{21}$$

where P is for padding borders, S for stride size, and W stands for the input volume size. In order to prevent over fitting and slower learning rates. Number of easily-learnable parameters in each layer that are impacted by the back propagation process are known as the trainable parameters of that layer. Learnable specifications in every layer are calculated using the algorithm in Equation (22). Only the convolutional and fully linked layers of network have these parameters. To decrease the dimensionality of the feature map and preserve the most important data, each pooling layer in suggested method conducts maximum sub sampling operation.

$$Trainable\ Parameters = ((filter\_size * Depth) + 1) * number\_of\_filters \tag{22}$$

Output from earlier convolutional layers is flattened before linking to final FC layers. Vector is first flattened, and then it is processed by an FC layer as well as 0.5-dropout layer, which removes 50 percent of network nodes at random to prevent overfitting issues. Output classification outcome of classes from associated datasets is represented in output layer of suggested network. Taught features are then fed into a brand-new, smaller classification method. Method has two FC layers to perform a classification process for multimodal recognition and learn more about a brand-new dataset.

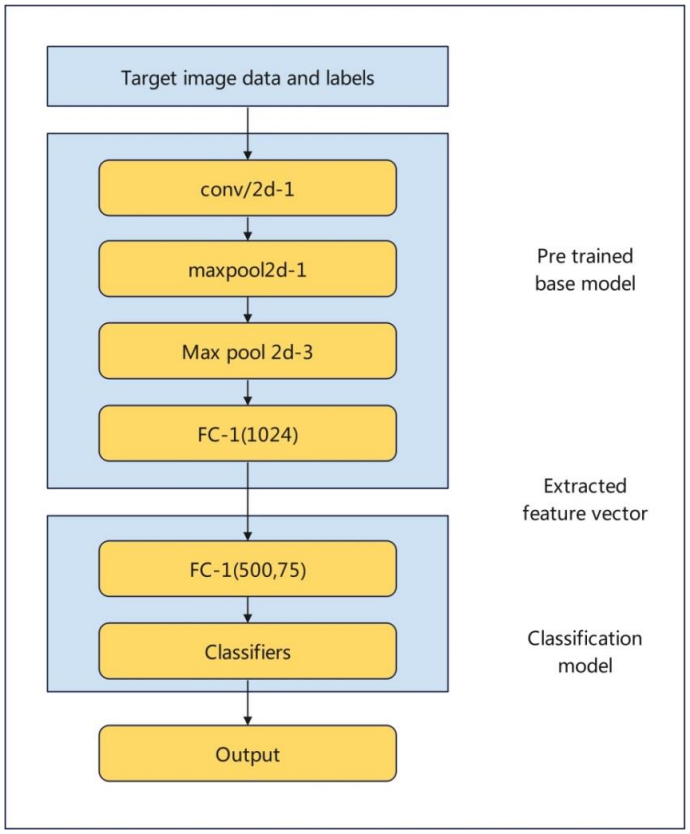

Fig. 2. Feature extraction from pre-trained model-based feature extractors and classification method

B. Metaheuristic Salp Cross Modality Swarm Optimization:

Suggested method combines derived features from two feature extractors and uses a feature-level fusion approach on all feasible combinations of facial and gait features to build most concatenated feature vectors possible for multimodal biometric detection. The classifier layer analyses the integrated feature vector during multimodal recognition.

"Leader" refers to the chain's front salp, and "followers" to the remaining salps. The variables and search space of the problem are represented by n dimensions, which are used to determine the salps' position. By looking for a food source, these salps can be used to pinpoint the swarm's intended target. The salp leader should update the position often, hence the following equation is employed to carry out this action:

$$x_j^1 = \begin{cases} F_j + c_1\left(\left(ub_j - lb_j\right)\times c_2 + lb_j\right)c_3 \le 0 \\ F_j - c_1\left(\left(ub_j - lb_j\right)\times c_2 + lb_j\right)c_3 > 0, \end{cases} \tag{23}$$

Additionally, specification c1 is a crucial coefficient in this algorithm because of how it balances the exploration and exploitation phases. It is derived using equation (24):

$$c_1 = 2e^{-\left(\frac{\pm}{l}\right)^2\Big)^2} \tag{24}$$

where the letters t and tmax stand for current iteration and maximum number of iterations, respectively. Following equation (25) is used by the SSA to begin updating the followers' position after updating the leader's position:

$$x_j^i = \frac{1}{2}\left(x_j^i + x_j^{i-1}\right) \tag{25}$$

In the j-th dimension, xi j is the position of i-th follower, where i is greater than 1. Lower boundary and higher boundary are first established for the benchmark function, while they are initially set for the benchmark dataset with a lower boundary of 0 and an upper limit of 1. Maximum number of iterations for the global optimisation problem is 500, whereas for the feature selection problem, it is 30. Last but not least, for the feature selection problem and the overall optimization problem, the population sizes are set at 20 and 50, respectively. The population size and iterations for the feature selection issue are set at modest values due to the complexity of the search space.

For the pre-owned benchmark dataset, the lower limit was initially set to 0 and the higher limit was initially set to 1 for the supplied information. The lower limit and upper limit are initially set in relation to the pre-owned benchmark capability. The maximum number of iterations for the global optimization problem is 500, whereas for the feature selection problem, it is 30. The final population size for the global enhancement issue is set at 50, while the include determination issue is set at 20. Due to the complexity of the search space as determined by eq. (26), the values of population size and iterations are chosen low for the feature selection issue.

$$F_D = F_R - F_T \tag{26}$$

$$F_{TD} = \sigma\left(GAP\left(F_D\right)\right)\text{e}\ \ F_T \tag{27}$$

$$F_{RD} = \sigma\left(GAP\left(F_D\right)\right)\text{e}\ \ F_R \tag{28}$$

4. Simulation Results and Discussion

We used backbone ResNet50 as well as Nvidia GTX1080Ti to implement proposed method in Pytorch. With 0.9 momentum and 0.0005 weight decay, we use Stochastic Gradient Descent (SGD) algorithm to improve network. Mini-batch for City peoples has two images, and we train the network for 30k iterations at a rate of 103 at the beginning and 104 after 6k iterations. The mini-batch for Caltech consists of eight images; we train network for 40 thousand iterations at an initial learning rate of 103, then decay it to 104 for 20 thousand iterations.

Based on the Cityscapes data, the diverse dataset City persons consists of 5000 images—2975 for training, 500 for validation, and 1525 for testing. It has 35,000 person and 13,000 ignore region annotations on a total of 5,000 images. Additionally, it observes that the train, validation, and test subsets all have the same person density. The Caltech Dataset is made up of approximately ten hours of video recorded at 640 x 480 pixels at 30 Hz from a car driving through normal traffic in an urban setting. An estimated 250,000 frames with 2300 distinct pedestrians and 350,000

bounding boxes were annotated. All of datasets have difficult settings, known as Heavy occlusion, in which probability of pedestrians' visible parts is less than 0.65. The KAIST dataset was utilized in our experiments. It was recorded both during the day and at night to account for shifts in light levels.

The test set is what we use to report the results. As a result, as part of ablation study, we use this dataset for a wide range of systematic experiments to determine the best configuration for our model. Momentum Optimizer is used to train the network, setting the momentum to 0.9 as well as initial learning rate to 0.01. After every 5k iterations, learning rate decreases by a factor of 10. Batch size was set to one, and Proposed method was trained on a system with two Nvidia Tesla P6 GPUs. On Nvidia Tesla P6 GPU, our end-to-end detection framework runs at 10 frames per second.

Parts of the network: A feature fusion unit combines outputs of two parallel Graph Attention Networks (GATs) that make up the MuFEm module's fundamental fusion unit. To gradually combine unimodal feature maps, we repeat these blocks. We fluctuate the quantity of combination units from 1 to 5. The SCoFA module's spatial and contextual components are altered for each of these configurations. Both of these are enabled simultaneously and separately during experiments. Our initial set of experiments shows that deformable convolution is significantly superior to regular convolution for these two encoders. As a result, we chose to use DCN early in the design process without conducting ablation studies on it with other network components to limit number of combinations.

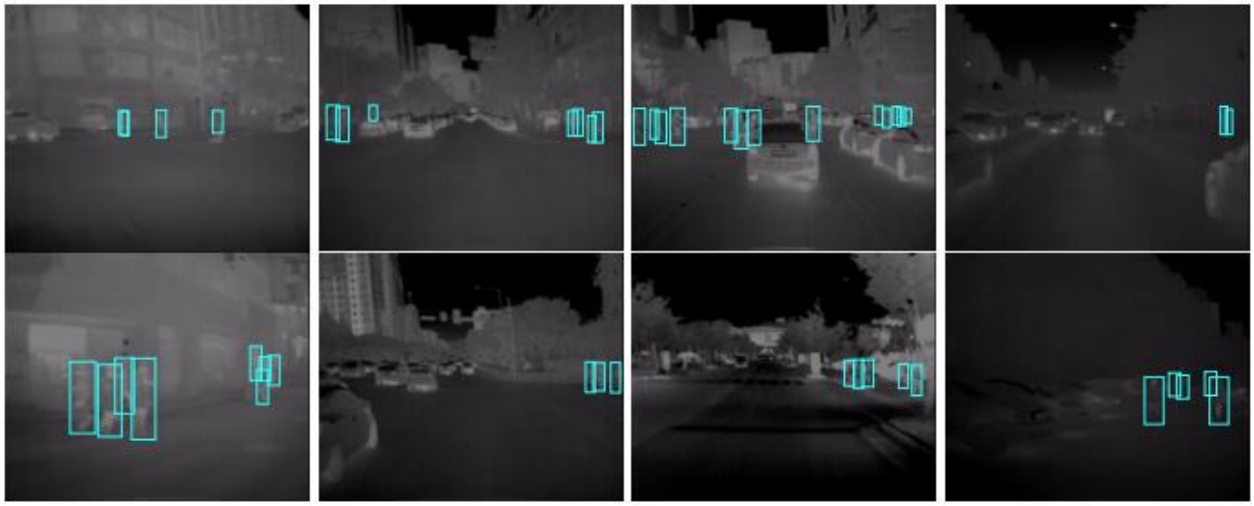

Fig. 3. Qualitative results of proposed method on KAIST dataset.

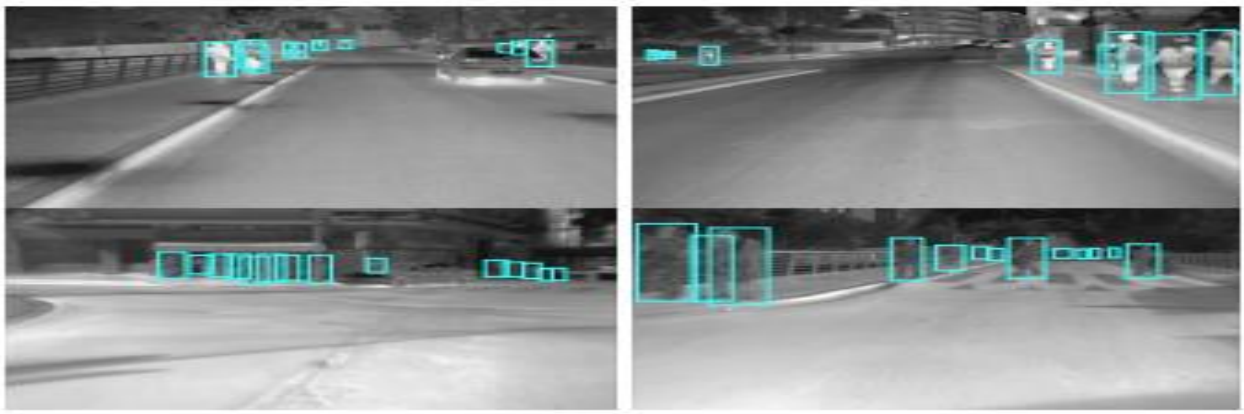

Fig. 4. Qualitative results on the Caltech Dataset

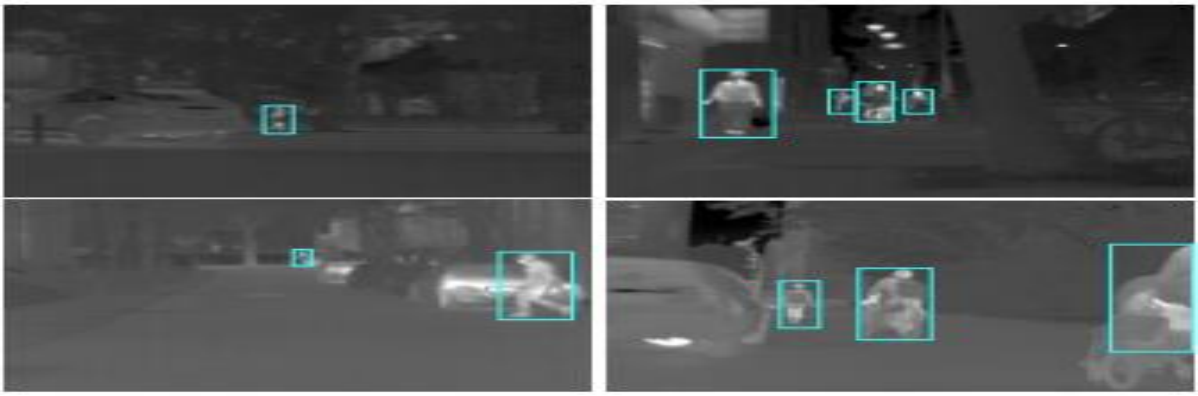

Fig. 5. Qualitative results on the City persons

On the KAIST, City persons, and Caltech datasets, respectively, the qualitative findings of the proposed model are shown in Figures 3, 4, and 5. In https, more qualitative results are provided: //youtu.be/FDJdSifuuCs. Figures 4 and 5's top and bottom rows depict daytime and nighttime scenes, respectively. Even in challenging circumstances, the proposed method accurately detects pedestrians. One such difficulty is small pedestrians, and Figure 5 demonstrates accurate detection of a small child and a far pedestrian. A crowded group of pedestrians with obstructions is another challenging scenario, and Figure 4 shows good detections. The SCoFA module helps learn context so that it can provide accurate detection in these situations. However, in groups of pedestrians, we occasionally observe a few missed detections. A group of pedestrians can sometimes be identified as a single pedestrian. Finally, when there is sufficient visibility, partially visible or occluded pedestrians are easily detected, as shown in Figure 5's bottom right image.

A. Proposed Analysis

Table 1. Proposed analysis based on various multimodal pedestrian image dataset

Dataset	Average precision	Log average miss rate	Accuracy	F1Score	Equal error rate
KAIST	91	77	55	45	52
Citypersons	93	79	59	49	55
Caltech	95	81	61	51	59

The above table-1 shows proposed analysis based on various multimodal pedestrian image dataset. Here the dataset analyzed are KAIST, City persons, Caltech in terms of average precision, log-average miss rate, accuracy, F1 score, equal error rate.

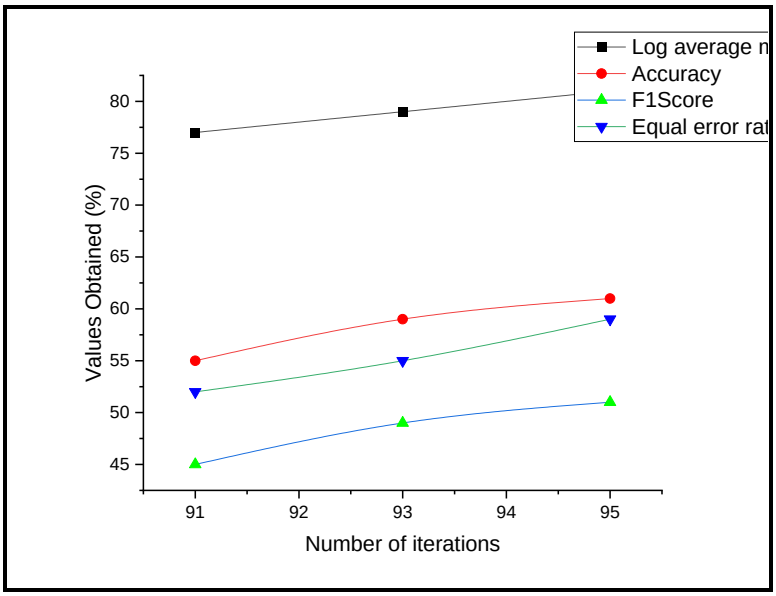

Fig. 6. Proposed analysis for various multimodal pedestrian dataset

From above figure 6 the proposed analysis based on various multimodal pedestrian image dataset is shown. The proposed technique attained average precision of 91%, log-average miss rate of 77%, accuracy of 55%, F1 score of 45%, equal error rate of 52% for KAIST dataset; proposed technique attained average precision of 93%, log-average miss rate of 79%, accuracy of 59%, F1 score of 49%, equal error rate of 55% for City persons dataset; proposed technique attained average precision of 95%, log-average miss rate of 81%, accuracy of 61%, F1 score of 51%, equal error rate of 59% for Caltech dataset.

B. Comparative Analysis

Table 2. Analysis based on various pedestrian multimodal dataset

Dataset	Techniques	Average precision	Log average miss rate	Accuracy	F1Score	Equal error rate
KAIST	**ORCNN [9]**	88	74	50	41	45
	SDSRCNN [12]	89	76	52	43	48
	MIA_PD_HMLM	91	77	55	45	52
Citypersons	**ORCNN [9]**	89	75	53	42	48
	SDSRCNN [12]	92	79	55	45	52
	MIA_PD_HMLM	93	79	59	49	55
Caltech	**ORCNN [9]**	91	77	55	44	51
	SDSRCNN [12]	93	80	58	49	55
	MIA_PD_HMLM	95	81	61	51	59

Table-2 shows comparative analysis based on various multimodal pedestrian image dataset. Here dataset analysed are KAIST, Citypersons, Caltech in terms of average precision, log-average miss rate, accuracy, F1 score, equal error rate.

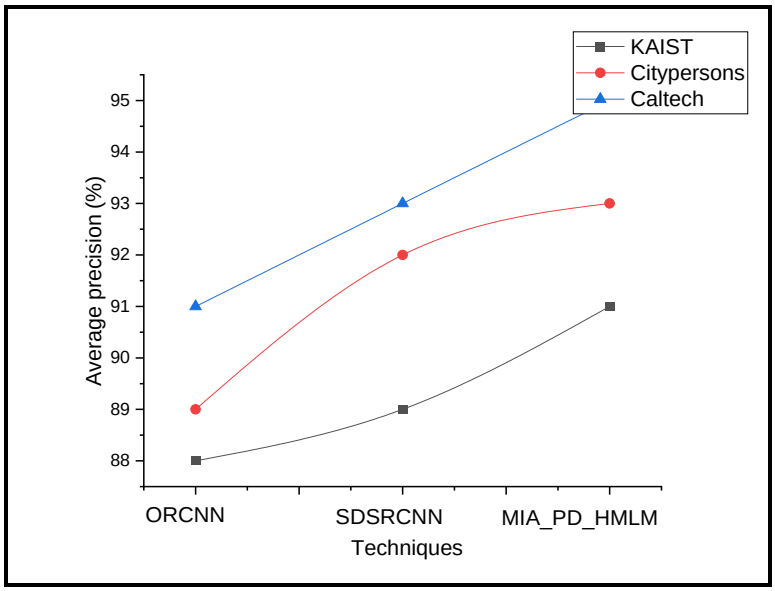

Fig. 7. Comparison of average precision

Figure 7 shows comparison of Average precision. for KAIST dataset proposed technique attained Average precision of 91%, existing ORCNN attained Average precision of 88% and SDSRCNN attained Efficiency rate of 89%; proposed technique Average precision of 93%, existing ORCNN Average precision of 89% and SDSRCNN Average precision of 92% for City persons Pedestrian dataset, for Caltech Square dataset proposed technique Average precision of 95%, existing ORCNN Average precision of 91% and SDSRCNN Average precision of 93%.

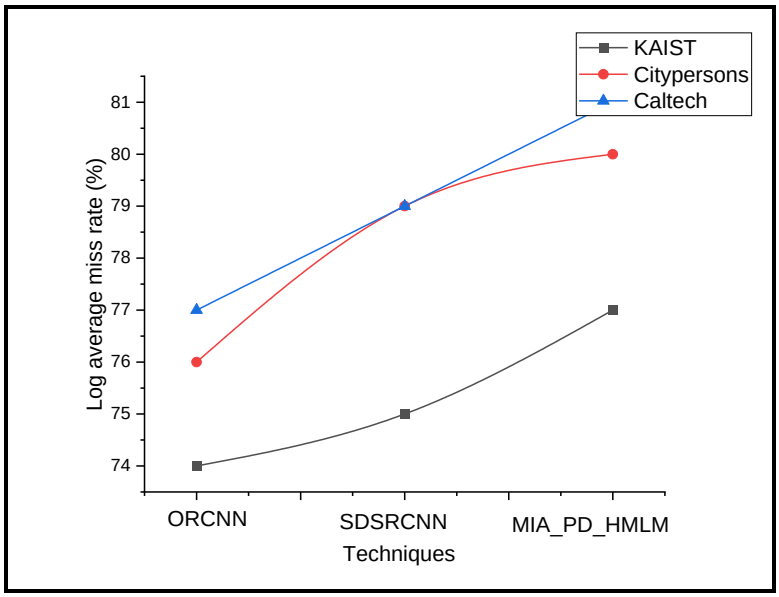

Fig. 8. Comparison of Log average miss rate

From above figure 8 comparison of Log average miss rate various pedestrian datasets. For KAIST dataset the proposed technique attained Log average miss rate of 77%, existing ORCNN attained Log average miss rate of 74% and SDSRCNN attained Efficiency rate of 76%; the proposed technique attained Log average miss rate of 79%, existing ORCNN attained Log average miss rate of 79% and SDSRCNN attained Log average miss rate of 75%; for City persons Pedestrian dataset proposed technique Log Average precision of 81%, existing ORCNN attained Log average miss rate of 77% and SDSRCNN attained Log average miss rate of 80%.

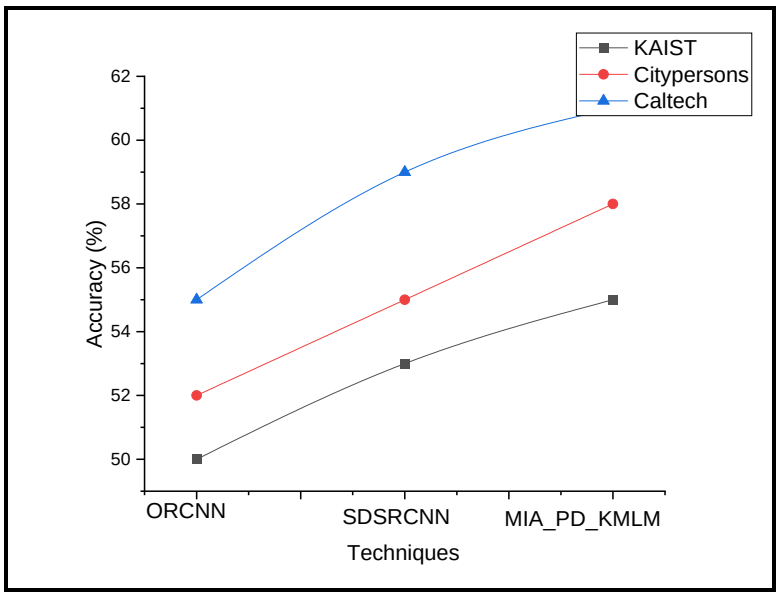

Fig. 9. Comparison of Accuracy

The above figure 9 shows comparison of Accuracy for KAIST dataset proposed technique Accuracy of 55%, existing ORCNN attained Accuracy of 50% and SDSRCNN attained Efficiency rate of 52%; proposed technique Accuracy of 59%, existing ORCNN Average precision of 53% and SDSRCNN Accuracy of 55% for City persons Pedestrian dataset, for Caltech Square dataset proposed technique Accuracy of 61%, existing ORCNN Accuracy of 55% and SDSRCNN Average precision of 58%.

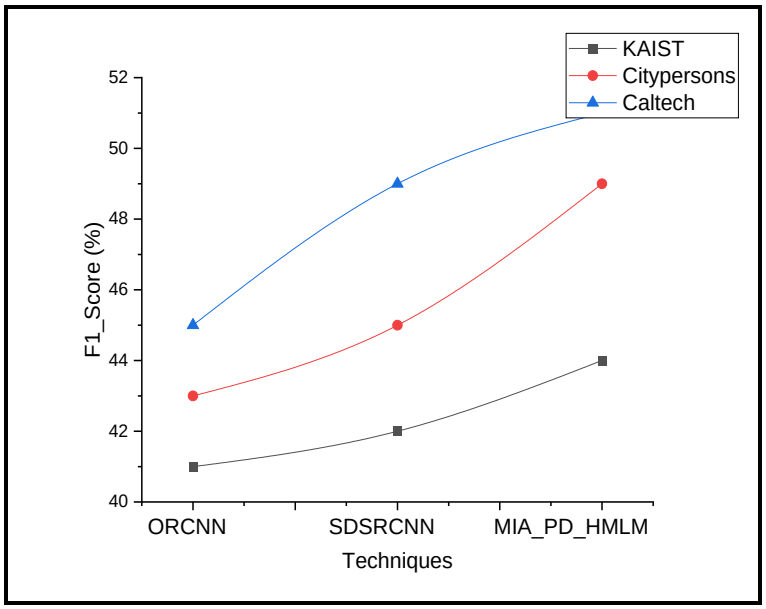

Fig. 10. Comparison of F1Score

From above figure 10 comparison of F1Score various pedestrian datasets. For KAIST dataset proposed technique F1Score of 45%, existing ORCNNF1Score of 41% and SDSRCNNF1Score of 41%; proposed technique F1Score of 49%, existing ORCNNF1Score of 42% and SDSRCNNF1Score of 45%; for City persons Pedestrian dataset proposed technique F1Score of 51%, existing ORCNNF1Score of 44% and SDSRCNNF1Score of 49%.

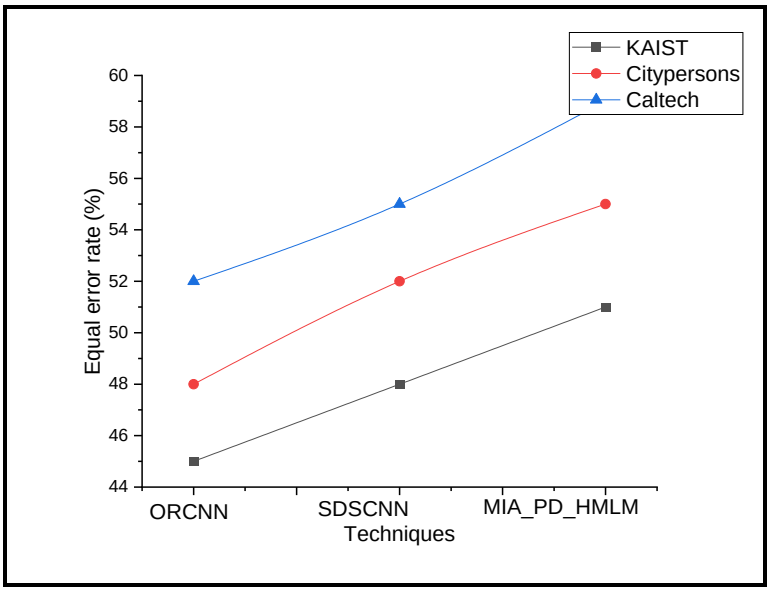

Fig. 11. Comparison of Equal error rate

Figure 11 shows comparison of Equal error rate. For KAIST dataset proposed technique attained Equal error rate of 52%, existing ORCNN attained Equal error rate of 45% and SDSRCNN attained Efficiency rate of 48%; the proposed technique attained Equal error rate of 55%, existing ORCNN attained Equal error rate of 48% and SDSRCNN attained Equal error rate of 52%; for City persons Pedestrian dataset, proposed technique attained Equal error rate of 59%, existing ORCNN attained Equal error rate of 51% and SDSRCNN attained Equal error rate of 55%.

We manually extracted a number of windows that were resized to 36 by 18 pixels and contained either pedestrians or parts of the background. The two final datasets consisted of 7072 windows for the first dataset, which included 3536 pedestrians and 3536 non-pedestrians, and 5236 windows for the second dataset, which included 2618 pedestrians and 2618 non-pedestrians. In order to encourage additional research on this subject, the two datasets will be made available to the general public. We extracted twelve distinct feature maps, each of which was the same size as the original images. We separated the two datasets into training and testing categories as a preliminary test. Sadly, the almost perfect results of the thermal-only based classifier prevented the fusion method's improvement from being statistically significant. The fact that the background is constant across all datasets suggests that this could be the case. Even though there was no

evidence of overlap between the training and testing windows, the training and testing sets had very similar background textures, which led to over-fitting and unreliable results. We decided to train on the first dataset and test on the second, and vice versa, because of this. In this manner, the background examples used in training and testing are sufficiently distinct from one another to guarantee that the performance achieved is representative of a typical circumstance in which the background is unknown a priori. The two background scenes share the same typology—urban outdoor environment—and are sufficiently similar to prevent results from being skewed.

5. Conclusion

This research proposes novel technique in multimodal image based pedestrian modal using machine learning techniques. The multimodal image has been classified using naïve spatio kernelized extreme convolutional transfer learning and optimized using metaheuristic salp cross modality swarm optimization. The performance indicators that are gathered by the optimization model are used to select the plans for available pedestrian facilities to meet time-varying demands of passengers. With aim of lowering employment costs and preventing excessive fatigue, optimization method chooses best pedestrian facility plans as well as generates corresponding staff assignment plan while taking into account workload equity and downtime. In addition, we introduce modal separation learning to accelerate training and improve performance further. Extensive testing demonstrates that, in comparison to current methods, our approach bridges the gap between simulated and real environments. Proposed technique attained average precision of 95%, log-average miss rate of 81%, accuracy of 61%, F1 score of 51%, equal error rate of 59%. Better results have been reached for pedestrian detection with the use of DL methods. However, the identification of tiny, intermediate, and obstructed objects is still a problem for the current methods. The aforementioned problems can be taken into account or addressed in the future. Additionally, there hasn't been enough research done on improving detection performance in poor lighting and weather. This issue can be resolved in the future by training both models using daytime and nighttime models as a single paradigm, which will improve their generalization capabilities.

References

[1] Jain, D. K., Zhao, X., González-Almagro, G., Gan, C., &Kotecha, K. (2023). Multimodal pedestrian detection using metaheuristics with deep convolutional neural network in crowded scenes. *Information Fusion*, *95*, 401-414.

[2] Kolluri, J., & Das, R. (2023). Intelligent multimodal pedestrian detection using hybrid metaheuristic optimization with deep learning model. *Image and Vision Computing*, 104628.

[3] Kim, J., Nirjhar, E. H., Kim, J., Chaspari, T., Ham, Y., Winslow, J. F., ...&Ahn, C. R. (2022). Capturing environmental distress of pedestrians using multimodal data: the interplay of biosignals and image-based data. *Journal of Computing in Civil Engineering*, *36*(2), 04021039.

[4] Zhang, H., He, B., Lu, G., & Zhu, Y. (2022). A simulation and machine learning based optimization method for integrated pedestrian facilities planning and staff assignment problem in the multi-mode rail transit transfer station. *Simulation Modelling Practice and Theory*, *115*, 102449.

[5] Dradrach, A., Konert, J., &Ruminski, J. (2023, April). Multimodal camera for pedestrian detection with deep learning models. In *2023 IEEE International Conference on Industrial Technology (ICIT)* (pp. 1-6). IEEE.

[6] Stanitsa, A., Hallett, S. H., & Jude, S. (2023). Investigating pedestrian behaviour in urban environments: a Wi-Fi tracking and machine learning approach. *Multimodal Transportation*, *2*(1), 100049.

[7] Wang, B., Zou, Y., Zhang, L., Li, Y., Chen, Q., &Zuo, C. (2022). Multimodal super-resolution reconstruction of infrared and visible images via deep learning. *Optics and Lasers in Engineering*, *156*, 107078.

[8] Huang, Z., Sun, S., Zhao, J., & Mao, L. (2023). Multi-modal policy fusion for end-to-end autonomous driving. *Information Fusion*, 101834.

[9] Hua, C., Sun, M., Zhu, Y., Jiang, Y., Yu, J., & Chen, Y. (2022). Pedestrian detection network with multi-modal cross-guided learning. *Digital Signal Processing*, 103370.

[10] Qiu, S., Zhao, H., Jiang, N., Wang, Z., Liu, L., An, Y., ...&Fortino, G. (2022). Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges. *Information Fusion*, *80*, 241-265.

[11] Ali, S., Li, J., Pei, Y., Khurram, R., Rehman, K. U., &Mahmood, T. (2022). A comprehensive survey on brain tumor diagnosis using deep learning and emerging hybrid techniques with multi-modal MR image. *Archives of Computational Methods in Engineering*, *29*(7), 4871-4896.

[12] Chen, C., Nishio, T., Bennis, M., & Park, J. (2022). RF-Inpainter: Multimodal Image Inpainting Based on Vision and Radio Signals. *IEEE Access*, *10*, 110689-110700.

[13] Zhang, C., & Berger, C. (2023). Pedestrian Behavior Prediction Using Deep Learning Methods for Urban Scenarios: A Review. *IEEE Transactions on Intelligent Transportation Systems*.

[14] Huang, Z., Mo, X., &Lv, C. (2022, October). ReCoAt: A Deep Learning-based Framework for Multi-Modal Motion Prediction in Autonomous Driving Application. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)* (pp. 988-993). IEEE.

[15] Wang, K., Zhou, T., Zhang, Z., Chen, T., & Chen, J. (2023). PVF-DectNet: Multi-modal 3D detection network based on Perspective-Voxel fusion. *Engineering Applications of Artificial Intelligence*, *120*, 105951.

[16] Wanchaitanawong, N., Tanaka, M., Shibata, T., &Okutomi, M. (2023). Multi-modal pedestrian detection with misalignment based on modal-wise regression and multi-modal IoU. *Journal of Electronic Imaging*, *32*(1), 013025-013025.

[17] Thakur, N., Nagrath, P., Jain, R., Saini, D., Sharma, N., &Hemanth, D. J. (2023). Autonomous pedestrian detection for crowd surveillance using deep learning framework. *Soft Computing*, *27*(14), 9383-9399.
[18] Yazdani, M., Sarvi, M., Bagloee, S. A., Nassir, N., Price, J., &Parineh, H. (2023). Intelligent vehicle pedestrian light (IVPL): A deep reinforcement learning approach for traffic signal control. *Transportation research part C: emerging technologies*, *149*, 103991.
[19] Srinivas, K., Singh, L., Chavva, S. R., Dappuri, B., Chandrasekaran, S., &Qamar, S. (2022). Multi-modal cyber security based object detection by classification using deep learning and background suppression techniques. *Computers and Electrical Engineering*, *103*, 108333.
[20] Iftikhar, S., Asim, M., Zhang, Z., & El-Latif, A. A. A. (2022). Advance generalization technique through 3D CNN to overcome the false positives pedestrian in autonomous vehicles. *Telecommunication Systems*, *80*(4), 545-557.
[21] A. K. M. Fahim Rahman, Mostofa Rakib Raihan, S.M. Mohidul Islam, " Pedestrian Detection in Thermal Images Using Deep Saliency Map and Instance Segmentation", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.13, No.1, pp. 40-49, 2021. DOI:10.5815/ijigsp.2021.01.04
[22] Anupam Dey, Fahad Mohammad, Saleque Ahmed, Raiyan Sharif, A.F.M. Saifuddin Saif,"Anomaly Detection in Crowded Scene by Pedestrians Behaviour Extraction using Long Short Term Method: A Comprehensive Study", International Journal of Education and Management Engineering(IJEME), Vol.9, No.1, pp.51-63, 2019. DOI: 10.5815/ijeme.2019.01.05

Authors' Profiles

Johnson Kolluri received the B.tech degree in Information Technology from KITS-Warangal, Telangana, India, in 2008, the M.Tech degree in Software Engineering from VNR VJIET , Bachupally-Hyd, Telangana, India, in 2011. He is currently a Research Scholar with the NIT Mizoram of Computer Science & Engineering department. His research interests include Artificial Intelligence, Pattern Recognition, Computer Vision and Image processing.

Ranjita Das has received the B. tech degree in computer science Engineering from NIT Agartala, India, in 2008, the M.Tech degree in information technology from Tezpur University, India, in 2011, and Ph. D. degree in computer science & Engineering from NIT Mizoram, India, in 2018. She is currently an Assistant Professor with the NIT Agartala of Computer Science & Engineering department. Her research interests include Artificial Intelligence, Evolutionary Computation, Pattern Recognition, NLP , Computational Biology, Computer Vision and Image processing, Image Captioning.

How to cite this paper: Johnson Kolluri, Ranjita Das, "Multimodal Image Analysis Based Pedestrian Detection Using Optimization with Classification by Hybrid Machine Learning Model", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.17, No.1, pp. 31-44, 2025. DOI:10.5815/ijigsp.2025.01.03