Hindawi
Advances in Multimedia
Volume 2020, Article ID 6153580, 15 pages
https://doi.org/10.1155/2020/6153580

Research Article

Pyramidal Part-Based Model for Partial Occlusion Handling in Pedestrian Classification

M. Thu and N. Suvonvorn

Department of Computer Engineering, Faculty of Engineering, Prince of Songkla University, Hat Yai, Songkhla 90112, Thailand

Correspondence should be addressed to N. Suvonvorn; nikom.suvonvorn@gmail.com

Received 25 September 2019; Revised 12 December 2019; Accepted 14 January 2020; Published 24 February 2020

Academic Editor: Martin Reisslein

Pedestrian detection and classification are of increased interest in the intelligent transportation system (ITS), and among the challenging issues, we can find limitations of tiny and occluded appearances, large variation of human pose, cluttered background, and complex environment. In fact, a partial occlusion handling is important in the case of detecting pedestrians, in order to avoid accidents between pedestrians and vehicles, since it is difficult to detect when pedestrians are suddenly crossing the road. To solve the partial occlusion problem, a pyramidal part-based model (PPM) is proposed to obtain a more accurate prediction based on the majority vote of the confidence score of the visible parts by cascading the pyramidal structure. The experimental results on the proposed scheme achieved 96.25% accuracy on the INRIA dataset and 81% accuracy on the PSU (Prince of Songkla University) dataset. Our proposed model can be applied in the real-world environment to classify the occluded part of pedestrians with the various information of part representation at each pyramid layer.

1. Introduction

Each year, there are about 1.3 million road-related fatalities worldwide, and crash rates are escalating in most urbanized countries [1–3]. Since the turn of the century, many researchers have focused on employing the wide range of applications such as intelligent transport systems, driver assistance system, intelligent video surveillance system, and automotive safety system with an aim to mitigate and alleviate road crashes. Computer vision could increasingly resemble human vision and emulate sensed images and videos through applications of a wide range of machine learning technologies that can rectify erroneous human vision and create a safer environment in the daily drive [4]. Of all types of road users, pedestrians are the most vulnerable. To protect them and to reduce other potential risks, pedestrian detection and classification systems are widely employed. Many pedestrian classifiers: holistic classifier [5–11], part-based classifier [12–20], and deep model classifier [19, 21, 22], have already been proposed and are actually in use, but several challenges remain to be solved, such as illumination changes, occlusion, variation of pose and shapes, variation of appearances, and inconsistency of surroundings. Besides, occlusion handling under complex backgrounds in the real-world environment may involve further difficulties [8, 14, 19–21, 23].

In the state-of-the-art approaches, most researchers have been working on two issues: feature extraction methods (handcrafted features and deep convolutional features) and classification through the machine learning algorithms. Some of the promising handcrafted feature extraction methods are the histogram of oriented gradients (HOG) [4–6, 14, 24, 25], Haar wavelet [2, 25–27], scale-invariant feature transform (SIFT) [28–30], edge templates [5, 23, 31, 32], adaptive contour features [2, 23, 33, 34], Gabor filters [15, 27], covariance descriptors [11, 15, 19, 35], and local binary pattern (LBP) [6, 11, 24, 36]. Among these handcrafted features, the histogram of oriented gradients (HOG) is a well-known feature descriptor for pedestrian detection due to the rich feature information under different illumination changes [14, 16, 20]. However, the traditional HOG detects well when the whole pedestrian's body appears in the system and performs poorly under occlusion and cluttered images [6, 8, 14, 20, 22]. Therefore, the enhanced

HOG detectors [15, 37] and complemented HOG with other methods [4, 25, 32, 38] had been proposed by modifying the HOG to solve the inviable occlusion of pedestrians.

In the classification method, the well-known techniques such as support vector machine (linear and latent) [4, 5, 10, 12, 25, 38], AdaBoost [7, 9, 34, 39], neural network [2, 3], random forest [3], and cascade [7, 17, 27] had been widely used. Among them, the linear SVM is one of the useful techniques that can compute faster than a latent SVM in terms of performance and efficiency. However, the classifier is leading to degraded performance because of misclassification of the partial occluded parts that are the assumed noise or backgrounds. The visibility of occluded parts (see Figure 1) can vary in the real-world environment, and the hypothesis of the prediction is dependent on the accurate estimation of the classifier. In order to study the linear SVM classifier, the confidence score values are used to examine whether a portion of a pedestrian is occluded. Taking the advantages of the linear SVM classifier, we observed that the classification score of different parts of the image could lead to an accurate prediction under the partial occlusion.

In the pedestrian classification model, there are two approaches: the full body (holistic) approach that relies on the whole body of pedestrians and the part-based approach that combines a set of specific parts of the human structure. In holistic approaches, some state-of-the-art pedestrian detectors: cascade learning model [6, 8, 17], hierarchical cooccurrence model [7, 11, 25], and decision tree model [2, 3, 28], were based on the block-based representation in which the feature extracted from each block uniformly responds to the classifier whether the occlusion occurs in that area. These models achieved excellent performance results on the pedestrian classification under the occlusion. Most of these approaches additionally used an occlusion map when the partial occlusion occurs and attained a more robust classifier to handle the partial occlusion problem.

On the contrary, part-based approaches had been introduced to detect occlusion parts of the human body, which were based on the part-based representation, and deformation of humans [12–20]. Some of the well-known part-based models are discriminatively trained part-based model [12, 15], deformable part-based model [14, 17, 19, 20], grammar model [13], mixture mask model [16], and part-boosting model [18, 19]. In these models, the images were segmented into the multicomponent templates with a low resolution for the root template and a high resolution for the part. Generally, the specific part of the human body (head, right/left shoulder, torso, and right/left leg) is defined for training to learn the features of each body part with the use of the latent SVM values for handling the occlusion problems. In the process of the training stage, the detection score map of the parts is obtained with the learned part filters and the deformable part score is calculated from the subtraction of the penalty of deformation and part detection score maps [14]. These part-based approaches attained a lot of attention on the excellent work to handle the occlusion parts in the detection window and achieved promising performance on the pedestrian detection. However, the part-based approaches require high computation costs to sum up the score values of the deformation parts, and the part score of the detector could be very low if one part is occluded [2, 3, 6]. As far as studying through the previous well-known approaches, the block-based representation methods add the additional information to know the occluded part in the detection window and require high computation time for each block with the additional information. On the contrary, the part-based representation needs a lot of computation time with the deformation of the selective structure of human parts to learn the information of each part. However, these well-known state-of-the-art models got the best choice as well as the excellent performance on pedestrian detection to handle the occlusion challenges.

In the recent years, the advanced development of deep convolution features has gained significant attention in terms of automatic end-to-end learning with the effective rich representations to enhance the ability of handcrafted features. Compared with the handcrafted features, the deep model and deep convolutional network have reached successful state-of-the-art performances in terms of image classification in computer vision due to the hierarchical representation of high-level features. The existing deep convolutional neural network models for image classification on a large-scale dataset, ImageNet, achieved excellent classification performance (e.g., VGGNet, GoogLeNet, and ResNet) [40–44]. Recently, the use of the pretrained model to get the boost on accuracy with small datasets got much attention [40–42, 45]. Among them, GoogLeNet (Inception v3) [40], with the ImageNet-weighted parameter values, is basically used as a benchmark in the image classification to retrain the pedestrian dataset with two class annotations. However, the performance of the deep convolutional network can impact the results when the input data contain noise [41, 42] and occlusion scenarios [22, 41, 42].

To deal with the above challenges, there are two issues considering the partial occlusion handling: how to estimate the location of the partially occluded parts and the accurate performance of the partially occluded parts in the detection window. Due to the useful idea of the local classifier, the hypothesis of the confidence score values is adapted as an occlusion inference to estimate the location of the partially occluded parts in the detection window. Furthermore, the conducted idea of the cascading concept is used to calculate the current position of the response score by comparing it with the corresponding threshold values at each layer of pyramids for the subsequent calculation. From this observation, this paper proposes the pyramidal part-based model (PPM) by adopting the advantages of the holistic classifier and part-based classifier to handle partial occlusion in the detection window. Firstly, a part classifier is used to estimate the partially occluded parts in the given image. The hypothesis of the confidence score from the classifier is assumed as the occlusion inference whether or not the partial occlusion arises in the detection window. Secondly, if the partial occlusion happens in the given input image, the pyramidal ensemble classifier (PEC) is used to predict the hypothesis of the individual confidence scores at each specified pyramid layer. Additionally, the input image is

Figure 1: Example representation of the visible partial occluded parts. (a) Upper body part. (b) Occluded lower body part. (c) Left body part. (d) Occluded right body part.

segmented into the pyramidal part representation to calculate the confidence scores of each part, in order to identify the partially occluded parts. Finally, the ensemble score is composed if the individual scores are confident enough by comparing them with the corresponding threshold values. Moreover, the majority vote via the pyramid structure is employed for the threshold decision from the confidence scores of each part of the PEC. In this paper, we evaluated our proposed method and compared its performance with the baseline HOG + SVM model and the benchmark pretrained GoogLeNet (Inception v3) model, for handling the occlusion under complex scenes. Experiment results on two datasets demonstrated that the proposed model achieved the promising performance for nonoccluded pedestrians and for the partially occluded pedestrians, but not for the seriously occluded pedestrians.

2. Related Works

A relevant role of features in images will be applicable for the detection of different-scale pedestrians. This is due to the fact that the feature information of a large-scale image is different from a corresponding feature extracted from a small-scale image [28]. Concerning related studies, the histogram of oriented gradients (HOG) is a well-known feature descriptor for pedestrian detection to calculate the gradient orientation, but this technique can detect a pedestrian only when the whole body of the pedestrian appears in the system. Dalal and Triggs [5] proposed the normalized histogram of gradient orientation features in person detection to reduce the false positive rates. The performance of the system improved the various descriptor parameters, scale of gradients, orientation binning, relatively coarse spatial binning, and high-quality blocks [5]. Anyway, the current single feature has been proved difficult to handle the occluded parts of the pedestrians, and a part-based model would be effective for the detection results with a local spatial invariance. However, the detection rates markedly drop if the attached objects cover some parts of the person's body [6, 14].

To solve this problem, the Wang et al. [6] method performed well to handle occlusion by using the histogram of gradients (HOG) and local binary pattern (LBP) as a feature set. To know whether the occluded part occurs in the detection window, an occlusion likelihood map has been additionally introduced to handle the occlusion problem. Marín et al. [8] followed the Wang et al. [6] method to enhance the classifier with the random subspace for human detection, which is trained with a subset of images with and without partial occlusion for the partial occlusion handling. The proposed method outperformed the evaluation performance on both partially occluded and nonoccluded data for classification and detection. Li et al. [7] modified the adaptive boosting algorithm and cascade detector by combining them with the histogram of oriented gradient (HOG) features for more accurate detection of the pedestrians. In order to identify the occluded part of a pedestrian, the local area-making map (LAMM), which is derived from the enhanced cascade scheme, was proposed to decode the occluded part in the detecting window.

Occlusion and deformation can vary on different visible human body parts under a complex background. A well-known method, the part-based detectors, proposed the whole body part into multiple component parts of specific structure of the human body to handle the occlusion problem. In the past decade, many evaluations have been done to improve the occluded pedestrian detection. Felzenszwalb et al. [12] proposed the mixtures of multiscale deformable part models, which introduced the notion of multiple components (i.e., head, right shoulder, left shoulder, upper leg, and lower leg) into the detector. This system relied mainly on the discriminative training that used latent information for matching deformable models to images as well. Choi et al. [14] also proposed a deformable part-based model (DPM) on humans, which had recently emerged as a useful and popular tool to detect a part of an object. Luo et al. [14] proposed the mixture representation of different body parts (i.e., body, head shoulder, upper body, and lower body) to explicitly model the complex mixture of visual variations at multiple levels. The proposed model learns hierarchical features, salience maps, and mixture representation of body parts. Wahyono et al. [18] proposed the part-based models for the detection of humans carrying baggages, by modeling the body parts of humans (i.e., head, torso, leg, and baggage parts). A mixture model was also built specially for the baggage part due to the significant variation of its shape, colour, and texture. The proposed model learns the parallelogram Haar-like features, Gaussian mixture model, and body part-boosting model for the detection and classification of the baggage carried by humans.

In contrast, learning from each specific part of the human body's structure and combining it with different features to make use of the contextual information are commonly used among the occluded pedestrian handling.

Most of these approaches [6–8] attained the promising performance to handle the misclassification of the occluded pedestrians and estimated the location of these parts in the detection window. Our method is different from these approaches because we adapted the pyramidal part representation to make use of the confidence score from the PEC without using additional contextual information to handle the partially occluded parts.

3. Pyramidal Part-Based Model

The challenging part of this research is to handle the partial occlusion when some parts of the pedestrian are invisible, and the inaccurate scores of the part classifier will affect the performance prediction. Therefore, a pyramidal part-based model (PPM) is proposed to obtain a more accurate prediction based on the majority vote of the confidence scores of the visible parts by cascading the pyramid structure from top to down. There are two classification stages in the proposed model: a part classifier (PC) and a pyramidal ensemble classifier (PEC) (see Figure 2).

A part classifier is basically used as the occlusion inference to identify the partial occluded parts of the input images in the range [+1, −1]. If the inference result given by the part classifier is not confident enough, then an occlusion inference process is applied. Thereafter, the inference process determines that there is a partial occlusion; a PEC is applied for classification in order to obtain an accurate prediction for the detection window based on the pyramidal part representation. After that, the ensemble score is achieved in terms of the majority vote via the pyramid structure from the confidence scores of the PEC. The detailed implementation of the PPM process is described in the following section.

3.1. Part Classifier. In order to identify the partially occluded parts in the image, the part classifier is used to get the response over the whole window that is described by the feature vector. We followed the Wang et al. [6] procedure to know if there is a partially occluded part of the human in the given image. There are two necessary steps for the procedure of the part classifier: feature extraction and classification, which are needed to obtain a response. In addition, the histogram of oriented gradients (HOG) and support vector machine (SVM) are used for the feature extraction and classification, respectively. If the response of the part classifier is not confident enough, the pyramidal part representation is applied in the next step. In order to identify the partially occluded part in the detection window, the occlusion inference process is applied depending on the response value of the part classifier. The detailed process is explained in the following section.

3.1.1. HOG Feature Descriptor. In this section, the detailed process of the HOG feature extraction is described (see Figure 3).

The histogram of oriented gradients (HOG) has been proposed by Dalal and Triggs [5], which was successfully applied for pedestrian detection, object detection, facial expression recognition, and pose recognition. To represent any pattern of gradients, the features play a vital role because the edges are insensitive to illumination changes and pose variations. The orientation of the gradients within a region and the spatial layout are obtained by dividing the images into nonoverlapping blocks at multiple resolutions [29]. In order to evaluate the HOG descriptor, the image is divided into a number of blocks, and the directional gradients from each block of histograms are concatenated to obtain the shape descriptor. The image is divided into a dense grid of rectangle cells under the detection window, and each cell is grouped into blocks.

A separate orientation of gradients ($\Theta_{G\ (x,\ y)}$) and the directional gradients (Grad_x and Grad_y) for each cell are computed by the gradient mask with the pixel values of $I(x, y)$. The horizontal and vertical gradients are given in the following equations, respectively:

$$\mathrm{Grad}_x(x, y) = I((x+1), y) - I((x-1), y), \tag{1}$$

$$\mathrm{Grad}_y(x, y) = I(x, (y+1)) - I(x, (y-1)). \tag{2}$$

The magnitude of the gradient is computed by the summation of the square root of the gradients, and the orientation is determined by the proportion of the y-gradient and x-gradient that are shown in the following equations, respectively:

$$\mathrm{Grad}(x, y) = \sqrt{\left(\mathrm{Grad}_x(x, y)\right)^2 + \left(\mathrm{Grad}_y(x, y)^2\right)}, \tag{3}$$

$$\theta_{\mathrm{Grad}}(x, y) = \tan^{-1}\left(\frac{\mathrm{Grad}_y(x, y)}{\mathrm{Grad}_x(x, y)}\right). \tag{4}$$

The gradient vote of each pixel is accumulated to the histogram orientation which has 9 bins of 0–180 degrees, and each vector of the magnitude is split between neighboring bins depending on the angle. The voted values are bilinearly interpolated between two neighboring bin centers that are given in the following equation:

$$\mathrm{Grad}_i = ((1-\gamma) * \mathrm{Grad}(x, y)), \tag{5}$$

where γ means the weight of each pixel. Each block of histograms is normalized to reduce the variations of illumination and contrast background. Each cell contains a histogram of 9 bins, and each block contains a histogram of 72 numbers (9 bins × 8 cells). This normalized value, L2_norm, is used for each block histogram disposed in the following equation:

$$\mathrm{L2\_norm} = \frac{\varphi_i}{\sqrt{\left(\|\varphi_2\|^2 + \varphi_2\right)}}, \tag{6}$$

where i is a number from 1 to 72, φ_i is the nonnormalized vector of the block, and φ is a small constant that avoids division by zero and is approximately assumed as 1.

3.1.2. SVM Classifier. The classifier we used is the baseline SVM algorithm [5, 6] as a part classifier. All of the

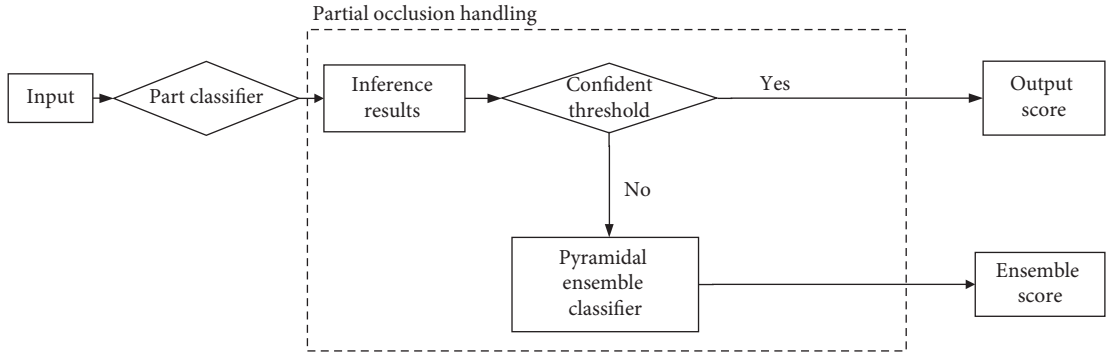

Figure 2: A proposed framework of the pyramid part-based model for occlusion handling.

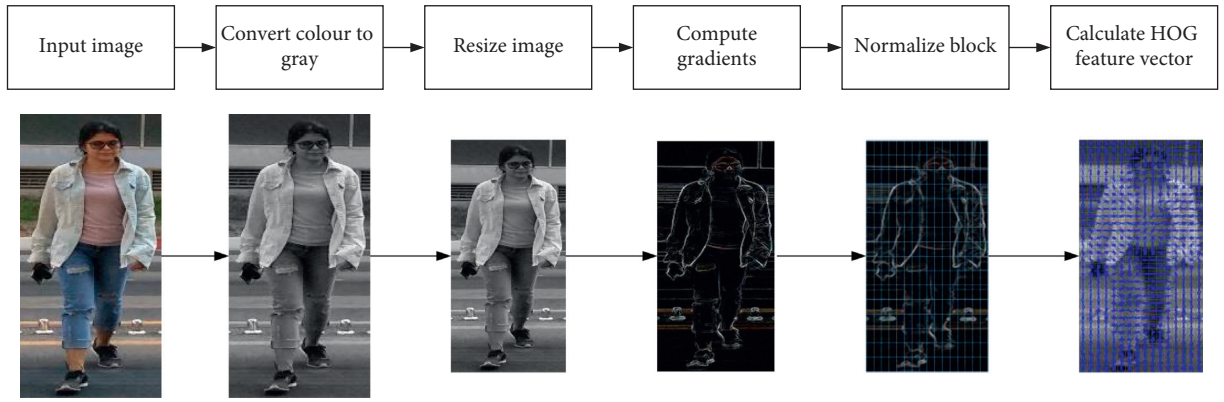

Figure 3: Process of feature extraction with the histogram of oriented gradients (HOG).

corresponding feature vectors extracted from the input image $I_1(x, y)$ are the feed for the SVM classifiers. The desired response R_1^1 of a part classifier is "1" for the pedestrian and "−1" for the nonpedestrian. The support vector classification (SVM::C_SVC) is used for allowing the imperfect separation of classes with the penalty multiplier C for outliers. A radial basis function (RBF) kernel is used instead of a linear kernel, applicable in this area, and the autoTrain function is used for adjusting the optimal parameter for classification.

3.1.3. Occlusion Inference. The procedure of the occlusion inference is used to determine whether the input image contains a partially occluded part or not in a detecting window, using the same procedure as Wang et al. [6]. We first consider the response value (R_0^0) of the input image (P_0^0) applied with a part classifier F that falls into the range $[-1, 1]$. If the response is preserved as a positive value, we observe that there is a pedestrian. If not, there may be a pedestrian with occluded parts. The algorithm is described as follows Algorithm 1.

3.2. Pyramidal Ensemble Classifier. The previous part classifier, using the traditional HOG and SVM methods, could achieve excellent performance on normal scenes, but the performance is reduced when partially occluded parts are contained in a detection window. To enhance the discriminative learning of a part classifier capability, a pyramidal ensemble classifier as a majority vote is proposed in order to learn additional structural information of confidence responses based on a pyramidal structure. There are two main steps: part representation and majority vote.

3.2.1. Part Representation. We model the whole image of the pedestrian as the pyramidal part representation with the pyramid layers to determine the accurate prediction if there may be partially occluded parts in the image. The occlusion may happen at different body parts and has various representations. Some research approaches combined assorted parts of the body to detect the human effectively under occlusion [21]. The aim of the pyramidal part representation of images is to decompose the original images into subimages with the same scales (see Figure 4).

We model the part representation without distinguishing different viewpoints of the image. A part representation model consists of n number of pyramidal layers, and i parts are assigned to each layer given by different sizes $\{P_i^n\}$. In this paper, we used three layers for the experiment to predict the confidence scores 1 for the pedestrian and −1 for the nonpedestrian. The confidence scores of each segmented part from the feature vector illustrate which portion of the image contains occluded parts. As an example of part representation, the original images $I(x, y)$ are resized as the images $P_0^0(I(x, y))$ into the specific window size for the base layer. The next step is

Input: image part $\mathbf{P}_0^0$.
Output: found partially occluded part
Procedure:
Calculate $\mathbf{F}(\mathbf{P}_i^n)$
Response $\mathbf{R}_i^n$: = $\mathbf{sign}(\mathbf{F}(\mathbf{P}_i^n))$
if $\mathbf{R}_i^n = 1$ then
  return true//Pedestrian
else
  return false//There is a partial occlusion in the image
end

Algorithm 1: Occlusion inference.

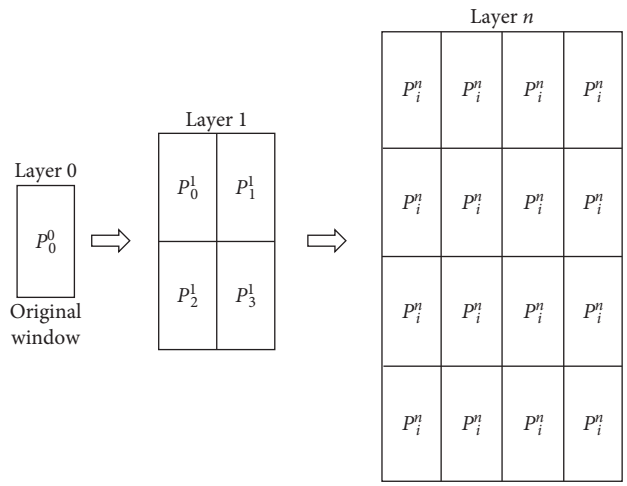

Figure 4: Pyramidal part representation. From left to right, the original input and that segmented into parts are shown.

to segment the original images into four parts: $\{P_0^1, P_1^1, P_2^1, P_3^1\}$, for the next layer. Later, there are sixteen parts to segment $I(x, y)$ in the same way as the previous layer. Likewise, the part representation process will follow the same procedure until the n^{th} pyramid layer.

3.2.2. Pyramidal-Based Majority Vote. The majority vote via the pyramidal structure is defined depending on the confidence score values obtained from the part classifier. The confidence is evaluated from the top to the bottom layers, hierarchically. An idea of the pyramidal ensemble classifier is proposed to accurately calculate the confidence score values via the pyramidal structure depending on the voting results from its upper layer parts (see Figure 5).

The algorithm of the pyramid ensemble classifier is described in Algorithm 2 and considers the part classifier set $\{P_0^0, \ldots, P_i^n, \ldots, P_{2^{2n}-1}^N\}$, where n is the number of pyramid layers and i is the number of parts. The normalized decision response N_i^n from the part classifier response R_i^n is given as a binary value $H_n \in \{1, 0\}$, defined as the following equation:

$$N_i^n = \begin{cases} 1, & \text{if } R_i^n \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Then, D is defined as the summation of normalized response representing the number of votes, which is given by

$$D = \sum_{i=0}^{2^{2n}-1} N_i^n. \tag{8}$$

Finally, the majority vote V, meaning sufficient confidence, is determined by a threshold value $(2^{2n})/2$, as described in the following equation:

$$V = \begin{cases} \text{pedestrian}, & \text{if } D \geq (2^{2n})/2, \\ \text{nonpedestrian}, & \text{otherwise.} \end{cases} \tag{9}$$

Example of results for the pyramidal ensemble classification are illustrated (see Figure 6) in comparison with classification response (see Table 1). In the illustration of the figure, the original image contains the pedestrian crossing a road in the near scale with a vehicle. The classification response of the 1st layer described a nonpedestrian (–1) in the image that was misclassified due to the overlapping with the vehicle parts. In the 2nd layer, the part classification responses in this image showed the same response in the 1st layer response. However, the 3rd layer classification response turned to describe as a pedestrian (1) for the input image, which contains the pedestrian. By analysing the classification response on this result, we found that some parts of the response were misclassified as the false positive values for the shadow of the human and the attached part of the vehicle. Due to this misclassification of noise, we can draw a conclusion that the 1st layer response was described as a nonpedestrian. However, this is not valid for all the testing images, especially some PSU data that contained complex backgrounds and various appearances (Figure 5).

4. Experimentation Results

In this section, the experimental setup, the conducted procedures of the proposed technique, and the evaluation results are described. The experimental setup was based on Intel® Core™ i7-8700K CPU @ 3.70 GHz, 32 GHz RAM, and Windows 10 environment using OpenCV 3.0.0 library and Keras with the TensorFlow backend. We performed our analysis on the publicly available INRIA person dataset [5] and PSU dataset [46].

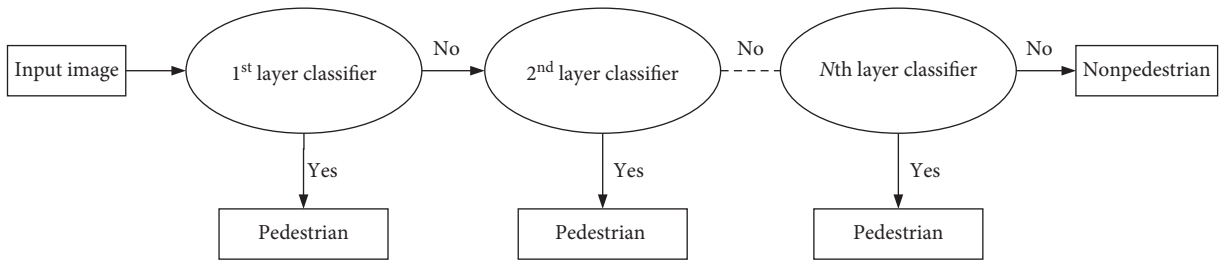

Figure 5: Flowchart of the pyramidal ensemble classifier (PEC).

Input: image part $\{\mathbf{P}_0^0, \ldots, \mathbf{P}_i^n, \ldots, \mathbf{P}_{2^{2n}-1}^N\}$
Output: pedestrian/nonpedestrian
Procedure:
For $n \in 0, \ldots, N$ do
  For $i \in 0, \ldots, 2^{2n} - 1$ do
    Calculate $\mathbf{F}(\mathbf{P}_i^n)$
    Response $\mathbf{R}_i^n := \mathbf{sign}(\mathbf{F}(\mathbf{P}_i^n))$
  End
  $\mathbf{N}_i^n = \begin{cases} \mathbf{1}, & \text{if } \mathbf{R}_i^n \geq \mathbf{0} \\ \mathbf{0}, & \textbf{otherwise} \end{cases}$
  $\mathbf{D} = \sum_{i=0}^{2^{2n}-1} \mathbf{N}_i^n$
  If $\mathbf{D} \geq (2^{2n})/2$ then
    Break; //Pedestrian
  Else
    Continue; //Nonpedestrian
  End
End

Algorithm 2: Pyramidal ensemble classifier.

For training the part classifier, we followed the same procedure as Dalal and Triggs [5] in which the part classifier is trained by feeding the SVM with the random number of positive and negative samples per image. Furthermore, the training data are divided into four subsets: reasonable, back/front upright, occlusion without complex scenes, and occlusion with complex scenes for estimating the various scenes' performance on the HOG + SVM classifier and the PEC. We performed the experiment with three layers of the pyramid model because when the depth of the pyramid increased, the number of parts segmented from this layer also increased and there may be less amount of the human structure component contained. The window size for the HOG feature descriptor is 128×256 for the 1st layer, 64×128 for the 2nd layer, and 32×64 for the 3rd layer (see Section 3.2), which consists of 16×16 block size with 8×8 cell size of 9 orientation bins for the INRIA data and PSU data. The HOG feature vector is normalized with the L2_HYS norm. We applied the same parameter values to extract features for each part of the pyramid fusion model. The following section presents the details of the experimentations.

4.1. Datasets. A comparison characteristic of the different datasets is described (see Table 2) with the example of sampled data in INRIA and PSU datasets (see Figure 7). A well-known standard dataset, INRIA person dataset [5], in which some of the pedestrians are roughly occluded, is used to assess the classifier without occlusions. To evaluate the performance of the classifier under partial occlusion, the PSU dataset [46] is used in which a significant number of partially occluded pedestrians are annotated.

4.1.1. INRIA Person Dataset. This dataset contains high-resolution static images of pedestrians with different poses and from different views and is proposed in [5], and it is still one of the most widely used datasets in human detection. The data are already divided into training and testing subsets. The annotations are provided for the original positive images (those containing pedestrians). The images come from a personal digital image collection, and pedestrians are shown with different poses against a variety of backgrounds (indoors, urban, and rural) in which people are normally standing or walking. Examples and counterexamples in the training set are downscaled to a height of 96 pixels (a margin of 16 pixels is normalized to 64×128 pixels, in which pedestrians are added around them).

4.1.2. PSU Dataset. This dataset also contains static images with occlusion parts and with different poses and complex backgrounds Pedestrian data are captured from multiview positions according to various postures: upright, walking, standing, cycling, motorbike riding, left, right, back, and occluded parts, of humans at different viewpoints around the marketplace areas, or on the PSU campus, and some data are taken from Google. There are four types of the image resolution: 64×64, 256×256, 720×960, and 960×720 pixels, which are resized to the original image resolution of 3120×3120 pixels for training and testing data to get the effective evaluation of the detection algorithm. Each pedestrian is provided with the useful properties: scene, view, pose, occlusion, and attachment information, and the scenes of the images contain pedestrians crossing a road, being in marketplace areas or at junction points of the city center, or on campus roads, to represent the real-world environment.

There are four groups of validation subsets from INRIA and PSU datasets (see Table 3). This is because the INRIA dataset contains mostly upright persons, but not many partially and heavily occluded parts are available. However, the PSU dataset contains partially occluded parts and heavily occluded parts with complex scenes. Therefore, four groups of validation sets are created for the experiment: Vad 1, Vad 2, Vad 3, and Vad 4, to evaluate the classification. Furthermore, we trained and tested with the PSU data

Table 1: Examples of results for the pyramidal ensemble classifier.

Pyramid layer	Part response	Layer response	Pyramid response
1st layer	−1	−1	
2nd layer	−1 −1 −1 −1	−1	1
3rd layer	1 1 −1 1 −1 1 1 1 −1 −1 1 −1 1 1 −1 1	1	

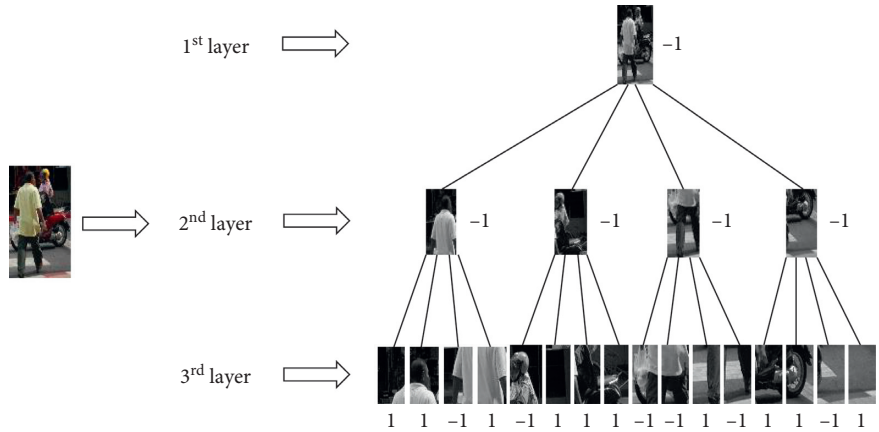

Figure 6: Illustration of the part representation process using the pyramidal ensemble classifier.

samples for Vad 1 and INRIA data samples for Vad 2. To get a validated result, we crossly trained and tested on both datasets for Vad 3, having PSU as training samples and INRIA as testing samples. For Vad 4, we trained with INRIA data samples and tested with PSU data samples.

4.2. Results and Discussion. This section describes the experimentation results, as well as the discussion of the proposed technique. Different experiment results are carried out, whereby various subsets of the validation are set up to enhance the capabilities of the classification.

4.2.1. Experiment Results on Pyramid Layers. The classification performance of the PEC according to different layers is shown in Figure 8. For the 1st layer, the whole body of the pedestrian is considered one part in the classification process and only 71% is obtained according to occluded parts. In the 2nd layer, the whole body is segmented into four parts, the prediction score of each part is individually estimated, and the classification accuracy achieved 6% improvement. This is because the misclassification of the occluded parts is assumed as noise in the 1st layer. For the 3rd layer, more subparts are used for the classification that leads to a better identification of the occluded parts. The performance of the 3rd layer reached 9% improvement compared with the 1st layer's accuracy. With the structural pyramidal representation, the classification performance improved, while the number of layers increased; however, the complexity has also increased. In our experiments, we choose the 3rd layer as the optimal pyramidal structure.

4.2.2. Experiment Results on INRIA and PSU Datasets. In our proposed framework, we have developed an idea of part-based model and ensemble learning with the use of the pyramidal part representation to handle the occlusion problem in the real-world environment. Four groups of training subsets: reasonable, front/back, occlusion without complex scenes, and occlusion with complex scenes, are used to evaluate the performance of our proposed classifier and to estimate the various scenes' performance. The reasonable subset contained only full body pedestrians with no occlusion under complex scenes. The subset of front/back contained only front and back views of the pedestrians' body because the previous classifier attained declined results when testing with the backside of the pedestrian. The other two subsets contained partially occluded parts of pedestrians without complex scenes and with complex scenes.

In this experiment, we choose HOG + SVM as the baseline classifier because of it being widely used in pedestrian detection and classification. However, the performance was limited under occluded parts and various scenes of the pedestrian's posture. We conducted the baseline classifier with the use of pyramid structure to learn additional structural information of confidence responses. We compared our proposed classifier (PEC) and the baseline classifier with different groups of validation subsets.

The performance comparison of four different training groups is described on INRIA and PSU datasets (see Table 4). The first column illustrates four validation subsets to show the performance of the comparison classifier under no occlusion and occlusion. Normally, INRIA contained full body pedestrian images without partial occlusion and was mostly used as the training data for the evaluation tests. However, partial occlusion can occur under the different kinds of conditions—with and without complex scenes—in the real-world environment. Training with different groups

Table 2: Comparison of pedestrian datasets.

Datasets	Imaging setup	Training			Testing			Properties						Publication
		#Pedestrians	#Neg. Images	#Pos. Images	#Pedestrians	#Neg. Images	#Pos. Images	Colour images	Occlusion labels	Illumination	Video seqs.	View labels	Pose labels	
INRIA	Photo	1208	614	1218	556	288	453	√						2005
PSU	Photo	1186	517	1051	1270	517	755	√	√	√		√	√	2018

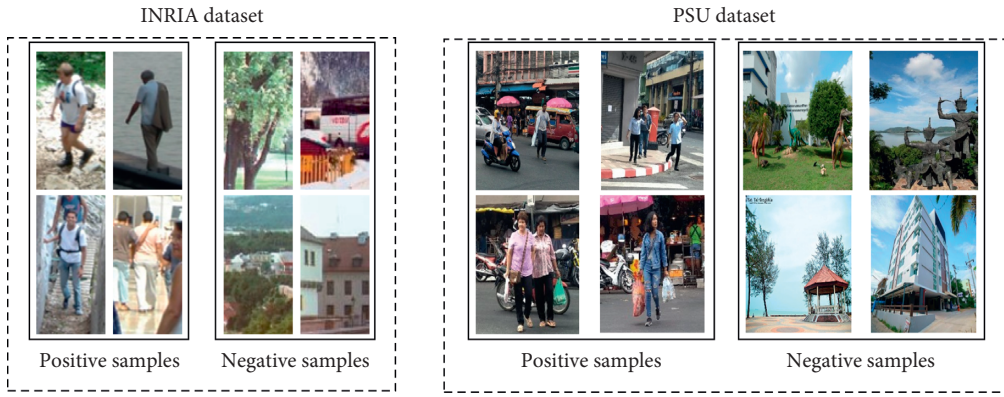

Figure 7: Example of images for training and testing samples from the INRIA person dataset and PSU dataset.

Table 3: Total number of sample data for the validation data.

Validation subsets	Dataset		Total no. of training data		Total no. of testing data	
	Training	Testing	Positive	Negative	Positive	Negative
Vad 1	PSU	PSU	630	1050	100	100
Vad 2	INRIA	INRIA	614	1218	100	100
Vad 3	PSU	INRIA	630	1218	100	100
Vad 4	INRIA	PSU	630	1050	100	100

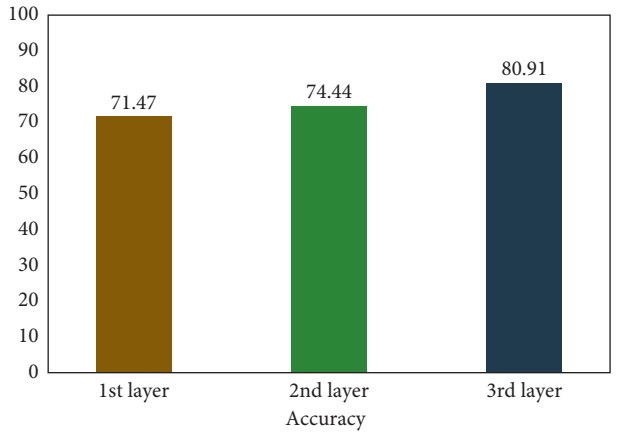

Figure 8: Performance comparison of each pyramid layer accuracy.

of subsets could help the validation performance upon the various scenes. For this reason, we crossly validated with the PSU dataset, containing real-world pedestrian images with partial occlusions under complex scenes. Furthermore, the previous approaches were mostly used to avoid the occluded pedestrian samples in the training stage because the noise in this sample will probably be unreliable to assist the classification.

As we can see in Table 4, the results declined in the general performance while training with the occluded samples on the Vad 1 and Vad 3 subsets especially for the occlusion in the complex scenes. The results displayed in the table demonstrate that the performance of our proposed technique is improved compared with the performance of the base one. This is because of the misclassification of the pedestrian as a background in the complex scenes, and the previous base classifier assumed it as a nonpedestrian in the 1st layer. In contrast, this misclassification part of the body is observed as a pedestrian while using the structural information of the pyramid structure.

A comparison of the overall performance between the HOG + SVM classifier and the PEC on INRIA and PSU datasets for four subsets of data is illustrated (see Figure 9). The results indicate that the precision result of the PEC for four subsets attained higher performance than the HOG + SVM classifier. It can clearly be seen that the overall performance of the proposed technique was 14% improved compared with that of the HOG + SVM technique. Especially for the occlusion with complex scenes, the improvement of the precision results remained significantly higher than that of the compared technique. Compared to that of the INRIA dataset, the precision of the PSU dataset was quite lower when the partial occluded part occurred in the complex background.

For the reasonable case, the proposed technique improved by approximately 7% compared with the HOG + SVM classifier and achieved the best performance when compared with the other subsets. The reason is that the reasonable case contained full body pedestrians without occlusion under complex scenes. In order to know the reason of declined performance for the case of front/back, we trained with the front/back pedestrians with the use of the HOG + SVM classifier and PEC. The results demonstrate that the performance slightly increased 15% while using the proposed techniques when the baseline classifier declined in the results. This is because some appearances of pedestrians' backside are mostly misclassified as nonpedestrians in the baseline classifier due to their appearance variation. However, the proposed technique still increased in the performance compared with the baseline one.

Table 4: Performance comparison on INRIA and PSU datasets (average precision %).

Validation subsets	Reasonable		Front/back		Occlusion without complex scenes		Occlusion with complex scenes	
	HOG + SVM	PEC	HOG + SVM	PEC	HOG + SVM	PEC	HOG + SVM	PEC
Vad 1	91.5	94	72	79	68.5	83.5	61.5	67.5
Vad 2	100	100	79.5	95.5	90.5	95.5	84	94
Vad 3	61	66	63.5	75	54	68.5	52	69
Vad 4	74	94.5	67.5	96.5	62	92	62	97.5

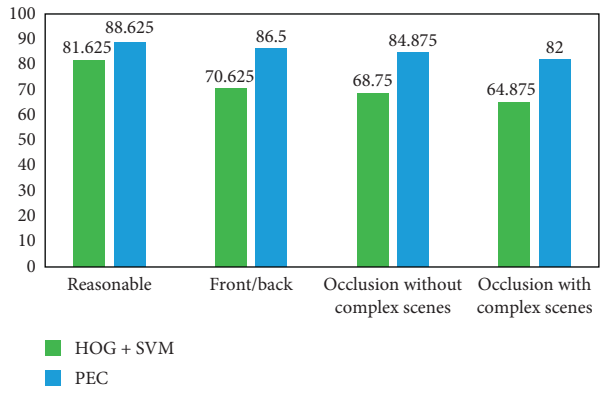

Figure 9: Comparison of the overall performance between the HOG + SVM classifier and the PEC.

For the case of occlusion without complex scenes, the average precision of the proposed technique attained 16% improvement, which is a better performance than the one obtained with the HOG + SVM classifier. Compared with the previous two cases, especially training with the occluded pedestrians decreased the classifier performance due to the noise caused to misclassify the pedestrian as a nonpedestrian.

As can be seen in the result, the average precision of the PEC achieved significantly 17% improvement compared with that of the baseline classifier for the occlusion with the complex scene case. This is because while we used the PEC for each pyramid part representation, the individual confidence score of each pyramidal part can estimate the accurate prediction of partially occluded parts and our model can support the increased performance accuracy. It was found that the average precision of the proposed classifier attained more promising results than that of the basic HOG + SVM classifier.

In order to effectively conduct our proposed technique, four groups of subsets were used to validate the evaluation of the overall performance compared with the baseline approach. The comparison of the validated performances for the HOG + SVM classifier and PEC on INRIA and PSU datasets is shown (see Figure 10). The results demonstrate that the performance of Vad 2 achieved 94% with the PEC, which shows better precision compared with the other subsets. This is because the training and testing with INRIA data samples contained only full body pedestrians without occlusion. It is clearly seen that the proposed technique achieved 8% higher performance than the baseline classifier.

One interesting point of this experiment is that when the train-test samples are crossly used in the experiment, the performance dramatically declined. As we can see from the result of Vad 3, it is found that the proposed PEC achieved 12% higher performance than the baseline classifier. However, when compared with all the validation subsets, the lowest performance is of the Vad 3 subset, which uses the PSU-INRIA data. It is interesting to note that more occluded and complex scenes are contained in the train data and less precision performance is attained. Therefore, the precision accuracy may be quite low compared with that of the other subsets in which less occluded parts are contained in the training set.

On the contrary, the performance of Vad 4 significantly increased by 28% when using the PEC as compared with the baseline classifier while using the train-test subset with INRIA-PSU. The reason is that the training data contained less noise with full body pedestrians without occlusion and complex scenes and the testing data contained partial occluded pedestrians with attached objects in complex scenes. Indeed, the baseline classifier decreased significantly in performance, whereas the proposed technique significantly improved the performance and performed well in the case of partial occlusion under complex scenes. Even for testing with the occlusion in complex scenes, Vad 4 performed well and reached less than 1% lower score in performance compared with Vad 2. From the results, it must be pointed out that the overall performance of the four validation subsets obtained 14% improvement while using the PEC, compared with the use of the HOG + SVM classifier.

A comparison performance of the well-known approaches and our proposed approach is illustrated (see Figure 11). State-of-the-art approaches also achieved good performance results to handle the occlusion of the pedestrians with different points of view. We used the HOG as the feature vector and PEC as the classifier for pedestrians on the INRIA dataset. All of these well-known approaches also used the HOG as the feature with the help of the occlusion map and deformation of the body parts to estimate the location of the occluded part on the INRIA dataset. The performance result of Vad 2 was chosen as the comparison evaluation because it uses the INRIA dataset for training and testing data, similar to the other approaches. The evaluation results clearly show that the proposed techniques outperform the state-of-the-art approaches [6, 7, 12, 14, 20, 25].

HOG + LAMM [7] used the HOG as the feature vector and enhanced cascade with the local area-making map (LAMM). The result shows that the performance of HOG + LAMM [7] was 18% lower in the classification as compared with our PEC

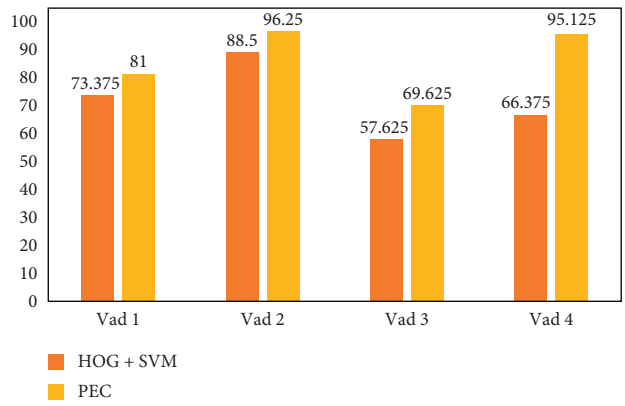

Figure 10: Comparison of the four validation performances between the HOG + SVM classifier and the PEC: Vad 1: train and test with PSU data samples, Vad 2: train and test with INRIA data samples, Vad 3: train with PSU data samples and test with INRIA data samples, and Vad 4: train with INRIA data samples and test with PSU data samples.

performance result. However, HOG + DWT [25] was proposed as a multifeature combined with the HOG and discrete wavelet transform. Some improvements can be observed compared with the previous method, but its performance is still 11% lower than the PEC performance. The DPM [12] is the well-known approach to solve the occlusion problem with the use of the deformation of the body part by using the HOG and latent SVM classifier. To reduce complex calculations, some variations were proposed, such as a faster DPM [14] and PLL-DPM [20]. Note that only the DPM was tested with the Pascal dataset [12]. The best performance of DPM-like methods was 6% lower than that of our method.

HOG + LBP [6] could also solve the partial occlusion problem by combining with the HOG and LBP feature vector with the help of the occlusion likelihood map. The decision of the higher confidence score was based on the global and part detectors to achieve the final classification. The difference between HOG + LBP [6] and our method is that our proposed classifier is dependent on the combination of the confidence scores of the pyramid layer of part representation. In comparison, our proposed technique outperforms the HOG + LBP approach with 5% improvement. Additionally, even using the partial occluded pedestrian with complex scenes, Vad 4, our proposed technique still improves its performance by up to 4%.

However, the ensemble classification response of our method may not perform well where the appearance of the pedestrian is quite similar to the background colour, the pedestrian under the serious occlusion, or the pedestrian present under a crowded scene. In summary, we showed in our experiments that the proposed technique could handle the classification of pedestrians under occlusion in complex scenes and obtains an increased performance result compared with the baseline classifier. Finally, our proposed technique outperformed 96.25% on Vad 2 compared with the other state-of-the-art methods.

Compared with other features, the fast fused part-based model (FFPM) [50], using the spatial deep features combined with six AdaBoost classifiers and Haar-like features, provides the lowest performance. Haar + SVM [9] proposed Haar features with two cascading stages of classifiers to detect the pedestrian in the detection window in the first stage and eliminate some false positives in the second stage. Its performance is comparable to that of the fast DPM [14] at 90%. Compared with some state-of-the-art deep learning approaches, the mixture mask model [16] used multiple mapping vectors to project the original feature matrix into different mask spaces for real-world pedestrian detection. The functional-link net [39] used the cascade AdaBoost detector and random vector functional-link net to enhance the detection accuracy and reduce the number of false positive rates for pedestrian detection. For the fast multiscale object detection, the multiscale CNN (MS-CNN) [47] was proposed to produce accurate object proposals on the detection network with the use of feature upsampling. Another promising YOLO-based architecture also attained improvement on detection of pedestrians. YOLO-based pedestrian detection (Y-PD) [48] and tiny-yolov3 [49] were proposed to modify the network's parameters and structures of the general YOLOv2 detector to identify the suitable characteristic of pedestrians for a better learning. Among these approaches, Y-PD provides the best performance at around 91%, which is 6.5% lower than that of our method.

4.2.3. Experiment Results on the Pretrained Model. For the pretraining approach, we choose the GoogLeNet (Inception v3) architecture among the convolutional neural network (CNN) architectures which is developed with a large ImageNet classification task. In this experiment, Inception v3 also consists of two parts: CNN part for feature extraction and classification part with the fully connected (FC) layer. To perform the fine tuning of the model, we freeze all of the base model's layers to leverage the knowledge by the network from the previous dataset. A new classifier is created to build a new output layer with our own number of classes for the pedestrian dataset. In order to evaluate the benchmark Inception v3 model, we follow the same input size of 224×224 for both training and testing images. For fine tuning the network, the same parameters of the benchmark architecture are used for the learning features. The adjustment of the network also requires the adjustment of the hyperparameters. The learning rates of 0.00001 with the Adam optimizer and the RELU (rectified linear unit) are used to accelerate the training phase. The dropout layer with the parameter of 0.5 is added for reducing overfitting and improving the generalization of the network.

In general, the training scheme of the Inception v3 pretrained model is as follows:

(1) The deep network is retrained by using the pedestrian classification task, i.e., using the image annotations of two classes from the INRIA and PSU datasets.

(2) The network is fine tuned for the pedestrian classification task in the fully connected (FC) layer for predicting labels (two classes). In the last FC layer,

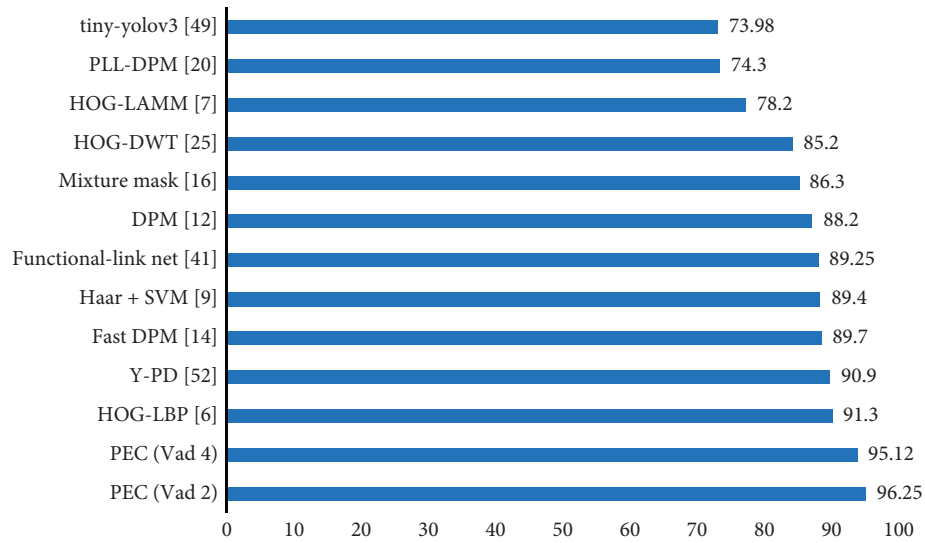

Figure 11: Performance comparison between the proposed approach and the state-of-the-art approaches.

we added a flatten layer, two fully connected dense layers, and a dropout layer with a probability of 0.5. The final softmax layer output is "pedestrian" or "nonpedestrian" from the probabilities of a given normalized class.

In order to conduct our proposed technique, the performance of the baseline HOG + SVM model and benchmark pretrained classifier is compared with that of our proposed classifier (see Figure 12). The evaluation performance is validated on the standard INRIA and PSU datasets. With the use of the standard Inception v3 architecture, the pretrained Inception v3 classifier outperformed the baseline HOG + SVM classifier. As we can see in the performance result, our proposed method achieved 16% improvement compared with Inception v3 on Vad 2 and 28% improvement on Vad 4. However, there is a decline of about 11% performance on Vad 3 compared with Inception v3. Inception v3 improves because of the discriminative feature representation on both local features and high extracted features using the inception module. The advantages of the pretrained deep learning model are the rich representation of local and global features with end-to-end learning. The retraining of model with our datasets might increase the accuracy for better results compared with the pretrained Inception v3, but the training takes a few days, whereas the pretrained model needs only a couple hours. In Vad 3, where considerable occluded pedestrians are included for the training phase, our proposed model degrades in performance during this stage to insufficient HOG features sensitive to noise compared with Inception v3. However, the performance is much better than that of the baseline HOG + SVM classifier.

The overall average performance attained is 71.47% on the HOG + SVM classifier, 78.73% on the benchmark Inception v3 pretrained classifier, and 85.5% on the PEC, respectively. For all classifiers, if noise is present in the input, especially partial occluded parts under complex background, the classification results degrade significantly. However, the

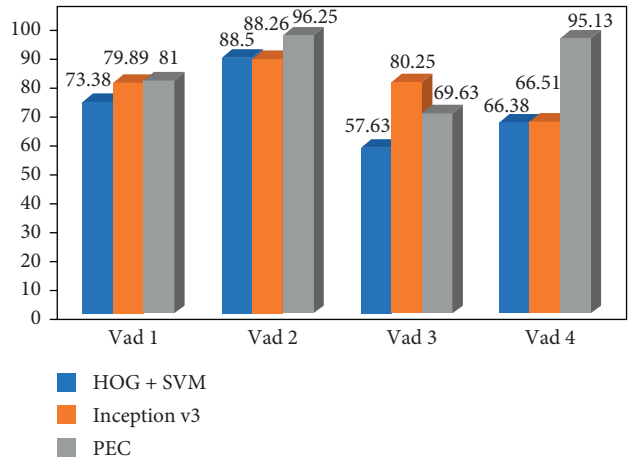

Figure 12: Comparison of the four validation performances between HOG + SVM, Inception v3, and PEC: Vad 1: train and test with PSU data samples, Vad 2: train and test with INRIA data samples, Vad 3: train with PSU data samples and test with INRIA data samples, and Vad 4: train with INRIA data samples and test with PSU data samples.

PEC uses different sizes of windows to estimate the final ensemble score with the sequential estimation of the pyramidal representation via the majority voting, which can better identify the occluded parts. Inception v3 uses the same size of window with noise containing in the training data cloud effecting to the classification performance, as shown in Vad 4.

5. Conclusions

In this paper, the problem of partial occlusion for pedestrian classification is studied. The pyramidal part-based model (PPM) is proposed to obtain a more accurate prediction based on the majority vote of the confidence score of the visible parts by cascading the pyramidal structure. In our

experiments, we trained and tested with different validation subsets on INRIA and PSU datasets. Compared to the state-of-the-art approaches, the average of overall performances achieved 14% improvement. The proposed technique performed 96.25% and 95.12% on datasets without or with occlusions, respectively, outperforming the baseline HOG + SVM and HOG + LBP [6] classifiers. Although we have developed the partial occlusion prediction framework, some works are still to be improved such as pedestrian classification under serious partial occlusion or in crowded scenes with the complex environment. In the future, we will investigate more the structural pyramidal representation with the deep CNN model according to the sequential representation of learning characteristics.

Data Availability

The PSU pedestrian dataset is available in [46], which is authorized only for noncommercial or educational purposes. The additional datasets supporting the current study are cited at relevant places within the text as reference [5].

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the scholarship award of the Higher Education Research Promotion and Thailand's Education Hub for Southern Region of ASEAN Countries (TEH-AC) Project Office of the Higher Education Commission.

References

[1] K. Goniewicz, M. Goniewicz, W. Pawowski, and P. Fiedor, "Road accident rates: strategies and programmes for improving road traffic safety," *European Journal of Trauma and Emergency Surgery*, vol. 42, no. 4, pp. pp433–pp438, 2016.

[2] D. Gerónimo, A. M. López, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1239–1258, 2010.

[3] P. Dollá, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: an evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[4] J. Hariyono, V.-D. Hoang, K.-H. Jo, and Y.-B. Yuan, "Moving object localization using optical flow for pedestrian detection from a moving vehicle," *The Scientific World Journal*, vol. 2014, Article ID 196415, 8 pages, 2014.

[5] N. Dalal and W. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 886–893, San Diego, CA, USA, June 2005.

[6] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision*, pp. 32–39, Kyoto, Japan, September–October 2009.

[7] W. Li, P. Liu, H. Ni, B. Fu, Y. Wang, and S. Wang, "Detection of partially occluded pedestrians by an enhanced cascade detector," *IET Intelligent Transport Systems*, vol. 8, no. 7, pp. 621–630, 2014.

[8] J. Marin, D. Vazquez, A. M. Lopez, J. Amores, and L. I. Kuncheva, "Occlusion handling via random subspace classifiers for human detection," *IEEE Transactions on Cybernetics*, vol. 44, no. 3, pp. 342–354, 2014.

[9] L. Guo, P.-S. Ge, M.-H. Zhang, L.-H. Li, and Y.-B. Zhao, "Pedestrian detection for intelligent transportation systems combining AdaBoost algorithm and support vector machine," *Expert Systems with Applications*, vol. 39, no. 4, pp. 4274–4286, 2012.

[10] X. Zhang, H.-M. Hu, F. Jiang, and B. Li, "Pedestrian detection based on hierarchical co-occurrence model for occlusion handling," *Neurocomputing*, vol. 168, pp. 861–870, 2015.

[11] Z. Zhao, Y. Zhang, L. Bai, Y. Zhang, and J. Han, "Multispectral target detection based on the space-spectrum structure constraint with the multi-scale hierarchical model," *Signal Processing: Image Communication*, vol. 68, pp. 58–67, 2018.

[12] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1627–1645, 2009.

[13] R. B. Girshick, P. F. Felzenszwalb, and D. Mcallester, "Object detection with grammar models," in *Proceedings of the Neural Information Processing Systems*, pp. 1–9, Granada, Spain, December 2011.

[14] P. Luo, Y. Tian, X. Wang, and X. Tang, "Switchable deep network for pedestrian detection," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 899–906, Columbus, OH, USA, June 2014.

[15] C. Zhu and Y. Peng, "Discriminative latent semantic feature learning for pedestrian detection," *Neurocomputing*, vol. 238, pp. 126–138, 2017.

[16] Y. Chen, L. Zhang, X. Liu, and C. Chen, "Pedestrian detection by learning a mixture mask model and its implementation," *Information Sciences*, vol. 372, pp. 148–161, 2016.

[17] J. Wen, X.-p. Wang, L.-f. Kong, and S.-h. Zhang, "Using weighted part model for pedestrian detection in crowded scenes based on image segmentation," *Proceedings of the National Academy of Sciences, India Section A: Physical Sciences*, vol. 86, no. 1, pp. 125–136, 2016.

[18] J. H. Wahyono, J. Hariyono, and K.-H. Jo, "Body part boosting model for carried baggage detection and classification," *Neurocomputing*, vol. 228, no. 8, pp. 106–118, 2017.

[19] W. Ouyang, H. Zhou, H. Li, Q. Li, J. Yan, and X. Wang, "Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 8, pp. 1874–1887, 2018.

[20] L. Geng, Y. Liu, Z. Xiao, Y. Li, and F. Zhang, "Fast pedestrian detection using deformable part model and pyramid layer location," *Journal of Electronic Imaging*, vol. 26, no. 3, Article ID 033020, 2017.

[21] W. Ouyang, X. Zeng, and X. Wang, "Modeling mutual visibility relationship in pedestrian detection," in *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 3222–3229, Portland, OR, USA, June 2013.

[22] D. Ribeiro, J. C. Nascimento, A. Bernardino, and G. Carneiro, "Improving the performance of pedestrian detectors using convolutional learning," *Pattern Recognition*, vol. 61, pp. 641–649, 2017.

[23] B. Wu and R. Nevatia, "Detection and segmentation of multiple, partially occluded objects by grouping, merging,

assigning part detection responses," *International Journal of Computer Vision*, vol. 82, no. 2, pp. 185–204, 2009.

[24] V. P. Viswanath, N. K. Ragesh, and M. S. Nair, "ACM based ROI extraction for pedestrian detection with partial occlusion handling," *Procedia Computer Science*, vol. 46, pp. 45–52, 2015.

[25] G.-S. Hong, B.-G. Kim, Y.-S. Hwang, and K.-K. Kwon, "Fast multi-feature pedestrian detection algorithm based on histogram of oriented gradient using discrete wavelet transform," *Multimedia Tools and Applications*, vol. 75, no. 23, pp. 15229–15245, 2016.

[26] A. Sharifara, M. S. Mohd Rahim, and Y. Anisi, "A general review of human face detection including a study of neural networks and haar feature-based cascade classifier in face detection," in *Proceedings of the 2014 International Symposium on Biometrics and Security Technologies (ISBAST)*, pp. 73–78, Kuala Lumpur, Malaysia, August 2014.

[27] Y. Wei, Q. Tian, J. Guo, W. Huang, and J. Cao, "Multi-vehicle detection algorithm through combining Harr and HOG features," *Mathematics and Computers in Simulation*, vol. 155, pp. 130–145, 2019.

[28] X. Fu, R. Yu, W. Zhang, J. Wu, and S. Shao, "Delving deep into multiscale pedestrian detection via single scale feature maps," *Sensors*, vol. 18, no. 4, p. 1063, 2018.

[29] Y. Liu, P. Lasang, M. Siegel, and Q. Sun, "Multi-sparse descriptor: a scale invariant feature for pedestrian detection," *Neurocomputing*, vol. 184, pp. 55–65, 2016.

[30] J. Yoo, S. S. Hwang, S. D. Kim, M. S. Ki, and J. Cha, "Scale-invariant template matching using histogram of dominant gradients," *Pattern Recognition*, vol. 47, no. 9, pp. 3006–3018, 2014.

[31] D. Sangeetha and P. Deepa, "A low-cost and high-performance architecture for robust human detection using histogram of edge oriented gradients," *Microprocessors and Microsystems*, vol. 53, pp. 106–119, 2017.

[32] C.-H. Zheng, W.-J. Pei, Q. Yan, and Y.-W. Chong, "Pedestrian detection based on gradient and texture feature integration," *Neurocomputing*, vol. 228, pp. 71–78, 2017.

[33] M. Sharif, M. A. Khan, T. Akram, M. Y. Javed, T. Saba, and A. Rehman, "A framework of human detection and action recognition based on uniform segmentation and combination of Euclidean distance and joint entropy-based features selection," *Eurasipian Journal of Image Video Processing*, vol. 2017, no. 1, p. 89, 2017.

[34] L. Guo, M. Zhang, L. Li, Y. Zhao, and Y. Lin, "Body parts features-based pedestrian detection for active pedestrian protection system," *PROMET—Traffic&Transportation*, vol. 28, no. 2, pp. 133–142, 2016.

[35] A. Sebti and H. Hassanpour, "Body orientation estimation with the ensemble of logistic regression classifiers," *Multimedia Tools and Applications*, vol. 76, no. 22, pp. 23589–23605, 2017.

[36] M. A. Muqeet and R. S. Holambe, "A collaborative representation face classification on separable adaptive directional wavelet transform based completed local binary pattern features," *Engineering Science and Technology, an International Journal*, vol. 21, no. 4, pp. 611–624, 2018.

[37] S. Kim and K. Cho, "Fast calculation of histogram of oriented gradient feature by removing redundancy in overlapping block," *Journal of Information Science and Enigneering*, vol. 30, no. 6, pp. 1719–1731, 2014.

[38] S. Zhang, D. A. Klein, C. Bauckhage, and A. B. Cremers, "Fast moving pedestrian detection based on motion segmentation and new motion features," *Multimedia Tools and Applications*, vol. 75, no. 11, pp. 6263–6282, 2016.

[39] Z. Wang, S. Yoon, S. J. Xie, Y. Lu, and D. S. Park, "A high accuracy pedestrian detection system combining a cascade AdaBoost detector and random vector functional-link net," *The Scientific World Journal*, vol. 2014, Article ID 105089, 7 pages, 2014.

[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, Las Vegas, NV, USA, December 2016.

[41] Y. Han, T. Jiang, Y. Ma, and C. Xu, "Pretraining convolutional neural networks for image-based vehicle classification," *Advanced in Multimedia*, vol. 2018, Article ID 3138278, 10 pages, 2018.

[42] B. Dai, Y. Wang, Y. Yao, W. Ye, and T. Chen, "Retracted: efficient object analysis by leveraging deeply-trained object proposals prediction model," *Journal of Visual Communication and Image Representation*, vol. 61, pp. 218–224, 2019.

[43] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object Detection with Deep Learning: A Review," *IEEE Transattion on Neural Networks Learning System*, vol. 30, no. 11, pp. 3212–3232, 2019.

[44] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7, pp. 53040–53065, 2019.

[45] C. Szegedy, "Going deeper with convolutions," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, Boston, MA, USA, June 2015.

[46] M. Thu, N. Suvonvorn, and M. Karnjanadecha, "A new dataset benchmark for Pedestrian detection," in *Proceedings of the 3rd International Conference on Biomedical Signal and Image Processing*, pp. 17–22, Bari, Italy, October 2018.

[47] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," *Computer Vision—ECCV 2016*, vol. 9908, pp. 354–370, 2016.

[48] Z. Liu, Z. Chen, Z. Li, and W. Hu, "An efficient pedestrian detection method based on YOLOv2," *Mathematical Problems in Engineering*, vol. 2018, Article ID 3518959, 10 pages, 2018.

[49] Z. Yi, S. Yongliang, and Z. Jun, "An improved tiny-yolov3 pedestrian detection algorithm," *Optik*, vol. 183, pp. 17–23, 2019.

[50] E. J. Cheng, "A fast fused part-based model with new deep feature for pedestrian detection and security monitoring," *Measurement*, vol. 151, Article ID 107081, , 2019.