Multimedia Tools and Applications (2020) 79:21389–21408
https://doi.org/10.1007/s11042-020-08864-z

A heuristic SVM based pedestrian detection approach employing shape and texture descriptors

Kaushal Kumar[1] · Ritesh Kumar Mishra[1]

Received: 4 April 2019 / Revised: 8 March 2020 / Accepted: 18 March 2020 /
Published online: 6 May 2020

Abstract
Pedestrian detection is a vital issue in various computer vision applications such as smart security system, driverless car, smart traffic management system and so forth. However, the issue of low detection accuracy and high computational complexity still makes a prompt topic of research. In the current scenario, Histogram of Oriented Gradients (HOG) with linear Support Vector Machine (SVM) is considered to be the most discriminative detector and has been adopted in various advance systems. In this paper, a novel method for pedestrian detection is proposed with the objective of improving the detection accuracy, precision and other metrics values. The proposed approach combines Histogram of Significant Gradients (HSG) and Non Redundant Uniform Local Binary Pattern (NRULBP) to generate a competent descriptor to be used in our detection model. The proposed approach is used in conjunction with various classifiers and the linear SVM classifier is found to provide better metric values over others. Different datasets like INRIA, TUD-brussels-motion pairs and ETH are utilized for performing experiments and to obtain detection results. Experimental results show that the proposed descriptor outperforms HSG by 2.59%, 8.97%, 8.5% and NRULBP by 3.19%, 39.55%, 19.66% in terms of detection accuracy, precision and F1 score respectively.

Keywords Computer vision · Histogram of gradients · Non redundant uniform local binary pattern · Pedestrian detection · SVM

1 Introduction

Object detection is gradually becoming an important part of smart video surveillance system. In the past, various researchers have proposed different algorithms to extract the

✉ Kaushal Kumar
kaushal.ec16@nitp.ac.in

Ritesh Kumar Mishra
ritesh@nitp.ac.in

[1] Department of Electronics and Communication Engineering, National Institute of Technology, Patna, India

corresponding features that would enable a better detection of various objects of interest. Among them, pedestrian detection is found to be a prompt topic of research due to its various real-time applications. In the field of computer vision, it is regarded as one of the most challenging issue because of various forms of clothing and poses a human exhibit. The commercial employment of human detection system is still not available due to its low precision and high complications.

A common technique of object localization is to scroll detection window over the entire image. The detection window results in binary outcome which is one when pedestrian exists in the image under consideration otherwise zero. The process is based on training a classifier to detect the pedestrian features which is deduced from a number of positive images containing pedestrian and negative images containing non-pedestrian images. Template matching and feature extraction combined with a classifier are the most common algorithms used for pedestrian detection. The most common classifier used for this purpose is Support Vector Machine (SVM). SVM is a differentiating classifier that precisely divides a group of data into various classes by introducing a hyperplane among them. That is, provided a labeled training data, the algorithm results in a perfect hyperplane which differentiates the test data. In this paper we employed various classifiers for cross-validating the model of our proposed technique and it is observed that SVM with linear kernel function gives highest accuracy among others to detect pedestrians from a given test datasets. The proposed method involves two stages i.e. training and testing as shown in Fig. 1. In training phase, classification model is created which is used in testing phase for detecting pedestrian.

The novel contributions of the work presented in this paper are summarized as follows:

- Illustrating the effectiveness of proposed descriptor that combines HSG and NRULBP descriptors. The features obtained from proposed descriptor are fed to various classifiers and linear SVM classifier comes out with best performance metrics values among all.

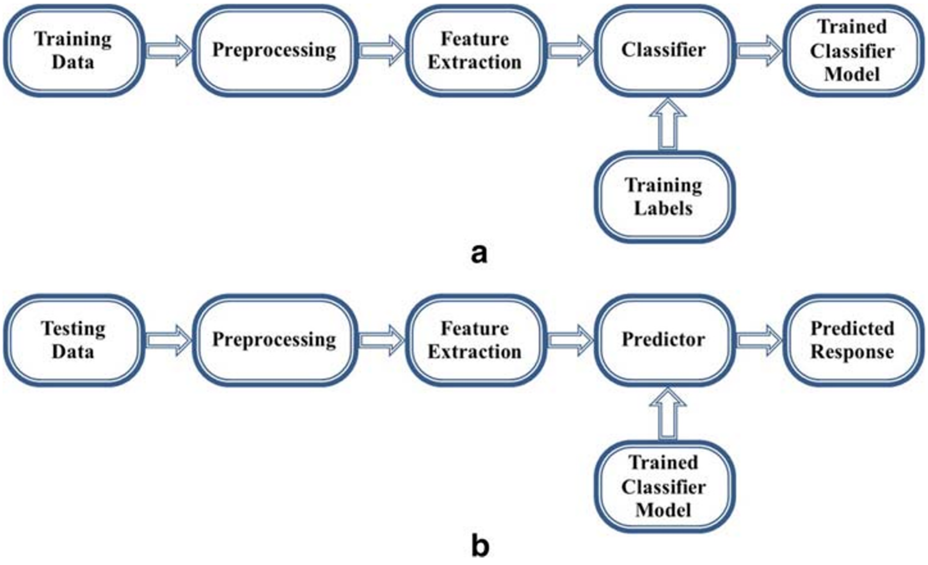


Fig. 1 Block diagram of pedestrian detection system: (**a**) Training stage (**b**) Testing stage

- Experiments on various dataset has been conducted and it is observed that the proposed framework has accomplished a better performance metric values than others and thus is robust for real time applications.

The rest of paper is categorized as follows: Section 2 describes various descriptors that have been used for extracting feature vector for the classifier. Section 3 provides an outlook of our approach to extract the feature for detecting pedestrians. Section 4 compares the results of our approach with various existing techniques mentioned in above sections. And Section 5 concludes the work highlighting important findings.

2 Related work

The significance of pedestrian detection in images and videos has fetched a lot consideration from computer vision people due to increasing multimedia and surveillance applications [13, 20, 29, 35, 37, 39]. The important units of a pedestrian detection framework are feature descriptor and feature classifier. These units play a significant role in deciding overall performance of the system in terms of accuracy, precision and recall. A practical feature descriptor must be capable of extracting all possible visual cues that human visual system utilizes to differentiate person and object under different situations.

Many studies have been done for extracting features efficiently to detect pedestrian in different scenarios. Some of the widely used descriptors are wavelets [19, 23], Local Binary Pattern (LBP) [16, 17, 32], Scale Invariant Feature Transformation (SIFT) [15], Speeded-Up Robust Features (SURF) [1], Histograms of Oriented Gradients (HOG) [4, 22] and Edge Orientation Histograms (EOH) [9]. Of these, HOG descriptor suggested by Dalal and Triggs [4] results in better performance and remains unaffected from variations in illumination, scaling and rotation. This descriptor gives the information about contour of objects present in image. Support Vector Machine (SVM) classifier is used with HOG descriptor for detecting pedestrian [2, 3, 36]. But this SVM-HOG model is subjected to high false alarm rate [6]. SIFT [15] was proposed by David G. Lowe et al. and are widely used in recognition and localization of targets. HOG and SIFT are similar in their implementation. SIFT is calculated on feature points acquired from difference of Gaussian detectors while in HOG, gradient vector is calculated over an overlapping windows. But this technique is limited by high computational complexity [33]. Herbert Bay et al. proposed SURF which reduces the computational complexity [1]. It uses box filters which improves SIFT features. EOH is also similar to HOG where histogram is build using directions of the gradients of borders [9]. Both SIFT and EOH has less performance measures compared to HOG [3]. Michael oren et al. used wavelets decomposition, containing information about shape of objects, for pedestrian detection [19]. They reported the presence of contours in image using the obtained coefficients. A person can also be characterized using its texture information. Ojala et al. proposed LBP descriptor to depict texture information [17]. Due to high complexity of this classic texture descriptor, it is not suitable for human detection. LBP was expanded in various ways. In order to enhance performance, some of the proposed extensions of LBP are Elongated Local Binary Patterns (ELBP) [12] and Rotation Invariant LBP [10]. Nguyen et al. [16] suggested a variant of LBP called NRLBP which is formed by taking the minimum of LBP code and its complement. Wang et al. recommended a combination of HOG and LBP features for creating histogram which handle occlusion [27]. Lie et al. also suggested the same combination for multi-part detection [14]. Wojek et al. proposed a combination of HOG, Haar and shape context features for increasing detector performance [30]. Xin et al. combined HOG and Haar features which results in high detection accuracy and speed [34]. Dangwei Li et al. come up with new Richly Annotated Pedestrian (RAP) dataset which can be employed for attribute based person retrieval and person re-identification problems [11]. Shanshan Zhang et al. has the analysis regarding the failures for top detectors on Caltech and KITTI datasets. The authors come up with new annotations of Caltech training and testing datasets to minimize the error of top detectors [38]. D.K. Vishwakarma et al. presented an approach for human action and activity recognition using the shape and rotation of human body [5, 24–26]. In brief, although HOG descriptor provide a better performance over others for detecting pedestrian, it is advantageous to use it in combination with other techniques at the cost of more computational complexity.

The accuracy of pedestrian detection is greatly varied with the type of classifier used. Dalal et al. used linear SVM in its HOG based pedestrian detection due to ease of implementation [4]. They reported an increase in performance by 3% at 10^{-4} FPPW using gaussian SVM but at lower speed. Viola et al. used Adaboost for training classifiers which reduces the computation time [23]. The classifier which results from Adaboost algorithm consists of a linear composition of sets of feature. In this paper, our aim is to design an improved pedestrian detector chiefly by enhancing the shape descriptor. It is found that when HSG is combined with NRULBP, the new improved descriptor is superior to either HSG or NRULBP.

2.1 Histogram of gradients (HOG)

HOG is shape descriptor. In this method, the searching window is segmented into various blocks which are overlapping in nature and these blocks are again segmented into individual cells as shown in Fig. 2. The area to be overlapped is usually taken as 50%, which results

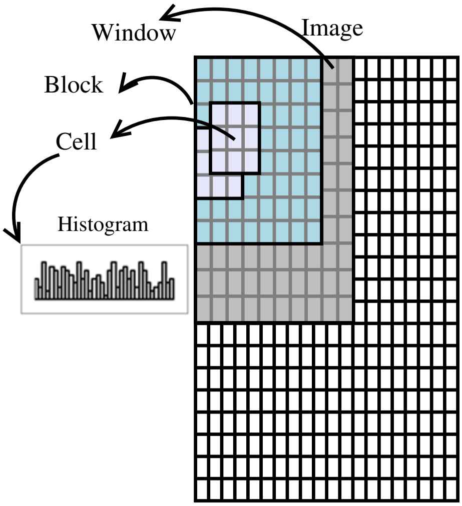


Fig. 2 Image division into cells for histogram calculations

in high density grid over the searching window. These blocks are further divided into cells and gradient information is extracted from these cells to create the histogram.

Let G(x,y) gives gives information about pixel value at (x,y), then its gradient vector is given by (1).

$$\nabla G(x,y) = \begin{bmatrix} g_x \\ g_y \end{bmatrix} = \begin{bmatrix} \frac{\partial g}{\partial x} \\ \frac{\partial g}{\partial y} \end{bmatrix} = \begin{bmatrix} g(x+1,y) - g(x-1,y) \\ g(x,y+1) - g(x,y-1) \end{bmatrix} \tag{1}$$

In order to get gradients g_x and g_y , a filter with coefficient of [1 0 -1] which is called central difference is applied in both horizontal and vertical directions of the window. The computed value of g_x and g_y are then utilized for the computation of two significant parameters i.e. gradient magnitude and gradient orientation using (2) and (3).

$$g_{mag}(x,y) = \sqrt{g_x(x,y)^2 + g_y(x,y)^2} \tag{2}$$

$$g_\theta(x,y) = \tan^{-1}\frac{g_y(x,y)}{g_x(x,y)} \tag{3}$$

Using magnitude and orientation matrix, histogram is built consisting of several bins, where every bin depicts a particular orientation in the class $[0, \pi]$. For a specific orientation of the pixel which corresponds to a specific bin of the histogram, the magnitudes of pixels are summed. This results in formation of a number of histograms corresponding to the each cell present in the window. The concatenation of all histograms leads to the formation of feature vector of the window. Pyramid Histogram of Oriented Gradients (PHOG) is a spatial pyramid description of HOG [28]. Some applications implemented with PHOG are found to have increased detection accuracy in contrast to HOG. This motivates the pedestrian detection to be also implemented with PHOG. In PHOG, the spatial layout of image is divided into subregions. The HOG descriptor is then applied into subregions. The final feature is the concatenation of features of subregions.

2.2 Local binary patterns (LBP)

Local Binary Pattern is a texture descriptor. By comparing the 3x3 neighbor pixel value with the center pixel value taken as threshold, the LBP feature is developed. The basic concept of LBP is shown in Fig. 3. The texture, X can be defined as the joint distribution of the $\rho + 1$ gray level image pixels given by (4).

$$X = x(P_c, P_0, ..., P_{\rho-1}) \tag{4}$$

where, P_c corresponds to center pixel. $P_\rho(\rho = 0, 1, ..., \rho - 1)$corresponds to the ρ equally spaced pixels on a circle of radius, r (r>0). Also, without losing information P_c can be subtracted from P_ρ as given in (5).

$$X = x(P_c, P_0 - P_c, ..., P_{\rho-1} - P_c) \tag{5}$$

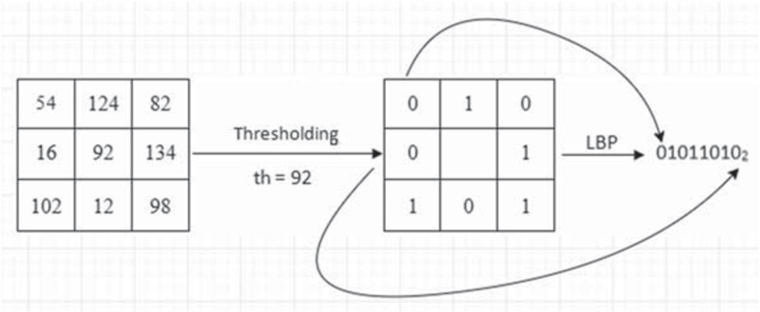


Fig. 3 Illustration of standard LBP operator on a pixel

Supposing the differences are not dependent on P_c, the above equation can be factorized as given in (6).

$$X \approx x(P_c)x(P_0 - P_c, ..., P_{\rho-1} - P_c) \tag{6}$$

$x(P_c)$ can be ignored as it gives luminance which is not related to local image texture. Equation (6) can now be rewritten as given in (7).

$$X \approx x(P_0 - P_c, ..., P_{\rho-1} - P_c) \tag{7}$$

The above equation remains unaffected from gray level shift but is affected by scaling. For incorporating the invariance with reference to any monotonic transformation of gray scale, only the sign of the differences are considered as given in (8) and (9).

$$X \approx x(s(P_0 - P_c), ..., s(P_{\rho-1} - P_c)) \tag{8}$$

where,

$$s(d) = \left\{ \begin{matrix} 1, d \geq 0 \\ 0, d < 0 \end{matrix} \right\} \tag{9}$$

To create LBP code, a binomial weight 2^ρ is assigned to each neighbor as given in (10).

$$LBP_{N,R}(x_c, y_c) = \sum_{\rho=0}^{N-1} s(P_\rho - P_c)2^\rho \tag{10}$$

In this paper, R is taken as 1, which results in 8 neighbors (N) of center pixel. So, the LBP code for a pixel formed by 3x3 neighbor with center pixel taken as threshold is given by (11).

$$LBP = \sum_{\rho=0}^{7} s(P_\rho - P_c).2^\rho \tag{11}$$

The LBP descriptor is advantageous in the sense that the required processing complexity is low and also is unaffected by change of neighbor pixel intensity value. But as the difference between values of neighbor pixel and center pixel is relatively small, generation of basic LBP patterns in uniform image patches may lead to unreliable results. This is generally solved by modifying the threshold as the mean/median of the 3x3 pixels or by adding a bias. The LBP for each pixel in window is calculated. Then the window is divided into 8x8 blocks and the histogram is calculated for each block. The number of bins of histogram depends on the variant of LBP (For basic LBP, 256 bins). The histograms of each block are concatenated to give feature of the window.

2.3 Non redundant uniform local binary patterns (NRULBP)

NRULBP is a variant or modification of the LBP descriptor used to decrease the size of the feature vector. Nguyen et al. claimed that the basic LBP has two main drawbacks: high number of possible patterns and sensitiveness to comparative changes between the foreground and background [16]. Thus, there is introduction of non-redundant local binary patterns (NRLBP). In case of NRLBP, the LBP patterns and its complements are believed to be same (for ex. 00101110 and 11010001) and thus NRLBP is considered as the minimum of LBP and its complement as given in (12).

$$NRLBP = min\{LBP, 2^N - 1 - LBP\} \tag{12}$$

Ojala et al. introduced the notion of uniform LBP by virtue of which the required number of possible LBP patterns reduce from 256 to 59 while keeping its discrimination power [18]. If the number of transitions between successive bits in a binary vector is at most two, then the LBP pattern is called as uniform LBP, shown by (13) and (14).

$$ULBP = \left\{ \begin{array}{cc} \sum_{\rho=0}^{N-1} s(P_\rho - P_c).2^\rho & if T(LBP) \leq 2 \\ 5 & otherwise \end{array} \right\} \tag{13}$$

where,

$$\begin{gathered} T(LBP) = \mid s(P_{N-1} - P_c) - s(P_0 - P_c) \mid + \\ \sum_{\rho=1}^{\rho-1} \mid s(P_\rho - P_c) - s(P_{\rho-1} - P_c) \mid \end{gathered} \tag{14}$$

Combining the two variants, a new variant is proposed termed as Non redundant uniform local binary pattern (NRULBP). In comparison to LBP, in NRULBP the number of bins required to create histogram reduces by approximately 88%. This variant is calculated by evaluating the intersection between ULBP and NRLBP, as shown in (15).

$$NRULBP = ULBP \cap NRLBP \tag{15}$$

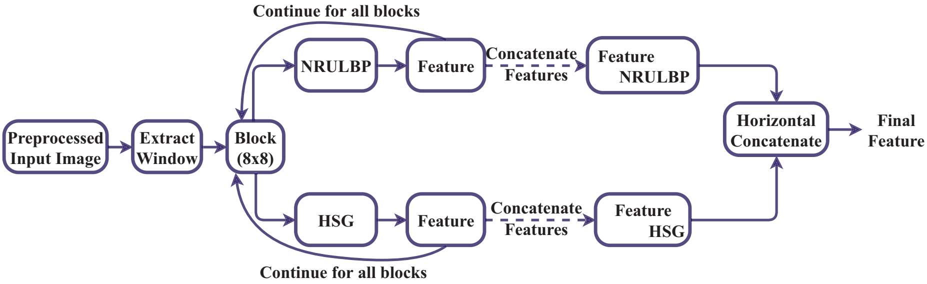


Fig. 4 Block diagram for the proposed approach

3 Proposed approach

In the previous section, we mentioned how each technique extracts the features of an image. Most of these techniques describe either the shape or the texture of the image. Combining the two descriptors will help to increase the performance and accuracy of the system. Thus we combine the features of a Histogram of Significant Gradients (HSG) descriptor and a NRULBP descriptor for detecting pedestrian as shown below in Fig. 4.

The HSG descriptor is a modification of the HOG descriptor. The image is partitioned into windows, which is further splited into blocks. In this work, a window of size 64 × 128 is used with a stride of 8 between each window. The block size is 8x8 with stride of 4 pixels. Thus there is a 50% overlap of blocks. The gradient magnitude and its orientation are manipulated for each pixel present in the block similar to HOG. Then instead of creating a histogram of the normalized sum of the gradient magnitude, the average gradient of each block is manipulated here. This average acts as a threshold value. Edges with gradient magnitudes above the average gradient value are considered to be more significant than the other edges. Only these edges are allowed to cast binary vote to the orientation histogram. It is similar to EOH as it allows only binary votes to be cast. But the HSG technique identifies the shape edges more effectively by recognizing only the more significant edges. The binary votes are cast for each cell to create a histogram. The histograms usually consist of 9 bins. The feature vector is generated by concatenating the histograms. NRULBP as mentioned previously is a variant of the LBP that has less number of bins. The fact that the number of bins of both HSG and NRULBP is low compared to other descriptors makes them a good choice as this reduces the feature size of the individual descriptors. The size of the feature determines the processing time of the classifier. In this work, we combine the features of the HSG descriptor and the NRULBP descriptor to form a new feature vector. These features are labelled and given to the classifier for training. The step by step process of proposed methodology is as follows:

```
  INPUT:   Image for detecting the presence of pedestrian.
image_w=width of image
image_h=height of image
window_width=64
window_height=128
number_win_x=image_w-window_width
number_win_y=image_h-window_height
for l= 0 : number_win_y
  for m= 0 : number_win_x
function $HSG_{descriptor}(Image)$
   Calculate gradients $G_x$ and $G_y$ along x and y direction
   Magnitude and direction of each pixel is calculated as:
   $G_{mag} = G_x + G_y$ and $G_{angle} = atan(G_y/G_x) * (180/\pi)$
   block_w=8
   stride=4
   number_block_x=(window_width-block_w)/stride
   number_block_y=(window_height-block_w)/stride
for c= 0 : number_block_y do
    for d= 0 : number_block_x do
        The average over ith block is given as:
        $M(B_i) = \frac{g_{mag}(x_1,y_1)+g_{mag}(x_2,y_2)+......+g_{mag}(x_8,y_8)}{64}$
        for i=1 : block_w do
            for j=1 : block_w do
                for n=1 : 9 do
                    if pixel, $P(x_i, y_j) >= n * 20 and P(x_i, y_j) < (n+1) * 20$ then
                        Bin for each pixel:
bin$(x_i, y_j) = n$
                    end if
                end for
            end for
        end for
        The value of $j^{th}$ bin for $i^{th}$ block is given by:
        $S_i(j) = \sum T(g_{mag}(x_n, y_n) > M(B_i))$
        where, $T(k) = \left\{ \begin{array}{ll} 1 & ;\ k = True \\ 0 & ; otherwise \end{array} \right\}$
        The histogram for $i^{th}$ block $(U_i)$ is given by:
        $U_i = |S_i(1)S_i(2)......S_i(9)|$
    end for
end for
   $H_1 = |U_1U_2.......U_i|$
return $H_1$
function $NRULBP_{descriptor}(Image)$
   Image is padded with pixel of value '0' all around
   w=width of image
   h=height of image
   block_size=8
   number of columns=width/block_size
   number of rows=height/block_size
```

```
for i=2 : h-1 do
    for j=2 : w-1 do
        Let $P(i,j)$ be center pixel
        $LBP(i,j) = \sum_{y=0}^{7} f(P_y - P_(i,j)).2^y$
        where, $f(x) = \begin{Bmatrix} 1 & ; x >= 0 \\ 0 & ; otherwise \end{Bmatrix}$
        $LBP_{COMP} = !(LBP)$
        $NRLBP = minimum(LBP, LBP_{COMP})$
        if the number of transitions between successive bits of NRLBP>2 then
            $NRULBP = 5$
        else
            $NRULBP = NRLBP$
        end if
    end for
end for
Let the value of $l^{th}$ bin for $m^{th}$ block is given as:
$V_m(l) = \sum T(NRULBP == l)$
where, $T(k) = \begin{Bmatrix} 1 & ; k = True \\ 0 & ; otherwise \end{Bmatrix}$
The histogram for $m^{th}$ block is given as:
$R_m = |V_0 V_1 ...... V_{127}|$
$H_2 = |R_1 R_2 ...... R_m|$
return $H_2$
Final feature, $H = |H_1 H_2|$
  end for
end for
   OUTPUT:  Feature, H is calculated for all windows of an image to detect pedestrian
            present in that window.
```

For classification of training dataset, various classifiers like SVM, naive bayes, LDA(Linear Discriminant Analysis), KNN(K Nearest Neighbor) and decision tree are employed in conjunction with proposed technique to find the classifier with best performance metrics. Among SVM, different kernel functions used in this paper are linear, quadratic, polynomial, Radial Basis Function(RBF) and MultiLayer Perceptron(MLP). Using experimentations, linear SVM is found to have better performance over others. Also linear SVM is simplest in its implementation. Let 'H' represents the final feature vector of n training images in dataset, given as $H^T = H_1^T + H_2^T + ... + H_n^T$ and their associated class labels represented by $L_i \in \{-1, +1\}$. Also limitations are required to be imposed for the instances to be correctly classified as given in (16) and (17).

$$wH_i + z \geq +1 \quad if L_i = +1 \tag{16}$$

$$wH_i + z < -1 \quad if L_i = -1 \tag{17}$$

Equations (16) and (17) can be equivalently written as given in (18). z represents the offset for the hyperplane. It corresponds to the bias augmented with vectors during calculation of hyperplane.

$$L_i(wH_i + z) \geq 1 \tag{18}$$

SVM is used to create hyperplane between the vectors present on boundaries (known as support vectors) with maximum distance between them. The distance or gap between the support vectors is given by (19)

$$G = 2/||w|| \tag{19}$$

From (19), it can be observed that for maximizing the gap, $||w||$ should be minimum or minimize $\frac{1}{2}||w||^2$. Thus, minimizing $\frac{1}{2}||w||^2$ leads to $L_i(wH_i + z) - 1 \geq 0$ for i=1,...,n. Thus, the classification function of the linear SVM classifier for a new feature H is given by (20).

$$f(H) = sign(wH + z) \tag{20}$$

Where, 'w' represents the weight vector of linear SVM expressed as $w^T = w_1^T + w_2^T + ... + w_n^T$. For higher dimension feature space, the classification function for a new feature H is given by (21).

$$f(H) = sign(wk(H_i, H) + z) \tag{21}$$

where, k represents kernel function. In case of linear SVM, k is dot product of H_I and H_j Some of the popular kernels like radial basis function (22), polynomial (23) used with SVM are given below.

$$k(H_i, H_j) = exp\Big(-\frac{||H_i - H_j||^2}{2\gamma^2}\Big) \tag{22}$$

where, γ represents width of kernel

$$k(H_i, H_j) = (1 + H_i H_j)^p \tag{23}$$

where, p represents the polynomial order.

Deep learning is an state-of-the-art technique in the field of object detection. So, the YOLOv2 model is also modified here for the detection of pedestrian [21]. YOLOv2 is trained using transfer learning method. It is also trained and tested on same dataset as used by other descriptors. The number of classes and filters in the last convolution layer is modified to use YOLOv2 for pedestrian detection. Since, the problem is related to binary classification, the number of classes is changed to 2 and the number of filters to 30.

4 Results and discussions

4.1 Dataset description

A number of pedestrian datasets, such as INRIA [4], Daimler [7], TUD-Brussels-Motion Pairs [31], ETH [8] are available for training and testing the performance of pedestrian detectors in various environments. In this paper, INRIA dataset is used for training and testing pedestrian detectors. The detection window is a rectangular window of size 64×128 pixels. INRIA consists of train and test images of sizes 96×160 (16 pixels margin around each side) and 70×134(3 pixels margin around each side) respectively. Margins are given to avoid boundary conditions while calculating the gradients and orientation of pixels. Daimler, TUD-Brussels-Motion Pairs and ETH are employed for obtaining the detection outcomes. The dataset consists of 5500 training and 1821 testing images. Out of 5500 training images, there are 500 positive images consisting of pedestrian and 5000 negative images having no pedestrian. The test dataset consists of 739 pedestrian images and 1082 non-pedestrian images.

4.2 Performance analysis

In order to measure the performance of proposed system and to do comparison with other state of the art detectors, various metrics are used. Such metrics are defined in terms of the components of confusion matrix i.e. TP(True Positive), TN(True Negative), FP(False Positive) and FN(False Negative). Confusion matrix is used to obtain the performance of classifier on a set of test data. It allows visualization of confusion between classes. It summarizes the number of correct and incorrect predictions. TP means positive images are predicted as positive. TN means negative images are predicted as negative. FP means negative images are predicted as positive and FN means positive images are predicted as negative. The confusion matrix for the proposed technique using linear SVM classifier is shown in Fig. 5. FP and FN are also known as type I and type II errors respectively which should be minimum compared to TP and TN for better classification. It can be observed that TP(709) and TN(1071) are relatively high compared to FP(11) and FN(30), which results in good detection accuracy. To illustrate the performance of detector, we plot Receiver Operating Characteristics (ROC) and Precision-Recall (PR) curves. The ROC curve computes the tradeoff between sensitivity and 1- specificity at different classification thresholds. Area Under Curve (AUC) gives an aggregate value of performance across different thresholds of ROC curves. It represents the degree of separability. Higher value of AUC represents the model is better in predicting ’0’ as ’0’ and ’1’ as ’1’. PR curve provides the tradeoff

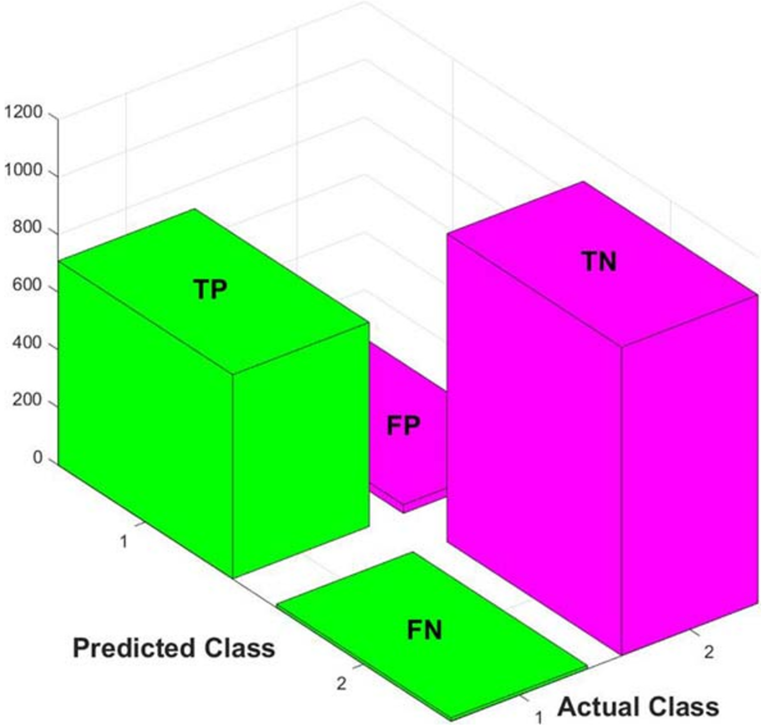


Fig. 5 Confusion matrix of the proposed framework using linear SVM classifier

Table 1 Comparison of performance of various descriptors using linear SVM classifier

Techniques → Metrics ↓	Block gradient	HOG	PHOG	HSG	NRULBP	YOLOv2	Proposed
Accuracy(%)	81.38	85.72	81.43	91.59	82.04	89.73	97.74
Precision	0.7674	0.8059	0.7235	0.9036	0.7056	0.9539	0.9847
Recall	0.7767	0.8539	0.8782	0.8877	0.9567	0.7848	0.9594
Specificity	0.8392	0.8595	0.7708	0.9353	0.7274	0.9741	0.9898
NPV	0.8462	0.8960	0.9026	0.9242	0.9609	0.8689	0.9728
Miss rate	0.2233	0.1461	0.1218	0.1123	0.0433	0.2152	0.0406
Fall out	0.1608	0.1405	0.2292	0.0647	0.2726	0.0259	0.0102
FDR	0.2326	0.1941	0.2765	0.0964	0.2944	0.0461	0.0153
For	0.1538	0.1040	0.0974	0.0758	0.0391	0.1311	0.0272
F1 score	0.7720	0.8292	0.7934	0.8956	0.8122	0.8612	0.9719
MCC	0.6148	0.7076	0.6375	0.8254	0.6752	0.7903	0.9533
Youden's index	0.6159	0.7134	0.6490	0.8230	0.6841	0.7590	0.9492

between precision and recall at different classification thresholds. In Table 1, performance comparison is performed between the proposed approach and other approaches employing linear SVM. It can be observed that accuracy is highest for proposed technique i.e. 97.74% followed by HSG (91.59%), YOLOv2 (89.73%), HOG (85.72%), NRULBP (82.04%), PHOG (81.43%) and block gradient (81.38%). Precision is also highest for the proposed technique (0.9847) which is the ratio of TP to the total predicted positive observations. Recall is highest for proposed approach (0.9594) followed by NRULBP (0.9567), HSG (0.8877), PHOG (0.8782), HOG (0.8539), YOLOv2 (0.7848) and block gradient (0.7767). It is the ratio of predicted correct out of all positive class. When a classifier has high precision and low recall, it indicates that we failed to detect most of the positive images but those predicted as positive are certainly positive.

In such situation, it is preferred to evaluate F1 score, which is a function of precision and recall. From Table 1, it can be observed that F1 score is highest for the proposed technique (0.9719) followed by HSG, YOLOv2, HOG, NRULBP, PHOG and block gradient. Specificity is the ability of classifier to predict negative result out of all negative class which is highest for proposed technique (0.9898). Negative Predictive Value (NPV) is the ratio of correct negative results to the total predicted negative results which is highest for proposed technique (0.9728) followed by NRULBP. Miss rate and fall out are errors in prediction and should be least. It is found that, both are lowest for proposed technique. False Discovery Rate (FDR) is the ratio of FP to the total predicted positive observations. False Omission Rate (For) is the ratio of incorrect negative results to the total predicted negative results. Both are minimum for the proposed approach. Matthews Correlation coefficient (MCC) evaluates the quality of binary and multi-class classification, whose value lies between -1 and +1. +1 indicates best prediction and -1 indicates poor prediction. MCC for the proposed technique is nearer to +1 i.e. 0.9533 which indicates good prediction compared to other techniques. Youden's index value lies between 0 and 1, where 1 indicates perfect prediction. It is also highest for proposed approach (0.9492) followed by HSG (0.8230), YOLOv2 (0.7590), HOG (0.7134), NRULBP (0.6841), PHOG (0.6490) and block gradient (0.6159). Figure 6 provides comparison of ROC curves and PR curves between proposed and other

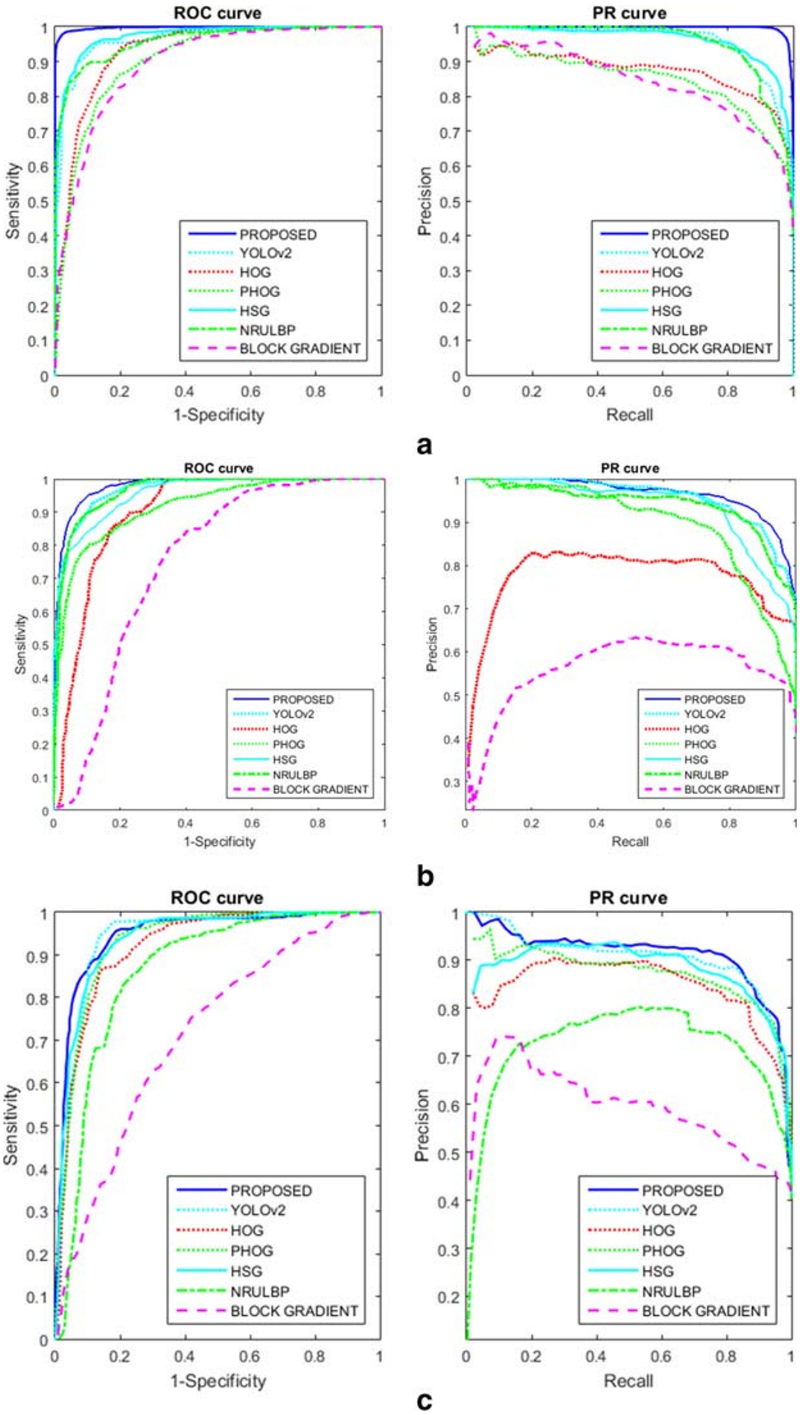


Fig. 6 Comparison of ROC and PR curves for the proposed and other detectors on (**a**) INRIA [4] (**b**) ETH [8] (**c**) TUD Brussels-motion pairs [31]

detectors on INRIA dataset. The AUC for ROC and PR curve is shown in Fig. 7. The AUC for the proposed framework is highest i.e. 0.9976 followed by HSG (0.9736), NRULBP (0.9607) and block gradient (0.8947). It indicates that the proposed detector has better

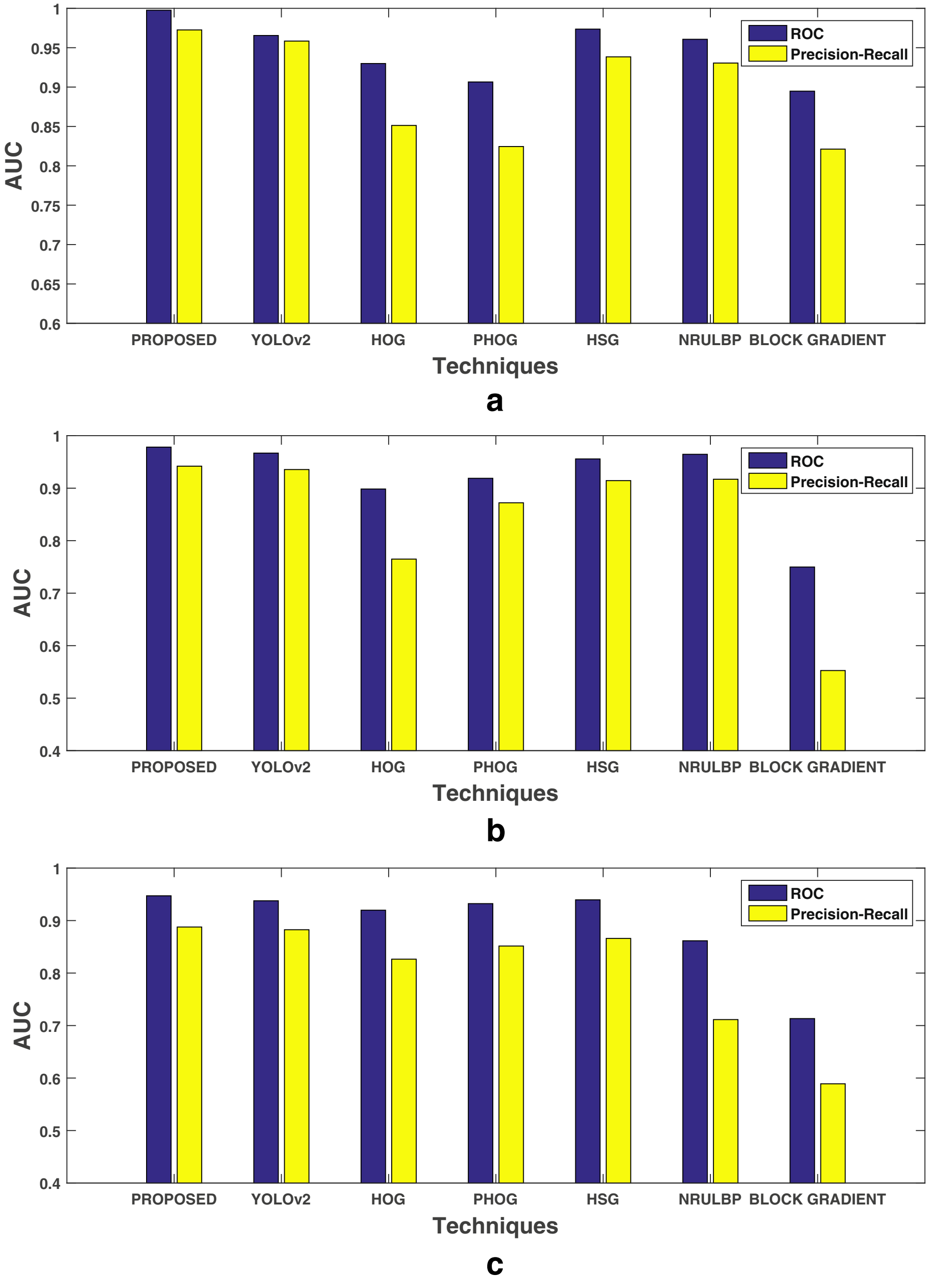


Fig. 7 Comparison of AUC for ROC and PR curves for the proposed and other detectors on (**a**) INRIA [4] (**b**) ETH [8] (**c**) TUD Brussels-motion pairs [31]

Table 2 Comparison of performance of various classifiers used with proposed descriptor

Classifiers → Metrics ↓	SVM linear	SVM quadratic	SVM polynomial	SVM RBF	SVM MLP
Accuracy(%)	97.74	94.45	92.03	59.41	76.11
Precision	0.9847	1	0.9967	NaN	0.6867
Recall	0.9594	0.8633	0.8065	0	0.7564
Specificity	0.9898	1	0.9982	1	0.7643
NPV	0.9728	0.9146	0.8831	0.5942	0.8213
Miss rate	0.0406	0.1367	0.1935	1	0.2436
Fall out	0.0102	0	0.0018	0	0.2357
FDR	0.0153	0	0.0033	NaN	0.3133
For	0.0272	0.0854	0.1169	0.4058	0.1787
F1 score	0.9719	0.9267	0.8915	0	0.7199
MCC	0.9533	0.8886	0.8414	NaN	0.5143
Youden’s index	0.9492	0.8633	0.8046	0	0.5208

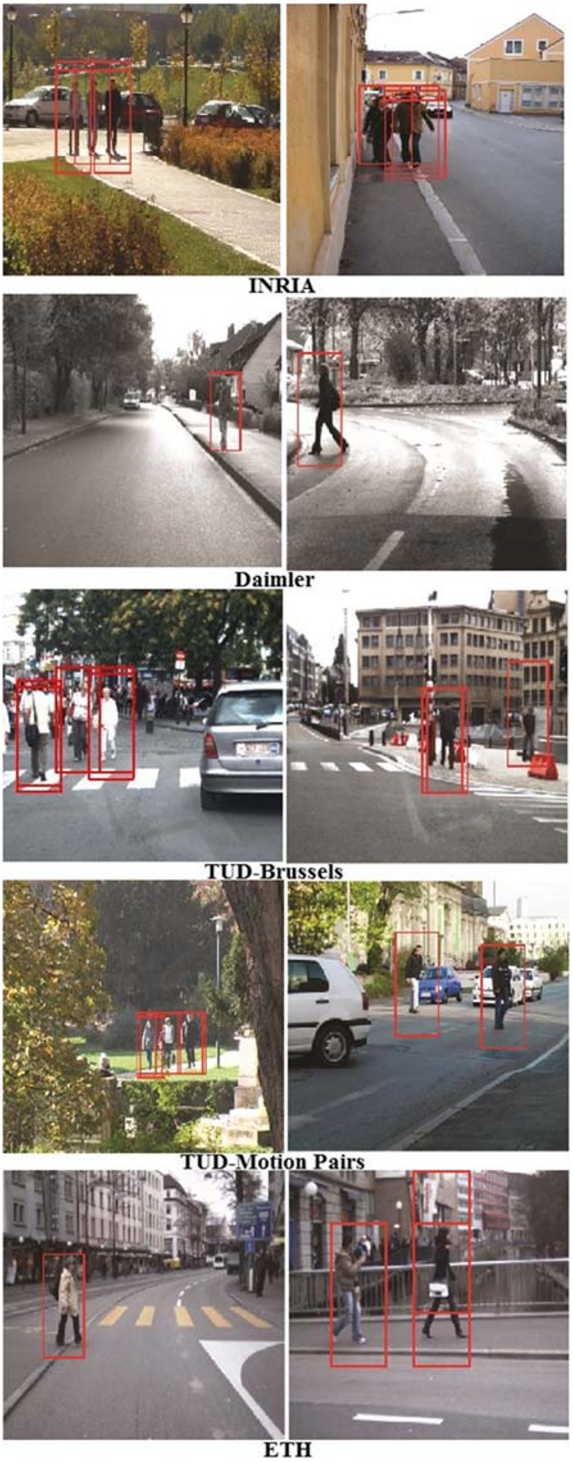
detection performance over other detectors. Tables 2 and 3 provides the performance comparison of various classifiers used with proposed technique. It can be observed that among SVM classifiers, linear SVM has highest accuracy i.e. 97.74 followed by quadratic (94.45), polynomial (92.03), MLP (76.11) and RBF (59.41). Precision is highest for quadratic SVM but its recall is less than proposed technique. In such situation, it is preferable to look for F1 score. It can be observed that F1 score is highest for proposed framework (0.9719) followed by quadratic SVM (0.9267), polynomial SVM (0.8915), naive bayes (0.8885), LDA (0.7886), KNN (0.7254) and others. Observing type I and type II errors i.e. miss rate and fall out, it is found that miss rate is lowest for naive bayes i.e. 0.0081 but its fall out is quite high (0.1645). Similar is the case for other classifiers but for the proposed approach, both are low.

Thus, linear SVM comes out to be the best classifier for the proposed descriptor compared to others and hence in this paper it is preferred over other classifiers for creating classification model. Figure 8 shows the detection results of proposed framework on INRIA, daimler, TUD-brussels-motion pairs and ETH datasets. Since a window detection stride of 4 is chosen, it leads to multiple detection for the pedestrian around the same area of the image. No window scaling has been performed and a fixed size of 64 × 128 was used.

Table 3 Comparison of performance of various classifiers used with proposed descriptor (continued)

Classifiers → Metrics ↓	Naive Bayes	LDA	KNN	Decision tree
Accuracy(%)	89.89	80.39	81.32	81.71
Precision	0.8046	0.7011	0.8998	0.8452
Recall	0.9919	0.9012	0.6076	0.6725
Specificity	0.8355	0.7375	0.9538	0.9159
NPV	0.9934	0.9162	0.7806	0.8037
Miss rate	0.0081	0.0988	0.3924	0.3275
Fall out	0.1645	0.2625	0.0462	0.0841
FDR	0.1954	0.2989	0.1002	0.1548
For	0.0066	0.0838	0.2194	0.1963
F1 score	0.8885	0.7886	0.7254	0.7491
MCC	0.8126	0.6279	0.6180	0.6180
Youden’s index	0.8274	0.6387	0.5614	0.5884

Fig. 8 Pedestrian detection outcome using the proposed framework on various datasets



5 Conclusion

In this paper, we predominantly concentrated on presenting an improved descriptor which combines HSG and NRULBP descriptors to provide more valuable feature information about pedestrian present in images. Due to the combined shape and texture descriptors, the algorithm has a very low failure rate, therefore this approach is ideal for pedestrian detection on roads. Also various classifiers like SVM, naive bayes, KNN, LDA and decision tree are used with proposed framework and it is found that SVM with linear kernel function provides superior metric values than others. Detection outcome is obtained on INRIA, daimler, TUD-brussels-motion pairs and ETH datasets while performance metrics are evaluated using experiments performed on INRIA datatset. Thus, the proposed model for feature extraction can be suited for usage in real time pedestrian detection systems.

References

1. Bay H, Ess A, Tuytelaars T, Gool LV (2008) Speeded-up robust features (SURF). Comput Vis Image Underst 110(3):346–359
2. Benenson R, Omran M, Hosang J, Schiele B (2015) Ten years of pedestrian detection, what have we learned? In: Agapito L, Bronstein M, Rother C (eds) Computer vision - ECCV 2014 workshops, pp 613–627
3. Bilal M, Khan A, Khan MUK, Kyung CM (2017) A low-complexity pedestrian detection framework for smart video surveillance systems. IEEE Trans Circ Syst Video Technol 27(10):2260–2273
4. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE Computer society conference on computer vision and pattern recognition 2005. IEEE, pp 886–893
5. Dhiman C, Vishwakarma DK (2019) A robust framework for abnormal human action recognition using R-transform and zernike moments in depth videos. IEEE Sens J 19:5195–5203
6. Elmikaty M, Stathaki T, Kimberm P, Giannarou S (2012) A novel two-level shape descriptor for pedestrian detection. In: Sensor signal processing for defence (SSPD 2012). IEEE, pp 1–5
7. Enzweiler M, Gavrila DM. (2009) Monocular pedestrian detection: survey and experiment. IEEE Trans Pattern Anal Mach Intell 31(12):2179–2195
8. Ess A, Leibe B, Schindler K, van Gool L (2009) Moving obstacle detection in highly dynamic scenes. In: Proceedings of ICRA, pp 56–63
9. Freeman W, Roth M (1995) Orientation histograms for hand gesture recognition. In: IEEE International workshop on automatic face and gesture recognition. IEEE
10. Guo Z, Zhang L, Zhang D (2010) Rotation invariant texture classification using LBP variance (LBPV) with global matching. Pattern Recogn 43(3):706–719
11. Li D, Zhang Z, Chen X, Huang K (2019) A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. IEEE Trans Image Process 28:1575–1590
12. Liao S, Chung ACS (2007) Face recognition by using elongated local binary patterns with average maximum distance gradient magnitude. In: Yagi Y, Kang SB, Kweon IS, Zha H (eds) Computer Vision – ACCV 2007. Springer
13. Liu H, Chen S, Kubota N (2013) Intelligent video systems and analytics: a survey. IEEE Trans Industr Inform 9(3):1222–1233
14. Liu W, Yu B, Duan C, Chai L, Yuan H, Zhao H (2015) Pedestrian-detection method based on heterogeneous features and ensemble of multi-view-pose parts. IEEE Trans Intell Transp Syst 16(2):813–824
15. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
16. Nguyen DDT, Zong Z, Ogunbona P, Li W (2010) Object detection using non-redundant local binary patterns. In: 17th IEEE International conference on image processing (ICIP) 2010. IEEE, pp 4609–04612
17. Ojala T, Pietikainen M, Harwood D (1994) Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In: International conference on pattern recognition. IEEE
18. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987

19. Oren M, Papageorgiou C, Sinha P, Osuna E, Poggio T (1997) Pedestrian detection using wavelet templates. In: Proceedings of IEEE computer society conference on computer vision and pattern recognition 1997. IEEE, pp 193–199
20. Porikli F, Yilmaz A (2012) Object detection and tracking. In: Shan C, Porikli F, Xiang T, Gong S (eds) Video analytics for business intelligence. Studies in computational intelligence, vol 409. Springer, Berlin, pp 3–41
21. Redmon J, Farhadi A (2017) YOLO9000: better, faster, stronger. In: Proc - 30th IEEE conf comput vis pattern recognition CVPR 2017 2017–January, pp 6517–6525
22. Sheenu JG, Vig R (2015) Histograms of orientation gradient investigation for static hand gestures. In: International Conference on computing, communication and automation (ICCCA) 2015. IEEE, pp 1100–1103
23. Viola P, Jones MJ, Snow D (2005) Detecting pedestrians using patterns of motion and appearance. Int J Comput Vis 63(2):153–161
24. Vishwakarma DK, Kapoor R (2015) Integrated approach for human action recognition using edge spatial distribution, direction pixel and R-transform. Adv Robot 29:1553–1562
25. Vishwakarma DK, Kapoor R, Maheshwari R et al (2015) Recognition of abnormal human activity using the changes in orientation of silhouette in key frames. In: 2015 International conference on computing for sustainable global development, INDIACo 2015, pp 336–341
26. Vishwakarma DK, Dhiman A, Maheshwari R, Kapoor R (2015) Human motion analysis by fusion of silhouette orientation and shape features. In: Procedia computer science, pp 438–447
27. Wang X, Han TX, Yan S (2009) An HOG-LBP human detector with partial occlusion handling. In: 2009 IEEE 12th international conference on computer vision. IEEE, pp 32–39
28. Wang J, Liu P, She MFH et al (2011) Human action recognition based on pyramid histogram of oriented gradients. In: Conference proceedings - IEEE international conference on systems, man and cybernetics, pp 2449–2454
29. Wang X, Wang M, Li W (2014) Scene-specific pedestrian detection for static video surveillance. IEEE Trans Pattern Anal Mach Intell 36(2):361–374
30. Wojek C, Schiele B (2008) A performance evaluation of single and multi-feature people detection. In: Rigoll G (ed) Pattern recognition, pp 82–91
31. Wojek C, Walk S, Schiele B (2009) Multi-cue onboard pedestrian detection. In: 2009 IEEE computer society conference on computer vision and pattern recognition workshops, CVPR workshops, pp 794–801
32. Wu J, Liu N, Geyer C, Rehg JM (2013) C4: a real-time object detection framework. IEEE Trans Image Process 22(10):4096–107
33. Yao S, Wang T, Shen W, Pan S, Chong Y, Ding F (2015) Feature selection and pedestrian detection based on sparse representation. PLoS ONE 10(8):e0134242
34. Yuan X, Shan X, Su L (2011) A combined pedestrian detection method based on haar-like features and HOG features. In: 2011 3rd international workshop on intelligent systems and applications. IEEE, pp 1–4
35. Yuan G, Zhang X, Yao Q, Wang K (2011) Hierarchical and modular surveillance systems in its. IEEE Intell Syst 26(5):10–15
36. Zhang G, Gao F, Liu C, Liu W, Yuan H (2010) A pedestrian detection method based on SVM classifier and optimized histograms of oriented gradients feature. In: 6th International conference on natural computation (ICNC) 2010. IEEE, pp 3257–3260
37. Zhang S, Wang C, Chan SC, Wei X, Ho CH (2015) New object detection, tracking, and recognition approaches for video surveillance over camera network. IEEE Sensors J 15(5):2679–2691
38. Zhang S, Benenson R, Omran M et al (2018) Towards reaching human performance in pedestrian detection. IEEE Trans Pattern Anal Mach Intell 40:973–986
39. Zhu C, Peng Y (2015) A boosted multi-task model for pedestrian detection with occlusion handling. IEEE Trans Image Process 24(12):5619–5629

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Kaushal Kumar received the M.Tech. degree in VLSI design from the Centre for Development of Advanced Computing, Mohali, Punjab, in 2015. He has two and half years of industrial experience. He is currently pursuing the Ph.D. degree from National Institute of Technology Patna, India. His research interests are Image processing, Machine learning and VLSI architectural design.

Ritesh Kumar Mishra is currently working as Assistant professor in the Department of Electronics and Communication Engineering at National Institute of Technology Patna, India. He has 15 years of teaching experience and 3 years of industrial experience. He is a member of IEEE, Institution of Engineers and is also connected with Indian Society of Technical Education. His research interests are Wireless Communication, Antenna and Computer Vision.